FUNDAMENTAL
PHYSICS

BY

Lloyd William Taylor

OBERLIN COLLEGE

WITH THE COLLABORATION, IN THE CHAPTERS ON MODERN PHYSICS, OF *Forrest Glenn Tucker*, OBERLIN COLLEGE

HOUGHTON MIFFLIN COMPANY

Boston New York Chicago Dallas Atlanta San Francisco

The Riverside Press Cambridge

The Riverside Press
CAMBRIDGE • MASSACHUSETTS
PRINTED IN THE U.S.A.

TO THE MEMORY OF

Carrie Brown Taylor

Preface

EIGHT MONTHS after the publication of an earlier textbook by the author of the present one, the United States became one of the belligerents in the most widespread and destructive war of all time. The author's earlier effort, like substantially all similar undertakings by other writers, had been prepared for an educational system designed to serve a world at peace. Since physics was already indicated as one of the academic subjects most useful in the war effort, the publisher almost immediately suggested that another textbook be prepared, adapted to serve the new requirements. The present work is the result.

In his earlier work, *Physics: the Pioneer Science,* the author tried to show how to break the vicious spiral initiated by narrow specialization in the education of physicists for times of peace. With the advent of war we have all had to turn aside temporarily, though with deep regret, from such peacetime objectives as do not happen to lie also in the direct line of the war effort. This war is frequently called "a physicist's war" and with much justification. Certainly the physicist is confronted with grave new responsibilities, the most exacting of which is the training of other and even better physicists at the maximum possible rate.

Right there lies an unprecedented opportunity which is not as widely recognized as it should be. We teachers of physics will be wise, with the deep wisdom of the preservation of a tolerable social order, if we keep acutely in mind the fact that we are training physicists, not only for the war, but for the even more exacting era that is to follow. After having won the war, the United Nations will be able also to "win the peace" only if our technical men, competent specialists though they must be, are also a great deal more than *mere* specialists. Men of science, above all others, must comprehend the scope and the broad implications of the scientific habit of thought as well as the limitations on the fields of its applicability. This comprehension does not automatically accompany the mastery of a particular field of science, though such mastery is one of the important elements in that kind of appreciation. But in one way or another the much-needed combination of technical knowledge with scientific and social perspective must be provided if the world is not to be herded again onto the highroad to self-destruction.

The most fertile ground for the growth of an appreciation of the scientific habit of thought is a familiarity with the struggles that were involved in developing it. Physics, in company with astronomy, which is really a specialized branch of physics, was involved far earlier and more deeply in this development than any other branch of science. It constituted the first large body of knowledge that was experimental, sequential, and cumulative. When the scientific millennium comes, it will begin in a widespread familiarity with the early struggles of physics as the pioneer science. Most great movements have grown out of similar bodies of tradition. But physicists must lead the way if general interest in scientific tradition is to develop.

In this book the strictly historical material has been reduced in extent and modified to meet the needs of the times. Classroom experience has demonstrated unmistakably that the historical approach is an exceedingly effective vehicle for the development of comprehension of subject matter. The student who experiences an introduction to physics which properly combines full technical exposition with the leavening of historical tradition will be a better physicist, and in addition a much more useful and far-sighted citizen, than the student whose training is confined to either aspect alone.

The most considerable change in this book, as compared with *Physics: the Pioneer Science*, is in the section on electricity. In the earlier book the historical approach dictated the study of static electricity before current electricity. In the present book the greater directness and practical advantage of going at once into the study of current electricity suggested a reorganization and rewriting of the section, with electrostatics reduced in extent and placed after the more utilitarian aspect of the subject.

The meter-kilogram-second system of units has been used in this as in the earlier text. Most physicists are aware that it was prescribed in 1935 by the International Committee on Weights and Measures to take effect January 1, 1940. It has been making steady gains and is coming to be recognized as a major factor in a long-needed simplification of units, especially electrical units. For any readers not already familiar with the system, a brief description is provided in the Appendix, incorporating a table comparing the M.K.S. system of units with the old C.G.S. system.

It has seemed desirable to provide citations for the many quotations and direct references to original sources. In place of scattering them out in footnotes, a serially numbered list of references to books has been included in the Appendix. Reference is made to these in the text by serial number followed by a second number to indicate the page. In case of works consisting of more than one volume, the number of the volume is interpolated in italics between the serial and the page numbers. For example: Faraday's *Experimental Researches in Electricity*, volume 3, page 160, would appear as 90:*3*:160. The list includes only books. References to periodicals are in footnotes.

Perhaps a word about the problems which follow most of the chapters will not be out of place. A majority of the problems are formulated in algebraic terms. Algebraic solution, while not necessary, should be encouraged from the beginning for the better students, and a certain amount of it required of all students as their maturity increases. Several sets of numerical data (usually four) are provided for each example. This gives material for additional drill, when required on a particular point, and helps to solve the literally perennial problem of duplication in successive years. For the convenience of teachers who may desire at least an admixture of the conventional type of problem, several of these have been included in each set. Answers are furnished, though the values are carried to only two significant figures, whereas three or even four may be required of the student. This, in the experience of the author, avoids both horns of an old dilemma.

The considerable extent to which the substance of *Physics: the Pioneer Science* has been carried over, with appropriate modifications, into this text will be clear to anyone familiar with both. Chapters 48 to 51 inclusive were written by one of the author's associates, Dr. F. G. Tucker. Chapter 47 was contributed by another associate, Dr. C. E. Howe. Acknowledgments of permission to use copyrighted material have been made at appropriate points.

LLOYD W. TAYLOR

OBERLIN COLLEGE

Contents

Mechanics

Heat

Sound

Light

Electricity

MECHANICS

CHAPTER 1

The Place of Mechanics in the Intellectual Enterprise

Physics in the World Crisis

Even in times of peace the study of physics as the pioneer science possessed a significance all out of proportion to the place accorded to it in the conventional Liberal Arts curriculum. But war, fought with the weapons of the nineteen-forties, is veritably "a physicists war." Physical meteorologists are in demand by the tens of thousands, radio technicians by the hundreds of thousands, and automobile and aviation mechanics almost literally by the million. In addition, men trained in other branches of applied physics are in demand in corresponding numbers. The present prospect is that the demand for such men will be scarcely diminished by the cessation of the war, for there are indications of a post-war technological era far surpassing in extent anything imagined heretofore. All this lays a grave responsibility on the little group of physicists out of whose profession almost every branch of engineering has directly or indirectly grown.

Yet there is another reason for the study of physics which is even broader and more compelling than the foregoing. Physics has, from its very inception, been the spearhead of the free, untrammeled search for truth. Physics and astronomy together established the basic scientific axiom that the problems of the physical world are solvable, and solvable by man. Ever since then physics has been the color-bearer of the fierce scientific conviction that man *can* become the master of his fate. And it has given to the rest of the world, with the aid of the other sciences, the doctrine that this can come about only at the price of rigorous intellectual integrity.

These two chief components of intellectual morale, man's confidence in his supremacy over nature and an utter devotion to intellectual integrity, have in recent years had to contend against fearful odds. The deepest issues of the present crisis of the world are intellectual. The day when humanity may really become the master of its intellectual fate has already been deferred by the totalitarian onslaught, and can still be lost. Totalitarian philosophy has always been actively anti-intellectual. The totalitarians do not like to have people use their minds and from the beginning have been committed to produce a society in which it would be impossible.

Even in this country it was a huge intellectual backslide that gave the totalitarians their chance — intellectual sloth, lack of imagination, and wishful thinking. Nor is the danger yet over. Lurking in the background of aroused consciousness of danger to the institutions of intellectual liberty are disturbing remnants of the anti-intellectual movement. Even without a totalitarian victory there is danger that the Axis powers may have done a permanent disservice to the race if the present intellectual defeatism, for which they are largely responsible, persists.

The Sciences Are Pioneers in the War for Intellectual Freedom

Yet this contest with intellectual sloth, defeatism, and appeasement is nothing new for physics. Physics was born in just such a contest, and has constituted the shock troops in an unremitting intellectual war ever since. Just as generals and admirals must know the history of military and naval strategy if they are to avoid the mistakes of the past, so one can with great profit study the three-hundred-year campaign which physics has waged to establish integrity and self-confidence as the main attributes of the scientific method, a campaign in which it has been ably abetted by the other sciences as they came into existence.

When science is viewed in this larger aspect, it is not its *effects*, profound and far-reaching though they are, that should be the primary interest of the student, but the nature of the instrument itself. It is utterly unique. Literature and the arts have been produced principally by special geniuses, and the rate of such production appears neither to grow nor to improve with passing time. Current masterpieces of art and literature are of no greater merit, nor are they being produced any more profusely today in proportion to the population, than the corresponding products of two thousand years ago. In the sciences, on the other hand, is the first large body of knowledge that is both *sequential* and *cumulative*. As a unified army, organized for a sustained assault upon the citadel of human ignorance, there has been nothing to compare with the sciences in the whole recorded development of human thought. It is possible to question the value of the material which the sciences discover; much of it seems trivial to the lay mind. One may also be fearful of the ultimate effect of scientific philosophy on human welfare; many thoughtful men hold science responsible for some of the major ills of the day. But whether for good or for evil, the fact that science dominates modern thought cannot be disregarded.

Historians have until recently been curiously blind to the scientific idea as a force influencing the trend of world affairs. But there are indications that they are waking from their exclusive preoccupation with imperialistic rivalries and military campaigns to an awareness of what is really the most significant factor in the history of the last three centuries. Thus, Preserved Smith says in his *History of Modern Culture* (117:1:606):[1]

> . . . whether as the new salvation or the new superstition, science has molded

[1] See the List of References in the Appendix.

the whole life of the modern world.... All modern production of wealth, all contemporary life, depend on the knowledge of nature acquired by science. But more than that, religion, politics, philosophy, art and literature have capitulated to science, or at least receded before her. There is no department of human activity today untouched with the spirit of experiment and of mathematics.

Difficulties in Attaining Scientific Appreciation

And yet it is not easy to know where to turn to acquire an appreciation of the scientific point of view or an informed and discriminating estimate of its effects. Historians interest themselves but slightly in that field. Philosophers, to whom one might naturally turn, seldom possess the first-hand familiarity with scientific subject matter which is a necessary condition for penetrating judgments in this field. Even men of science themselves are usually too busily engaged on the scientific battle front to give serious attention to explaining to visitors the larger aspects and significances of the campaign, even in the cases of those unusual men who know what these are themselves.

There is, moreover, one handicap under which all approaches to this problem from the historical, philosophical, or general-survey directions seem inevitably to labor. That is the fact that they are largely limited to *talking about* science, whereas one of the inescapable conditions of acquiring an appreciation of the nature of science is to come closely to grips with a representative portion of it. This is scarcely surprising, since it is also true of any other intellectual discipline. It is hard to imagine that one could arrive at an appreciation of literature or music or art without making a first-hand study of some of the more representative works in these various fields. One who felt that he possessed an adequate appreciation of any of these fields through merely discussing the generalizations made by various commentators, in the absence of any first-hand acquaintance on his own part, would simply be self-deceived. For this reason it seems necessary to return to some one of the specific sciences.

Physics as the Key to All the Sciences

Every science shares, each in its characteristic degree, the heritage called "the scientific method." By no means, however, is every science equally qualified to stand as an example of this heritage. There is, in fact, a wide range of gradations in the extent to which the various sciences have participated in the formation of the world view which lies at the foundation of the scientific era. Other things being equal, there is every reason for choosing the science which has been the most prominent in this respect, and which is today influencing such thought the most profoundly, namely physics.

Physics is, by common consent, the fundamental science. This is not merely because it fell to the lot of physics, in company with astronomy, to carve the place for science out of the highly resistant intellectual world of the sixteenth to the nineteenth centuries, but more particularly because the younger sciences, without exception, consider themselves scientific to just the extent that their concepts are logically reducible to those of physics. In addition, the basic technique of laboratory observation and measurement used by all the other sciences consists primarily of adaptations of the methods of physics. Consider, for example, the astronomical telescope, the chemical balance, the biological and the petrographic microscopes, the geological seismograph, the medical electrocardiograph and X-ray, and almost the entire equipment of the engineer. The basic instruments from which these were devised were born in the physics laboratory, and most of them, except for minor adaptations, were perfected there.

Evolution of the Subject Matter of Physics

Until about a hundred years ago the subject matter of all the physical sciences, including physics, astronomy, chemistry, and engineering, remained in an undifferentiated mass called natural philosophy. Men interested in that general field worked sometimes in one section of it and at other times in another. The body of knowledge in all these subjects was so slight that it was possible for a diligent and capable man to be an authority in all of them.

But early in the nineteenth century all this was changed. The invention by James Watt, which in 1769 made the steam engine practicable, created the practical engineers, a group which grew more and more important as the internal-combustion engine was developed a century later, at about the time that technical applications of electricity also became of importance. John Dalton, in 1808, placed the ancient concept of the atom on a firm experimental basis, whereupon chemistry graduated from the alchemy stage and became a real science. In the meantime, astronomy, while using more and more the concepts and tools of physics, was also becoming progressively specialized in its own field. The remainder of what had been the working material of natural philosophy was thus left to physics.

This remainder was heterogeneous enough. Mechanics, heat, sound, light, electricity, and magnetism — to the uninitiated these appear to be almost entirely distinct subjects. Yet as long ago as the first part of the nineteenth century a fair start had been made in unifying them. Just before the century opened, Rumford, the Massachusetts Yankee who became a Bavarian count, had performed the first experiments to show that heat is a mode of molecular motion, and had thereby broken down the barrier between heat and mechanics. In 1829, the Frenchman Ampère (under the stimulus of a suggestive experiment by the Danish physicist Oersted) showed that magnetism could be reduced "to effects purely electric" (30:248). Thus the way was opened to the close association between electricity and magnetism which characterizes the modern knowledge of physics. The demonstration that sound, known to be born in

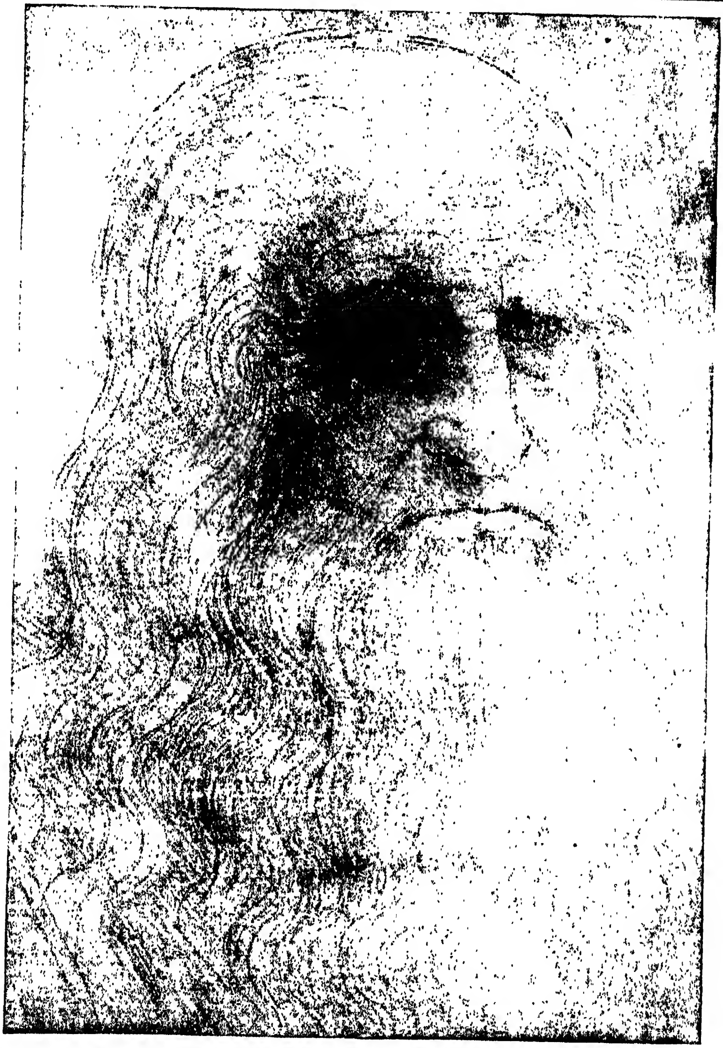


FIG. 1. LEONARDO DA VINCI
A portrait drawn by himself. (Courtesy of Fisher Scientific Company.)

motion, consisted of vibrations in the air, and that light was fundamentally an electromagnetic phenomenon, were discoveries reserved for the last half of the century. They went far toward knitting into a homogeneous logical system the mass of phenomena which superficially appear so unrelated. As the nineteenth century rounded into the twentieth, the unification was carried still further by the positive identification of the electrical structure of atoms. With that discovery, the previous tendency to interpret electrical phenomena exclusively in terms of mechanics gave way to the attempt to interpret the properties of matter in terms of electrical concepts.

Mechanics as the Key to Physics

Permeating this entire course of development is one fact which is of great significance: the place occupied by mechanics at the logical foundation of the whole structure. At first sight the reason for this seems somewhat perplexing. Mechanics, concerned, as Mach puts it, "with the motions and equilibrium of masses" (74:1), seems to have little to do with heat and sound, and still less with electricity and light. But the experimental demonstration by Rumford that heat is merely the energy of random molecular motion, and the attribution by Helmholtz of musical tones to periodic motions of molecules of the air, definitely associated these two groups of phenomena with "motions and equilibrium of masses," and hence identified them as mechanical in nature.

The fundamental units of both electricity and magnetism were defined in terms of forces between electrical charges, or produced by the magnetic interaction of adjacent currents, from which these two subjects became capable of treatment in terms of mechanical concepts.

The subject of light seems to have withheld itself the longest from mechanical treatment, though its identification as an electromagnetic radiation suggested its ultimate subjugation. It finally succumbed, however, toward the close of the first quarter of the twentieth century, to what might be termed an attack from the rear, in the successful application of the mechanical concepts of conservation of energy and conservation of momentum of photons or "light darts," which were shown to be inescapable attributes of what had for a century been considered pure waves.

There is, however, another reason for the prominence which mechanics possesses in any treatment of physics. An intuitive knowledge of the basic phenomena of mechanics is part of the equipment of every individual, however non-technical his education may be. The knowledge is hazy and unanalyzed, to be sure, but it is intensely practical. It is learned in the hard school of experience: the bumps of babyhood and all the complex muscular reactions acquired in early years. Every normal individual senses the substance of Newton's laws of motion sufficiently to control movements in accordance with them, even though he may not comprehend them in sufficient generality to phrase them in words. It is perhaps for this reason, more than for any other, that mechanics constitutes the alphabet of the physical sciences. Like all alphabets, it may be found, when scientific maturity is acquired, to be logically indefensible, but having been learned at the knee of mother nature, mechanics is irretrievably built into the structure of scientific thought before we acquire the degree of discrimination necessary to question its justification. For better or for worse, mechanics is basically important in any study of physics.

Questions for Self-Examination

1. What might be an advantage in approaching physics from the point of view of its contributions to the non-scientific intellectual world?
2. Tell why a comprehension of the *nature* of the sciences is at least as important as the more common study of their subject matter.
3. What are some of the bases for the common statement that physics is "the fundamental science"?
4. Trace the development of physics from its old realm, called natural philosophy, to its present form.
5. Why is mechanics commonly regarded as the foundation of physics?

CHAPTER 2

Measurement

The Establishment of Standards

The science of mechanics is, as has already been pointed out (page 8), concerned "with the motions and equilibrium of masses." Any adequate description of motion requires the measurement of length and of time, and hence has brought about the establishment of standard units of length and time. Similarly, a standard unit of mass is required before numerical values can be specified in that field. The hand-to-mouth necessities of trade forced the establishment of regional standards of length and mass before the dawn of history, and formal specification of the day as a unit of time is without doubt even older.

Little by little over many centuries the numerous regional units of length and mass have given way to units applicable over wider areas. Finally, in 1875, international standards of length, mass, and time were agreed upon in a convention meeting in Paris at which most of the civilized countries of the world were represented. Today those standards have been made official the world around.

The international standards thus adopted are based on what is termed the *metric system.* Though use of the metric system is permissive everywhere and has in practice been very generally put into common use, it is unfortunate that this use is not yet quite universal. There are two nations, and only two, who have not yet joined the otherwise complete roster. These are Great Britain and the United States. Even in these countries the official standard units of length and mass are the meter and the kilogram. Those in common use, however, the foot and the pound, are defined as .30480 meter and .45359 kilogram respectively, and in all except scientific practice completely displace the metric units.[1]

The Standard of Length

The international standard meter is the distance between two marks on a certain platinum-iridium bar preserved in the International Bureau of Weights and Measures at Sèvres, near Paris. Originally it was intended that this standard of length should be exactly one ten millionth of the dis-

[1] In this text, English units are used, for the most part, in the early chapters. After Chapter 10 metric units are used almost exclusively.

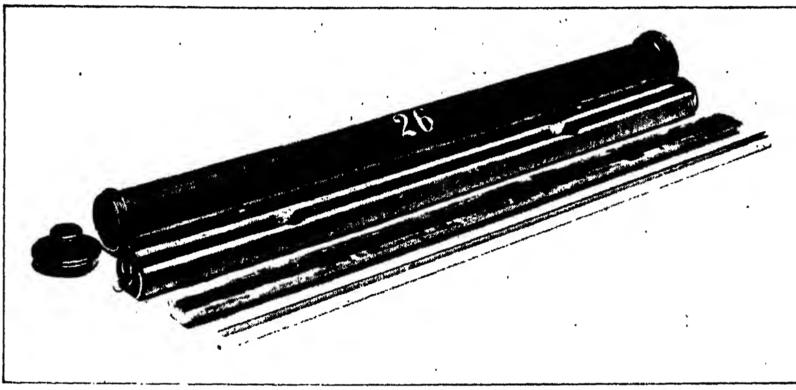


FIG. 2. ONE OF THE ORIGINAL STANDARD METER BARS

This is No. 26. It was made at the same time as No. 6, later chosen as the international standard of length. No. 6 has, however, never been photographed.

tance between the equator and the pole. But after the meter had been established as the world's standard of length, more accurate measurements of the size of the earth showed that the length of this standard was not quite the one ten millionth it had been supposed. Hence, its original qualification of bearing a simple ratio to the size of the earth had to be abandoned, but the standard remains. Since then it has been determined to be 1,553,164.13 wave lengths of the red spectroscopic line of cadmium (83:85). Thus, if some catastrophe should destroy the world's primary standard of length it could be reproduced with an accuracy of one part in 155 million. Secondary standards, copies of the primary standards, have of course been made in considerable numbers and distributed to the standardizing agencies of all the countries in the world. In the United States each state is provided with a tertiary standard copied from our national secondary. Figure 2 shows one of the international standard meter bars (in the foreground).

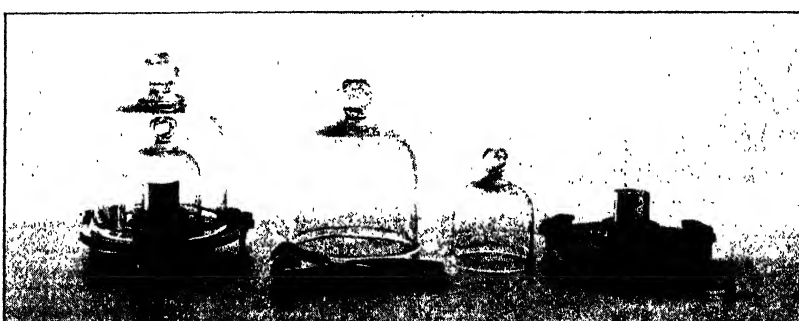


FIG. 3. TWO OF THE UNITED STATES STANDARD KILOGRAMS

The pound is .45359 kilogram.

The Standard of Mass

The United States made several gestures toward establishing its own standard of mass, beginning with George Washington's first annual address to Congress in 1790. Such standards as were adopted were for the most part taken over from the English. In 1875, the International Standard Kilogram became our standard along with the International Meter. When the French originally established the kilogram in 1792, they had intended it to represent the mass of one hundred cubic centimeters (one liter) of water at the temperature of maximum density (about 4° C.). But again, increasing accuracy of measurement showed this standard, like the original meter, not to be exactly what was intended and this standard, too, had to be re-established arbitrarily. Figure 3 shows (uncovered, on the right) the kilogram of the United States. The one on the left (covered) is another standard showing the double bell-jar protection given these standards when not in use.

The Standard of Time

The basic unit of time is the sidereal day: that is, the time from the passage of a star across a meridian to the next passage of the same star. Because of the revolution of the earth around the sun, sidereal time differs from mean solar time by one part in $365\frac{1}{4}$. The solar day, instead of the sidereal, is the unit of time upon which human activities are scheduled. The length of the solar day is subject to several varieties of fluctuation and cannot be advantageously used as a basic standard. But for practical purposes the overall average of these fluctuating solar days is termed the mean solar day. It is set as 1.00274 sidereal days, the solar day being thus nearly four minutes longer than the sidereal day. The mean solar second, the International Standard time interval, is then defined as $\frac{1}{86400}$ mean solar day ($\frac{1}{60} \times \frac{1}{60} \times \frac{1}{24}$).

Derived Units

It is necessary to have units that are both smaller and larger than the International Standards of length, mass, and time. For length and mass the derived units are decimal divisions or multiples of the standards. Thus, the *centimeter* (cm, one hundredth of a meter, about two fifths of an inch) and the *millimeter* (mm) are the most common subdivisions of the meter. For microscopic lengths the *micron* (μ), 10^{-6} m,[1] and the *millimicron* ($m\mu$) are in common use. A still smaller unit of length, the *angstrom unit*, (10^{-10} m), is commonly used to specify the wave lengths of light

[1] Powers of 10 are commonly used as shorthand representation of extremely large multiples and extremely small subdivisions. Thus

$$10^{6} = 1{,}000{,}000$$

$$10^{-6} = \frac{1}{1{,}000{,}000}, \text{ etc.}$$

and other submicroscopic quantities. The most common multiple of the meter is the *kilometer* (km, 1000 meters, about .6 mile).

Common metric units of volume are the *cubic meter* (m^3), the *liter* (l, the volume of a cube 10 cm on an edge, almost exactly a quart), and the cubic centimeter or *milliliter* (ml).

For mass, the most common subdivisions are the *gram* (g, 10^{-3} kg, about $\frac{1}{30}$ oz) and the milligram (mg). The most common multiple of the kilogram is the *metric ton* (1000 kg, about 2205 lbs).

Density

Two of the simplest and most common combinations of the units described above are *density* and *velocity*.[1] Density is a very old and very common concept. Archimedes (287–212 B.C.) seems to have been the first to glimpse its real implications. By long tradition they are supposed to have occurred to him in the famous bathtub episode referred to in Chapter 8.

Density is defined simply as mass per unit volume. In the English system the density of water is conventionally 62.4 pounds per cubic foot. In the metric system it may be stated either in grams per milliliter or, as will be done here, in kilograms per liter. The numerical value is the same in either case. Its value for water is very nearly unity, fluctuating slightly with changing temperatures.

Velocity

The second simple combination of basic units, a combination of length and time, yields the concept of velocity.

A motor car may be said to be traveling at a velocity of fifty miles per hour. It is thereby implied that if the velocity should remain unchanged, the car would in the course of an hour travel a distance of fifty miles, in two hours one hundred miles, and so forth. Such uniformity of velocity would be almost impossible for a motor car, though it might be approximated by an airplane. The theoretical possibility of realizing it is, however, one way of giving intelligibility to the concept of instantaneous velocity, the logic of which may otherwise appear somewhat troublesome to the beginner, notwithstanding the fact that the mere intuitive concept of instantaneous velocity is the common possession of every modern child.

Velocity, the rate of traversing distance, may thus be defined as the ratio of distance traveled to time elapsed. Not always is the fact that the words "rate" and "ratio" come from the same root so clearly shown. Miles per hour, feet per second, kilometers per hour, meters per second: these common expressions, though cumbersome in use, illustrate the basic idea of velocity as ratio of distance to time. Communication would be made easier if names were provided for some of the common units of velocity, to replace, for example, the awkward expression "miles per hour"; but this

[1] Strictly speaking, a velocity is not completely specified unless direction is given as well as magnitude. At this point only the magnitude of velocity is involved.

has been effected in nautical practice only, in the term *knot*. The knot, instead of being a unit of distance, the seagoing brother of the mile, as many landsmen perhaps think, is a unit of velocity. It means simply one nautical mile (6080 feet) per hour.

For the case of uniform velocity (non-uniform velocity will be treated in some of the following chapters), a convenient algebraic relation between distance, velocity, and time may be deduced from the definition of velocity. Since

$$\text{velocity} = \frac{\text{distance}}{\text{time}},$$

then

$$\text{distance} = \text{velocity} \times \text{time}.$$

This relation was first formally recorded by the Moslem natural philosophers of the twelfth century (2:105). Also the obvious remaining permutation of this relation

$$\text{time} = \frac{\text{distance}}{\text{velocity}}$$

is occasionally useful, though less frequently so than the previous two. To express these relations algebraically, where t represents time, v represents velocity, and s represents distance,

$$v = s/t \qquad s = vt \qquad t = s/v. \tag{1}$$

Strictly speaking, a velocity is not completely specified until direction as well as magnitude has been given. In such a case a velocity of forty miles per hour north would be considered different from one of forty miles per hour south. The distinction is a real one, but cannot be advantageously developed until the basic laws of motion have been studied (Chapter 10). Some writers reserve the word *speed* to indicate the mere magnitude of a velocity without reference to its direction.

Questions for Self-Examination

1. Name the world's standard of length, tell how it evolved, and briefly describe how it is preserved.
2. Name the world's standard of mass, tell how it evolved, and briefly describe how it is preserved.
3. Define density.
4. Define velocity and state some common units of velocity.
5. Discuss the time-velocity-distance relation for uniform motion.

CHAPTER 3

Free Fall

Significance of the Study of Free Fall

Motions characterized by uniform velocity are far less commonly encountered than is often supposed. Non-uniform velocities are almost always in evidence. But, though motions possessing such non-uniformities of velocity are usually irregular and complicated, some are capable of fairly simple description. One of the most common of these is the motion exhibited by bodies falling freely, that is, without hindrance of any kind.

There are several considerations that render the motion of freely falling bodies worthy of serious study. One is, of course, the fact that it is very frequently encountered, being involved as it is, not merely in cases of falling

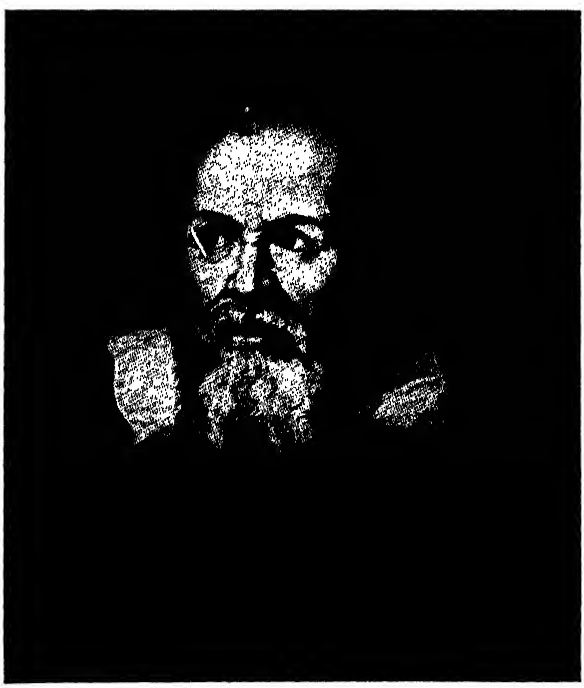


FIG. 4. GALILEO GALILEI (1564–1642)
(Courtesy of *Scripta Mathematica*.)

bodies as such, but as the principal factor in the motion of all projectiles. More important, perhaps, is the fact that the laws of motion of falling bodies are applicable to many motions in which there is no element of free fall involved. That is to say, free fall is even more significant as a type of motion than it is in its own right. But most important is the fact that it was in the conquest of the laws of falling bodies that physics, in the character which it bears today, won its spurs. When Galileo (1564–1642) near the end of the sixteenth century discovered how bodies fell, the significance of his accomplishment far transcended the content of his discovery, important though the content was.

The main thread of Galileo's procedure will be followed in the present approach to the study of falling bodies, though with some deviations. It is possible now, for example, to make a direct study of a falling body, which Galileo could not do because of lack of means for observation and measurement, a lack characteristic of his time.

Rate of Fall the Same for All Bodies in a Vacuum

That a freely falling body, dropped from rest, acquires a progressively greater velocity in the course of its fall is a fact of common observation. That the behavior of light bodies would be the same as that of heavy bodies if it were not for the effect of air resistance is a fact which is, even today, not always clear, and which up to the time of Galileo had proved to be an almost insurmountable obstacle to clarity of thought on this subject. With the aid of a vacuum pump it is possible now to remove most of the resistance presented by air to the free fall of bodies and thus to demonstrate, for example, that a feather will keep abreast of a coin in the course of a fall in a vacuum (Fig. 5).

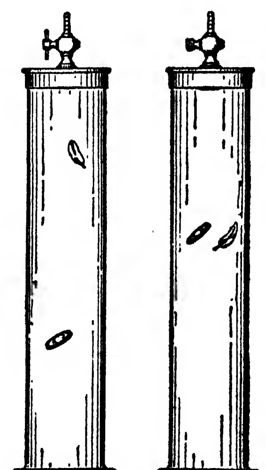


FIG. 5. THE FAMOUS "GUINEA AND FEATHER" EXPERIMENT

Since the air pump was not devised until 1650, sixty years after the time of Galileo's experiments on falling bodies, he did not have an entirely incontrovertible basis for his conviction that the behavior of all falling bodies would be the same in a vacuum. But the assiduity of his experimentation and the penetrating quality of his reasoning therefrom are indicated by his following remark (46:72):

> ... the variation of velocity in air between balls of gold, lead, copper, porphyry, and other heavy materials is so slight that in a fall of one hundred cubits a ball of gold would surely not outstrip one of copper by as much as four fingers. Having observed this, I came to the conclusion that, in a medium totally devoid of all resistance, all bodies would fall with the same velocity.

This statement may have originated in some of Galileo's exploits on the famous leaning tower of Pisa, about which the volume of unsubstantiated tradition far outstrips authenticated history. It seems to have been demonstrated quite conclusively (28) that Galileo never made the sensational public demonstration before the students and faculty of the University of Pisa with which modern writers like to credit him. On the other hand, his references (46:62, 72, 166) to a height of one hundred cubits, almost the exact height of the leaning tower, suggest that he may have used the tower privately in the course of his experiments. Indeed, it is almost inconceivable that he did not do so, his nature and his interests being what they were. But it is a curious fact that in the entire extant collection of the works of Galileo there are only two scant references to the experiments for which he is perhaps most famous.[1]

Notwithstanding the fact that Galileo could not actually observe free fall in a vacuum, we know now that his experiments, wherever he may have made them, gave him a basis for his convictions which was perhaps even better than he realized. In 1917–18, some very careful measurements were made for the United States War Department on the courses that would be followed by bombs dropped from airplanes (36:65). This study made it evident that compact bodies would fall for nearly six hundred feet before the effect of air resistance became large enough to be at all measurable even with the accurate instruments which were used for those observations. Since the leaning tower of Pisa is only 180 feet in height, Galileo had better justification than he was in a position to adduce for his conviction that heavy and light bodies would fall at the same rate in a vacuum.

The similarity in the behavior of all freely falling bodies removes what would otherwise be the necessity for finding how the rate of fall depends on weight or size and opens up the possibility of finding, once and for all, the mode of motion of all falling bodies, whether heavy or light, whether large or small.

The Principal Experimental Difficulty

The hypothesis which Galileo formulated and proposed to test was that "velocity goes on increasing, after departure from rest, in simple proportionality to the time" (46:167). This was found, though somewhat indirectly, to be borne out by experiments, a repetition of which, with improvements appropriate to the times, have a place in all laboratory manuals of physics. But, though improved apparatus will eliminate some of the elements of indirection of Galileo's original experiments, one such element is as insurmountable now as it was then, namely, the impossibility of direct measurement of the velocity of fall. An experimental verification of the assertion that "velocity goes on increasing, after departure from rest, in simple proportionality to the time" naturally involves direct observation of time intervals and velocities, followed by a comparison to test their al-

[1] Florian Cajori, *Science*, *52*, 409, 1920.

leged proportionality. Unfortunately the problem is not quite as simple as that.

In these days of motor cars and airplanes, each equipped with speedometers, it may be somewhat perplexing to be reminded of the difficulties inherent in a direct measurement of velocity. This is possibly not the place to dwell on the fact that velocity is more of an interpretive concept than an object, and correspondingly difficult to measure. Perhaps it will be more convincing to observe that even if speedometers gave direct and accurate measures of velocity, which they do not, to utilize them in the present problem would pose the difficult task of so attaching them to falling bodies as to avoid all interference with freedom of fall. That difficulty seems as insurmountable today as it was three hundred and fifty years ago, and necessitates recourse to an indirect approach to the problem in place of the logically more desirable direct approach.

The Difficulty Surmounted

The difficulty would disappear if, in place of the desired relation between velocity and time, a corresponding relation between distance and time could be substituted; for distance, unlike velocity, is a readily measurable quantity. This Galileo did, stating the ratio in the following form and deducing it from his earlier hypothesis by the geometrical reasoning which was characteristic of his day. Having devised the term "uniformly accelerated" (46:162, 169) to describe the motion specified in his hypothesis, he proposed the following as the required substitute:

> The spaces described by a body falling from rest with a uniformly accelerated motion are to each other as the squares of the time-intervals employed in traversing these distances [46:174].

There are two stages in Galileo's deduction of this theorem. The first is a proof that the distance traveled by a uniformly accelerated body in a given time is the same as that traveled by a body having a uniform speed, the value of which is half that possessed by the former body at the conclusion of the given time interval. This replacement of a changing quantity by an equivalent steady quantity (the concept of *mean value*) is an extremely useful device in scientific thought, and Galileo's use of it at the end of the sixteenth century is one of the evidences that he was far ahead of his time in methods of thought.

Average Velocity in Free Fall

On page 14, the first of equations (1) gave the definition of velocity in algebraic terms as $v = s/t$. This assumes that the velocity is uniform. When the velocity is not uniform, the mean value, usually termed the *average velocity*, may be defined by a similar equation, namely,

$$\bar{v} = \frac{s}{t} \qquad (1)$$

in which the bar over the v now indicates the mean value of a changing quantity. This definition is entirely in line with common practice, as when we say that a car which has traveled four hundred miles in ten hours has averaged forty miles an hour, notwithstanding its having had speeds all the way between zero and perhaps seventy miles an hour.

Galileo was confronted with the fact that the velocity of a falling body was not constant and set himself the problem of determining the value of the average velocity, $\bar{v}$. He used a geometrical method of deducing that for a body falling from rest the average velocity was one half of the final velocity. In his diagram (Fig. 6) vertical distances represent times of fall, measured downward from A. Horizontal distances, measured to the left of AB only as far as the oblique line AE, represent velocities. Since velocities were proportional to times of fall by his main hypothesis ("velocity goes on increasing, after departure from rest, in simple proportionality to the time"), the line AE was straight. The distance BE represented the final velocity of a freely falling body, and FB (and hence GA and IC) half its final velocity. The area of the rectangle $GABF$ being *velocity* times *time* for uniform motion represented total distance moved by a body having half the final speed of a freely falling body. (See $s = vt$, p. 14.) The area of the triangle AEB similarly represented the total distance fallen by the freely falling body. The two areas being equal, the equality of the distances traveled by the two bodies was demonstrated. Thus, Galileo demonstrated that the $\bar{v}$ of equation (1) possessed the value

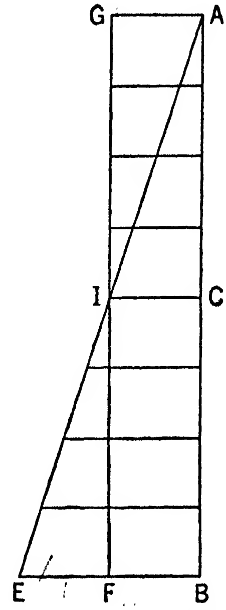


FIG. 6. GALILEO'S CONCEPT OF AVERAGE SPEED (101:1/3)

$$\bar{v} = \frac{v}{2}, \tag{2}$$

where v represented the final velocity of a body falling freely from rest.

But Galileo had already given the name *acceleration* to the rate of change of velocity and hence his main hypothesis (the proportionality of velocity to time in free fall) was expressible.

$$g = \frac{v}{t}, \text{ or } v = gt, \tag{3}$$

where g represents the acceleration of free fall.[1] Substitution of gt for v in equation (2) gives

$$\bar{v} = \frac{gt}{2} \tag{4}$$

[1] The commonly accepted value of g in the English system of units is a change in speed of 32.2 feet per second each second of fall. The common scientific way of stating this is that the ac-

The Distance-Time Relation

The second stage in Galileo's deduction of his hypothesis that distances of fall from rest would be found proportional to the squares of the times of fall, consisted in essence of substituting the value $gt/2$ for $\bar{v}$ from equation (4) into equation (1) and solving for s.

Thus,
$$s = \bar{v}t = \frac{gt}{2}t = \tfrac{1}{2}gt^2. \qquad (5)$$

It was this hypothesis which Galileo subjected to test, thus assuring himself (46:178) "that the acceleration actually experienced by falling bodies is that above described." He did not, however, actually utilize freely falling bodies in the course of this test, but diluted gravity, so to speak, by substituting balls rolling down inclines, after he had satisfied himself that this, no less than free fall, was an example of uniformly accelerated motion.

The Velocity-Distance Relation

Thus were established two laws of falling bodies; that of constant acceleration embodied in equation (3) relating terminal velocity to time of fall and that embodied in equation (5) relating distance to time of fall. A third law relating terminal velocity to distance is

$$v = \sqrt{2\,gs}. \qquad (6)$$

It is not an independent law, being deducible from equations (3) and (5) by eliminating t between them. Though thus deducible from the first two laws, the third law does not seem to have occurred to Galileo. It has subsequently proved to be very useful. Christiaan Huygens (1629–95) is said (74:148) to have been the first to appreciate the significance of this relation.

Even though Galileo did not recognize the substance of equation (6) explicitly, it is important to note that he discovered and utilized what might be termed an important corollary of this relation. In his *Two New Sciences* he said (46:184):

> If a body falls freely along smooth [1] planes inclined at any angle whatsoever, but of the same height, the speeds with which it reaches the bottom are the same.

This is really an amazing discovery for Galileo to have made. Even today it raises grave doubts in the mind of the thoughtful student when he first encounters it. But a careful consideration of equation (3) will show that it at least does not contradict this assertion of Galileo's. The equa-

celeration of gravity is 32.2 feet per second per second. The corresponding value in the metric system is 9.81 meters per second per second. Naturally the value will vary with latitude, with altitude, and, to a lesser degree, with other geographic and geological factors.

[1] By "smooth" Galileo clearly implied "frictionless."

tion states that the final velocity of any body which passes through a certain *vertical* distance is proportional to the square root of that distance. The additional assertion made by Galileo was that the simultaneous *horizontal* distance of travel, necessitated by the motion's being along an inclined plane, had no effect on the final speed. This statement is significant as a prelude to some major investigations undertaken by Galileo himself and later by Huygens.

Equations (3), (5), and (6) describe the motion of falling bodies which start from rest. There is frequent occasion for describing the motion of falling bodies which are not at rest when the observations begin. Using a zero subscript to specify the initial vertical velocity, a repetition of the line of reasoning which yielded equations (3), (5), and (6) will give the following:

$$v = v_0 + gt \tag{7}$$

$$s = v_0t + \frac{gt^2}{2} \tag{8}$$

$$v = \sqrt{v_0^2 + 2gs} \tag{9}$$

In applying equations (7) to (9) to specific numerical cases, attention must be given to sign. The directions of v, s, and v_0 may be either upward or downward. The acceleration of gravity, g, will of course always be downward. The simplest way to deal with these is to adopt a convention, associating positive signs with the numerical values of downward quantities, and negative signs with numerical values of upward quantities. The same convention could have been applied also to numerical values substituted in equations (3), (5), and (6). It was scarcely worth while to mention it, however, since for freely falling bodies starting from rest, all quantities have the same sign, only downward directions being involved.

Principles of Motion of Projectiles

A problem which is obviously associated with free fall is that of the motion of projectiles. In view of Galileo's clarification of the problem of falling bodies, it will not create any surprise to note that he made a similar contribution toward the solution of that of the motion of projectiles. Considering the very prevalent impression even today that the high speed of a projectile somehow exempts it from the downward acceleration which other objects experience when in midair, one can appreciate the necessity which must have existed in Galileo's time for a pioneer study of the problem.

Fundamentally, what Galileo did was simply to pursue the implications of his hypothesis on free fall, especially the second form, the proportionality of distance to the square of the time. Naturally, this applied only to the vertical portion of the motion of the projectile. To the horizontal motion he applied a principle which he had formulated in connection with his experiments in rolling balls down inclines. It was quite revolutionary and, indeed, entirely subversive of the doctrine which had theretofore received unanimous assent.

This principle was that "the horizontal motion remains uniform" (46:250) except as influenced by friction or air resistance. Galileo had been led to this conclusion by noting that the retardation of balls rolling up an inclined trough became less as the slope was rendered less steep until, when the trough became horizontal, the retardation seemed to him small enough to be accounted for entirely by friction. Hence, he concluded that if friction could be eliminated, the horizontal motion of bodies would continue indefinitely, unabated.

Hitherto the accepted doctrine had been that the most natural state was one of rest and that the maintenance of any motion whatever required the continued exercise of force. It was supposed that, entirely aside from such retarding agents as friction, any motion, whether horizontal or otherwise, would necessarily cease with the cessation of the force which caused it. The resulting view of projectile motion was, of course, quite naïve. Perhaps the best summary of it is to be found in a work by Santbeck written in 1561.[1] He represents the course of a projectile as in Figure 7. The force on the projectile was somehow conceived as continuing during the first or "violent" part of its flight, resulting in a straight-line course. When the force ceased — for reasons as obscure as the reasons for its continuance up to that point — "natural" motion of vertical fall ensued. Let any modern who is inclined to ridicule this concept inquire honestly and insistently into his own or his friends' ideas on bullet flight and then make comparison! Santbeck was at least consistent in his application of prevailing scientific doctrine, which is more than can be said of some of his contemporaries. Figure 8 shows how other writers, uneasy about the angle at the top of this type of trajectory, modified it. They theorized that projectile motion was compounded of three parts: (1) *modus violentus*, (2) *modus mixtus*, and (3) *modus naturalis*. But as early as 1537, Nicholas Tartaglia, "the father of ballistics," had shown the trajectory to be curved throughout and that the maximum range in a vacuum existed for an angle of forty-five degrees.

[1] E. N. da C. Andrade, "Science in the Seventeenth Century," *Proceedings, Royal Institution of Great Britain, 30,* 212 (1938).

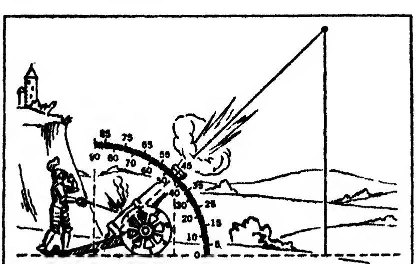


FIG. 7. SANTBECK'S CONCEPTION OF THE PATH OF A PROJECTILE (1561)

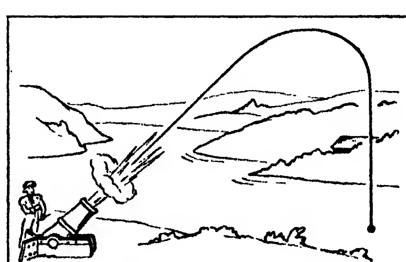


FIG. 8. A MODIFICATION OF SANTBECK'S THEORY OF PROJECTILE MOTION

Galileo and the Trajectory in Vacuo

It was in 1637, exactly a century after Tartaglia's work, that Galileo put the theory of the flight of the projectile on a firm foundation, at least as far as flight in vacuum was concerned. He simply disregarded the earlier incorrect or inadequate treatments, substituting his own doctrine: that of the continuance of motion unchanged in the absence of a retarding force. Throughout his entire treatment of the motion of projectiles this principle appears time after time. It is the essence of what we now know as Newton's first law of motion. Unfortunately Galileo failed to envisage all its implications. It apparently did not occur to him to extend the principle beyond the particular case of the motion of projectiles. That generalization was made by Newton, who, by a highly symbolic coincidence, was born the year that Galileo died, 1642. Hence, Galileo's principle that the horizontal portion of the motion of a projectile remains uniform possessed an even greater degree of validity than Galileo realized.

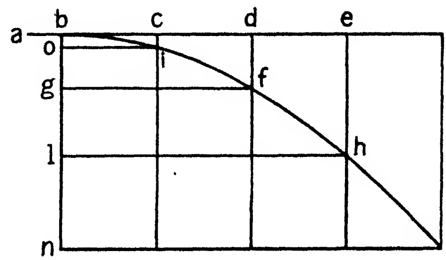


FIG. 9. GALILEO'S DIAGRAM OF THE FORM OF A TRAJECTORY (101:249)

To facilitate the application of his two principles, one on the vertical portion of the motion of a projectile and the other on the horizontal portion, Galileo proceeded as follows (46:248–49):

> Let us imagine an elevated horizontal line or plane *ab* along which a body moves with uniform speed from *a* to *b*. Suppose this plane to end abruptly at *b*; then at this point the body will, on account of its weight, acquire also a natural motion downwards along the perpendicular *bn*. Draw the line *be* along the plane *ba* to represent the flow or measure of time; divide this line into a number of segments, *bc*, *cd*, *de*, representing equal intervals of time; from the points *b*, *c*, *d*, *e*, let fall lines which are parallel to the perpendicular *bn*. On the first of these lay off any distance *ci*, on the second a distance four times as long, *df*; on the third, one nine times as long, *eh*; and so on in proportion to the squares of *cb*, *db*, *eb*, or, we may say, in the squared ratio of these same lines. Accordingly we see that while the body moves from *b* to *c* with a uniform speed, it also falls perpendicularly through the distance *ci*, and at the end of the time-interval *bc* finds itself at the point *i*. In like manner, at the end of the time-interval *bd*, which is the double of *bc*, the vertical fall will be four times the first distance *ci*; for it has been shown in a previous discussion that the distance traversed by a freely falling body varies as the square of the time; in like manner the space *eh* traversed during the time *be* will be nine times *ci*; thus it is evident that the distances *eh*, *df*, *ci* will be to one another as the squares of the lines *be*, *bd*, *bc*.

Having previously shown that points whose vertical and horizontal displacements were in a quadratic ratio lay on a curve termed a parabola, Galileo concluded that the path of a projectile bore the form of a parabola. This would be true for any initial velocity and also, as later demonstrated, for any angle of projection. This not only identified the form of the trajectory described by a projectile, but also, what was more important, rendered possible the application of the large body of mathematical knowledge about the parabola, which was available even in Galileo's time, to the study of such trajectories. Naturally, this applied only to the ideal case of motion in a vacuum, but even so, it was of considerable practical importance.

Elements of the Trajectory

In contemporary practice, a shorter and more informing deduction of the path of a projectile would be as follows. Take equation (8) as an expression of the distance-is-as-the-square-of-the-time theorem on vertical motion under gravity, and write it in the form

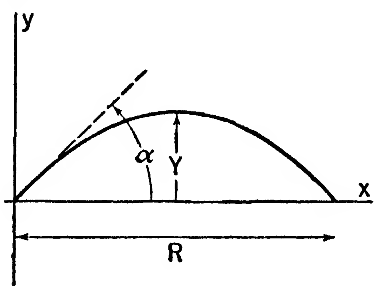


Fig. 10. Elements of a Trajectory in Vacuo

$$y = v_y t + \tfrac{1}{2} g t^2 \qquad (10)$$

The term y represents the distance of a projectile above or below the point from which it was fired (taken as the origin) after the lapse of t seconds, and v_y is the vertical component of the initial velocity. In substituting numerical values it will be well to adopt the more common sign convention for equation (10), and those that follow; namely, that the positive sign indicates upwardly directed quantities and the negative downwardly. Thus, g will be -32.2, etc.

Now express Galileo's constant-speed theorem for the horizontal part of the motion in the equation

$$x = v_x t, \qquad (11)$$

the notation being analogous to that for equation (10). Solving equation (11) for t and substituting in equation (10), there results

$$y = \frac{v_y}{v_x} x + \frac{g}{2v_x^2} x^2. \qquad (12)$$

This equation will be recognized as that of a parabola, as Galileo discovered.

If the angle of inclination of the gun with the horizontal is θ,[1] then v_y/v_x has the value $\tan\theta$. If the initial velocity of the projectile is V, v_x has the value $V\cos\theta$. Substitution in equation (12) gives

$$y = x \tan\theta + \frac{gx^2}{2V^2\cos^2\theta}. \qquad (13)$$

[1] Greek letters are very commonly used in physics, especially to designate angles. This letter is *theta* (pronounced "thayta"). The entire Greek alphabet, with the names of the letters, will be found in the Appendix.

Several properties of the trajectory are deducible from equations (10), (11), and (13). It will be at once evident that, given the value of g, V, and θ, the coordinates of the projectile may be established for any given number of seconds after it starts on its flight, through equations (10) and (11). The height of the projectile may, through equation (13), be found for any given horizontal distance of travel. The converse is also true, though the calculation is not so simple.

The *range* R (Fig. 10) is

$$R = \frac{2\,V^2 \sin\theta \cos\theta}{g} = \frac{V^2 \sin 2\,\theta}{g}. \tag{14}$$

Equation (14) results from substituting $y = 0$, $x = R$ in equation (13) and solving for R, taking g as positive in this and the following equations.

The *maximum ordinate* Y is

$$Y = \frac{V^2 \sin^2\theta}{2\,g}. \tag{15}$$

Equation (15) results from substituting $x = R/2$ in equation (13), using the first value of R given by equation (14).

The *time of flight*, T, is

$$T = \frac{R}{V \cos\theta} \tag{16}$$

or

$$T = \frac{2\,V \sin\theta}{g}. \tag{17}$$

FIG. 11. THE ARTILLERY "DIRECTOR" AS A MODERN APPLICATION OF THE THEORY OF PROJECTILE MOTION

(Courtesy of Sperry Gyroscope Company.)

Equation (16) is deducible from equation (11) by appropriate substitutions, and equation (17) follows when the value of R is substituted in (16).

The fact that the maximum range occurs when $\theta = 45°$ follows from the fact that in equation (14) the maximum R results when the value of $\sin 2\theta$ has its greatest value. The largest possible value for the sine of an angle is unity, which occurs when the angle is 90°. Hence the maximum value of R occurs when $2\theta = 90°$; that is, when $\theta = 45°$.

In deducing the foregoing and many other properties of the motion of projectiles *in vacuo*, Galileo used the cumbersome geometrical methods of his day instead of the more concise and informing algebraic treatment which was developed later. It was only because he possessed one of the most penetrating minds of all time that he was able to accomplish what he did with the crude intellectual tools at his disposal. Since Galileo's time the mathematical treatment of the flight of a projectile has become an extensive branch of applied mathematics called *ballistics*. For the practical purposes of artillery fire it has become an exact science. The problem having been solved mathematically, the next step (natural in the machine age) has been taken. Computing machines called "directors," electrically attached to the guns, automatically keep them trained on the target and elevated at the necessary angle, making all corrections such as those for difference in height of gun and target, motion of target, and windage. All that is necessary is for the observer to keep the telescope of the director trained on the target and to make the settings corresponding to the corrections. Figure 11 illustrates such a director in use.

The requirements of anti-aircraft fire and of bombing from airplanes are, of course, far more exacting than for ordinary artillery fire. Mechanical directors have been devised for both these purposes. Those made in the United States are reported to be greatly superior to any others. Their construction is a closely guarded secret.

Questions for Self-Examination

1. Why was the discovery of the laws of free fall an especially significant episode in the history of physics?
2. Why did Galileo not subject to experimental test his first hypothesis about falling bodies, namely, that "velocity goes on increasing in simple proportionality to the time"?
3. What hypothesis did Galileo adopt for an experimental test of the behavior of falling bodies and what experiment did he perform to verify it?
4. What two principles guided Galileo in his studies on projectile motion? What form did he deduce for the path of a projectile?
5. From the measurements on a falling body the speed (v) and the total distance (s) at the end of successive seconds (t) is found to be:

t	0	1	2	3
v	0	32	64	96
s	0	16	64	144

Calling 32 by the letter g shows that all entries in the v line are describable by the relation $v = gt$ and all entries in the s line by the relation $s = \frac{1}{2} gt^2$.

6. What was Galileo's main contribution to the knowledge of motion of projectiles?
7. Perform the operation leading to equation (12).
8. Perform the operation leading to equations (14), (15), and (16).

Problems on Chapter III

Remarks on Problems

These problems, and those pertaining to following chapters, are intended to illustrate the principles developed in the text and to help clarify them in a way which only a treatment of numerical cases will do. If you handle them thoughtfully, they will make a real contribution to your grasp of the subject. Though the physical principles upon which solutions are to be based will be found in the text, very seldom will a "formula" be found that can be used directly. Ability to substitute values in a given formula has been termed the lowest order of mathematical intelligence. Whatever its mathematical level may be, it can scarcely be said to constitute physics at all. Look for the basic principle, not for a formula.

The fact that many of the problems are stated in algebraic terms furnishes you, indeed, with an opportunity to develop your own formulas. If you are capable of doing this (and your instructor will probably prefer that you do it instead of the numerical part if you can), you will not only get a better grasp of the principles, but will be freed of a considerable amount of detailed computation. But you can do it only if your grasp of physics is somewhat better than the average and if your high-school algebra is in well-oiled working condition. A little trigonometry is involved at some points, but not more than can be furnished in a single class hour.

Numerical answers are given, but only to two significant figures. Your computations should be correct to three figures if done by slide rule. The data are presented with sufficient precision to warrant carrying the computations to four significant figures if desired. The approximate answers are given to help you to decide whether you are "on the right track."

Try to think a problem through before undertaking the actual solution of it. Unless you can keep a sense of perspective as you go, you will lose most of the value of the labor that is involved. *Give yourself time.* Haste is fatal to maintenance of perspective. When you feel as if you are groping in a fog, not able to see the desired destination or the path leading to it, you are wasting your time. Put away your work and return to it when your head is clearer or after you have received suggestions from some competent source.

Problems

In problems involving the acceleration of gravity, assume the value 32.2 ft/sec². For problems with tabular values, the table bearing the problem number, as 6 and 7 below, accompanies the problem correspondingly numbered.

1. An automobile increases its speed from 20 miles an hour to 50 in 15 seconds. What is the average acceleration in feet per second per second (ft/sec²)? 3 ft/sec².
2. What distance does the automobile of the preceding problem travel during the given interval, assuming the acceleration to be constant? 770 ft.
3. An automobile traveling 50 miles an hour runs head-on into a wall. Are the chances of survival of its passenger greater or less than those of a man who falls from the top of a five-story building (70 feet)?
4. A test pilot making a vertical power dive at 500 miles an hour stops his descent in

4.5 seconds. If the (negative) acceleration is uniform, how is it related to that of a freely falling body, both in direction and in magnitude? 5*g*.

5. A stone thrown horizontally from the top of the Empire State Building takes 8.64 seconds to reach the street. How high is the building? 1200 ft.

6. A stone thrown down from the edge of an overhanging cliff with an initial speed of v feet per second, strikes the ground in t seconds. What is the height h of the cliff in feet?

	v	t	h		v	t	h
6.	30	4	380	7.	30	4	140
	40	5	600		40	5	200
	50	6	880		50	6	280
	60	7	1200		60	7	370

7. Solve the preceding problem on the assumption that the stone is thrown vertically upward, instead of down.

8. A stone is launched downward at an angle of α degrees with the horizontal and with an initial speed of v feet per second, from the top of a cliff h feet high. How many seconds t elapse before it strikes the ground? How many feet d from the foot of the cliff does it strike?

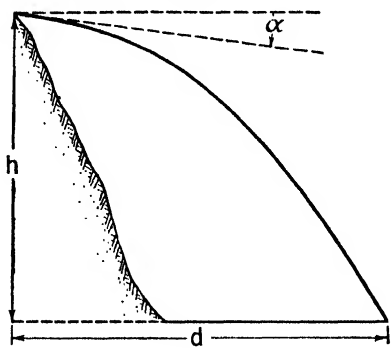


FIG. 12

v	α	h	t	d
60	30°	400	4.1	220
60	40°	400	3.9	180
60	50°	400	3.8	150
60	60°	400	3.6	110

9. Solve the preceding problem on the assumption that the stone is thrown upward at an angle α with the horizontal, instead of downward.

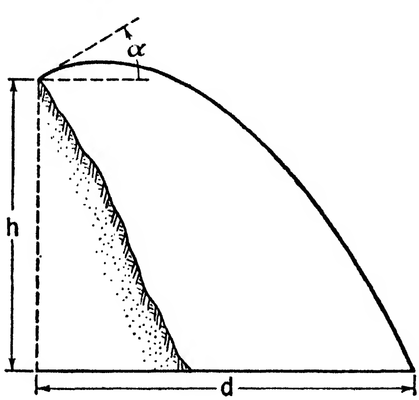


FIG. 13

v	α	h	t	d
60	30°	400	6.	310
60	40°	400	6.3	290
60	50°	400	6.6	260
60	60°	400	6.9	210

10. Two bombs are dropped simultaneously from an airplane. One strikes the edge of an h-foot cliff. The other, just clearing the edge, strikes at the bottom t seconds later. How many feet H was the airplane above the top of the cliff?

	h	t	H
10.	1000	1	15050
	1000	2	3403
	1000	3	1264
	1000	4	536.0

	v	α	R	Y	T
11.	2250	10°	10.20	2400	24
	2250	15°	14.91	5300	36
	2250	20°	19.16	9200	48
	2250	25°	22.84	14000	59

11. A bullet having a velocity of v feet per second is discharged at an angle α above the horizontal. What is its range R in miles, its maximum distance Y in feet above the point of discharge, and its time of flight T in seconds, neglecting air resistance?

12. What is the maximum range R in miles, in a vacuum, of a howitzer which has a muzzle velocity of v feet per second?

	v	R
12.	400	.94
	800	3.8
	1200	8.5
	60000	210.

	R	α	h
13.	1500	0° 11.	1.2
	4500	0° 34.	11.
	9000	1° 8.	45.
	15000	1° 54.	120.

13. The bullet from an army rifle has a muzzle velocity of 2700 feet per second. At what angle α with the horizontal must the rifle be inclined for a range of R feet in a vacuum? How many feet h would the bullet rise and fall?

14. Find the angle of departure θ, the maximum ordinate Y in feet, and the time of flight T in seconds for the trajectory *in vacuo* when the given range is R *feet*, the initial velocity being 2250 ft/sec.

	R	θ	Y	T
14.	15,000	2° 45′	170	6.7
	30,000	5° 30′	720	13.
	45,000	8° 15′	1600	20.
	60,000	11° 15′	2900	27.

	x	y	t
15.	10000	1400	4.5
	18000	2100	8.1
	27000	2400	12.
	40000	1800	18.

15. Find the height y of the projectile in feet, and the lapse of time t in seconds after firing, for each value of x when the angle of departure is 10° and the initial velocity is 2250 ft/sec.

Equilibrium; Resolution of Vectors

The Roots of the Science of Mechanics

Most early machines had for their object a more advantageous application of muscular effort than was possible without them. Their operation called into play complicated combinations of forces, large and small and in varying directions. Their inventors must have encountered the analysis and combination of such forces in manifold practical form. It seems almost inconceivable that the general problem of the mathematical treatment of these forces should not have received attention much earlier than it did. Yet the first recorded efforts in this direction, those of the ancient Greeks, notably Archimedes (74:510; 25:6), were neglected and lost to the Western world for more than a millennium and a half. One would suppose that the Romans, an enterprising and intensely practical people, would have realized the value of such studies. Yet it is probable that in those days, as in these, the discoveries which were the most valuable, judged even from a limited utilitarian standpoint, were seldom made by the "practical" man.

Whatever the reason, the problem of force combinations was not solved in its full generality until the seventeenth century. It must have been "in the air," as were so many other fundamental scientific problems at that time, for the first solution was announced by three different men in the same year, 1687, in apparent independence of one another. These three were Isaac Newton (1642–1726), of whom we shall hear much, Pierre Varignon (1654–1722), and Bernard Lamy (1645–1716) (74:36).

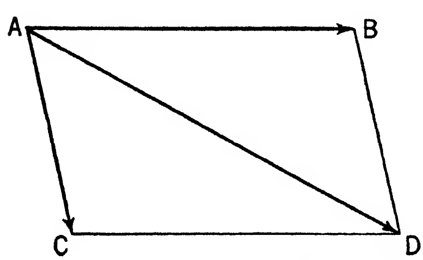


FIG. 14. NEWTON'S ILLUSTRATION OF COMPOSITION AND RESOLUTION

Vectors and Scalars

Figure 14 is the illustration which accompanied Newton's introduction of the idea of combinations of forces. Perhaps the most significant point is his use of arrows to represent the forces. This is such a common expedient today that one is likely to miss the full significance of its early use. Forces constitute examples of an important class of physical quantities that have both direction and magnitude. These are commonly termed *vector quantities*, a term originated by the Irish mathematician, W. R. Hamilton, in the middle of the nineteenth century. An arrow is a natural device for representing vector quantities and when so used, the term *vector* (as a noun) is applied to it. It clearly indicates direction and its length may be made to represent magnitude. Thus, a vector may not only represent a force but a velocity (which obviously has direction as well as magnitude), displacement, momentum, and, equally useful though less obvious at this stage, several electrical entities. Vectors are thus tools for the treatment of some of the principal concepts encountered in physics.

In contrast to vector quantities there are *scalar* quantities (a term also due to Hamilton), those which possess magnitude but no direction. An example which will come to mind immediately is volume. Such entities require only one number to describe them. Moreover, when two such entities are combined, the result is merely the arithmetical sum of the two. Two weights of four and three pounds respectively add to seven pounds. This is not necessarily true of vectors. A man walking three miles an hour across the width of a ship which is moving at four miles an hour forward would be moving obliquely with reference to the surface of the earth at a rate of five miles an hour. Of course, the *resultant*, as it is called, of two vectors of magnitude four and three is not necessarily five. It is five only if, as in this case, the two components are at right angles. Its value would be different from five for any other angle between the two vectors being compounded. For example, if the man walked toward the bow of the ship, the resultant — that is, his actual motion with respect to the surface of the earth — would be seven miles an hour, the arithmetical sum of the components. And if he walked toward the stern, his resultant velocity would be one mile an hour. Methods of combining vectors and of computing their resultants will be illustrated and used in Chapter 5. For the present, a grasp of the twofold nature of a vector and of the distinction between vectors and scalars will suffice.

Resolution of Forces

There are two aspects to the treatment of forces which are applied at one point. One is the breaking-down or *resolution* of a given force into two or more *components* which, acting jointly, constitute the equivalent of the given single force. The other is the combination or *composition* of two or more forces into a single force called the *resultant* which is the equivalent of the given individual forces. The first aspect, resolution, may be illustrated by a simple example.

When a sled is being drawn along level ground, the rope by means of which the force is being exerted upon it is usually at an oblique angle. The consequence is that not all of the force transmitted through it to the sled is effective in producing horizontal motion. The smaller the angle that the rope makes with the ground, the larger is the useful component of the force until, when the rope is actually horizontal, the maximum effect is produced. The larger the angle that the rope makes with the ground, the smaller the useful component until at the vertical none of the force acts horizontally. All of it acts toward balancing the weight of the sled. Suppose that the magnitude of the force transmitted by the rope were known, as might easily be effected by means of some such device as a spring balance. Suppose the balance indicated a force of magnitude of ten units which, for present convenience, will be called ten pounds.[1] It then appears natural to ask, since the magnitude of the horizontal component is related to the obliquity of the rope, what is the relation between the two. Knowing the oblique force and the angle with the horizontal, how can the horizontal component be computed?

Representing the ten-pound oblique force by a vector (Fig. 15), which is the diagonal of a parallelogram (in this instance the parallelogram becomes a rectangle, since the components are at right angles with each other), the horizontal component will be represented simply by either of the horizontal sides of this rectangle. The problem is solved in principle. All that remains is to make a careful drawing to scale, and by measuring the length of the horizontal vector to deduce the magnitude of the useful component of the force. If the angle were 37°, for example (chosen because it yields the convenient 3–4–5 triangle), the magnitude of the horizontal component would be found to be 8 pounds. From the same figure, the vertical component could be measured as 6 pounds. Hence, with an oblique force of 10 pounds applied at the given angle, 8 would be effective in a horizontal direction. Six pounds would act toward balancing the weight of the sled; that is, if the sled should be drawn across a platform balance, its apparent weight would be six pounds less than would be registered if the given force were not acting upon it.

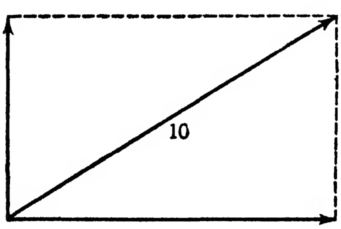


FIG. 15. RESOLUTION OF A FORCE INTO TWO RECTANGULAR COMPONENTS BY GRAPHICAL METHOD

A Better Method of Resolving Forces

But the graphical method of finding the components of a given vector, though it embodies the principle, is of very little use except in the hands of a skilled draughtsman. A method which is free from this objection and which is far more accurate than the graphical method, even at its best, can be devised by constructing the geometrical equivalent, taking the foregoing graphical method as the guide. The procedure would then be as follows:

[1] The potential confusion involved in regarding the pound or the kilogram as a unit of force will be discussed in Chapter 9. It need create no difficulty at this point.

Construct on the vector representing the given force as a hypotenuse a right triangle whose sides are parallel to the two rectangular components which it is desired to know. Identify the side corresponding to the desired component. The value of the component will then be the product of the magnitude of the original force by the ratio of the length of that side to the length of the hypotenuse. In the case above, for example, the horizontal component is

$$x = 10 \times \tfrac{4}{5} = 8,$$

and the vertical component is

$$y = 10 \times \tfrac{3}{5} = 6.$$

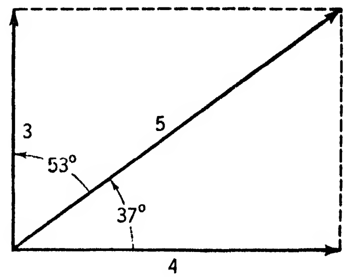


FIG. 16. RESOLUTION OF A FORCE BY GEOMETRICAL METHOD

This method is perfectly general and applies to all cases by substituting for the fractions $\frac{4}{5}$ and $\frac{3}{5}$ the corresponding ratios of the sides of whatever triangles may be involved in any particular case. It is important to observe that the ratio always involves the same two sides: the hypotenuse as the denominator of the fraction and the side, which together with the hypotenuse fixes the angle, as the numerator. This ratio will be recognized as the *cosine* of the angle between the side and hypotenuse. Complete trigonometric tables including a table of cosines will be found in the Appendix. The resolution of vectors is greatly facilitated by the use of such tables.

The General Procedure in Resolving Forces

A rule for resolution may now be stated:

Given a vector, to find the component at a desired angle with it, multiply the magnitude of the vector by the cosine of the angle between the vector and the direction of the desired component.

This method, though apparently so simple, seems to present considerable difficulty when the beginner encounters it for the first time. The reader can perhaps best lay a foundation for a ready comprehension of the following steps if he will carry through several computations similar to the above, using different values for the forces and angles. To reduce arithmetical labor, the angle may be chosen in each case so as to give a right triangle with integral sides and the magnitude of the force taken commensurate with the hypotenuse. The 3–4–5 case may be utilized further by taking the complementary angle, 53°. Similarly, the approximate angles of 23° or 67° are associated with a 5–12–13 triangle and of 28° or 62° with an 8–15–17 triangle.

This method of resolution of vectors is of immense utility. Its power will become especially evident in dealing with the present problem of forces in equilibrium. The combination of the method of resolution with the principle of equilibrium is possibly more powerful than any other tool in the whole of mechanics; certainly than any other in that part of mechanics, known as *statics*, at present under survey.

The principle of equilibrium requires statement in a form in which it can be applied to the problems of statics. ***A body is said to be in static equilibrium when its velocity does not change.*** This is equally true of an automobile traveling sixty miles an hour on a perfectly smooth and straight road and of any stationary structure such as a bridge. Since no change is occurring in the velocity of either, the principle of static equilibrium applies equally to both. However, in practice, the principle is more readily applicable to stationary than to moving bodies, partly because it is not easy to measure the forces, such as friction and air resistance, which act on moving bodies. For present purposes, the most useful special case will be that of a body at rest under the mutually neutralizing effects of two or more forces. It should be kept in mind, however, that this is a very restricted special case, to which the general principle of equilibrium is by no means limited.

Utilizing the Principle of Equilibrium

Apply this restricted principle to the case of a three-cornered tug-of-war. Two forces, of magnitude 200 and 100 (Fig. 17), make with each other an angle of 60°. Required, the magnitude (F) and direction (α) of the single force which will just balance these two; that is, which will produce a state of equilibrium. Since, by hypothesis, the junction point is stationary under the balanced action of these forces, one of the directions in which it is *not moving* is the horizontal. Hence, it follows that, among others, the algebraic sum of the horizontal components of the three forces must be zero. These three components have the following values:

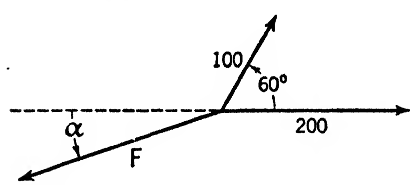


FIG. 17. THREE FORCES IN STATIC EQUILIBRIUM

for the *200* force,	$200 \cos 0° = +200;$	
for the *100* force,	$100 \cos 60° = +50;$	(1)
for the F force,	$-F \cos \alpha$	

The signs indicate direction, right or left: + for the right and − for the left. The assumption of equilibrium requires the algebraic sum of these quantities to have the value zero. Therefore,

$$+200 + 50 - F \cos \alpha = 0. \qquad \text{or } F \cos \alpha = 250. \tag{2}$$

This does not yield the desired information, however, namely, the values of F and α. One of the first principles of algebra is that to find two independent unknown quantities, two independent equations are required. Here we have only one equation; another must be sought.

The first equation was educed by considering the horizontal components. The second may be similarly educed by considering the vertical components. These are:

$$\left.\begin{array}{ll}\text{for the } \mathit{200} \text{ force,} & 200\cos 90^\circ = 0; \\ \text{for the } \mathit{100} \text{ force,} & 100\cos 30^\circ = 86.6; \\ \text{for the } F \text{ force,} & -F\cos(90-\alpha),\end{array}\right\} \quad (3)$$

where the signs indicate direction, up or down: + for up, − for down.

Simplifying the Equilibrium Equations

The expression cos $(90 - \alpha)$ requires comment. It represents the ratio a/c in Figure 18. As related to α, it is the side ***opposite*** α divided by the hypotenuse. (The cosine was the side *adjacent* to α divided by the hypotenuse.) This ratio will be recognized as the *sine* of α. Hence, for the term *cos* $(90 - \alpha)$ the equivalent term *sin* α may be substituted. Like cosines, values of the sines of angles have been tabulated. A complete table will be found in the Appendix.

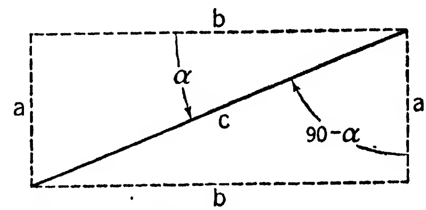


FIG. 18. GEOMETRICAL RELATIONS OF EQUILIBRANT OF FIG. 17

Making this substitution in the statement of the values of the vertical components

$$\left.\begin{array}{l}\text{for the } \mathit{200} \text{ force, } 200\cos 90^\circ = 0 \\ \text{for the } \mathit{100} \text{ force, } 100\cos 30^\circ = 86.6 \\ \text{for the } F \text{ force, } \quad -F\cos(90-\alpha) = -F\sin\alpha\end{array}\right\} \quad \text{from} \quad (3)$$

The assumption of equilibrium requires the algebraic sum of these quantities to have the value zero.

Therefore, $0 + 86.6 - F\sin\alpha = 0$ or $F\sin\alpha = 86.6$. (4)

Equations (2) and (4) may now be solved:

$$F\cos\alpha = 250 \qquad \text{from} \quad (2)$$
$$F\sin\alpha = 86.6 \qquad \text{from} \quad (4)$$

From the fact, evident in Figure 18, that

$$a^2 + b^2 = c^2, \text{ whence } \left(\frac{a}{c}\right)^2 + \left(\frac{b}{c}\right)^2 = 1, \text{ whence } \sin^2\alpha + \cos^2\alpha = 1,$$

it will be evident that these equations can be solved for F by squaring and adding; whence,

$$\begin{array}{ll}F^2(\sin^2\alpha + \cos^2\alpha) & = \overline{86.6^2} + \overline{250^2}; \\ \text{whence } F^2 & = 70{,}003^-; \\ \text{whence } F & = 265^-.\end{array}$$

The value of α may now be found by substituting this value of F into either equation (2) or equation (4), solving for cos α or sin α, and then using the table to find α, which will be found to be 19° 6′ approximately.

Hence, the two given forces will be equilibrated by a single force of magnitude slightly less than 265, directed down and to the left at an angle of 19° 6′ with the horizontal. Thus the method of resolution of forces in conjunction with a special case of the principle of equilibrium has made it possible to compute the magnitude and direction of the force required to balance two given forces. It can be applied with equal facility to any number of forces in one plane meeting in a point. Stated in its full generality, ***translational equilibrium of concurrent forces is realized when the vector sum of all the forces is zero.***

Application to a Simple Bridge Truss

This is a powerful principle. It will repay further attention, the more so in that it may appear somewhat elusive to the beginner. The same procedure, applied successively at different joints, may be made to yield information on the stresses in a bridge truss such as that shown in Figure 19. Suppose, to simplify the calculations, that the lengths of the members of the truss are proportioned in the ratios 3–4–5. The angles are then known to be approximately 37° and 53°. Disregard the weight of the truss itself and suppose a load of 80 tons to be resting at the middle. This is, of course, distributed equally between the two end supports, each end of the truss, therefore, bearing down on its respective support with a force of 40 tons. The principle of equilibrium then requires, if the bridge is not to collapse, that the supports themselves shall push up on each end of the truss with a force of 40 tons. It is this aspect of the situation, *the forces acting from outside* on the body under consideration, which is the working material from which the equations are to be constructed. Confine attention now to one end of the truss, say the left. The two members meeting at that point can exert forces only along their length (a statement which applies equally to all other points where bridge members meet). Represent these forces by vectors and include as a third vector the upward thrust of the support at that point. The vector diagram will then be as in Figure 20.

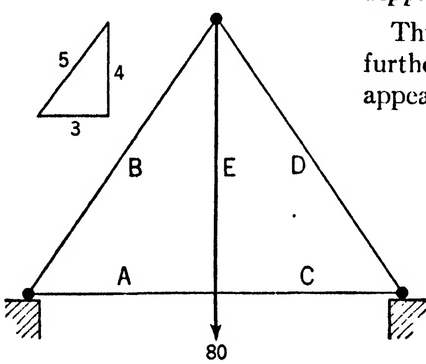


FIG. 19. A "KING-POST" TRUSS WITH LOAD AT CENTER

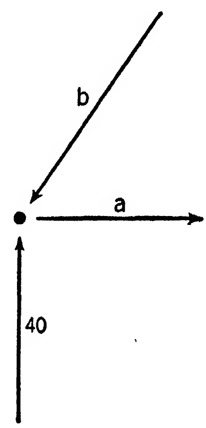


FIG. 20. VECTORIAL ANALYSIS OF FORCES ON LEFT PIN OF THE TRUSS OF FIG. 19

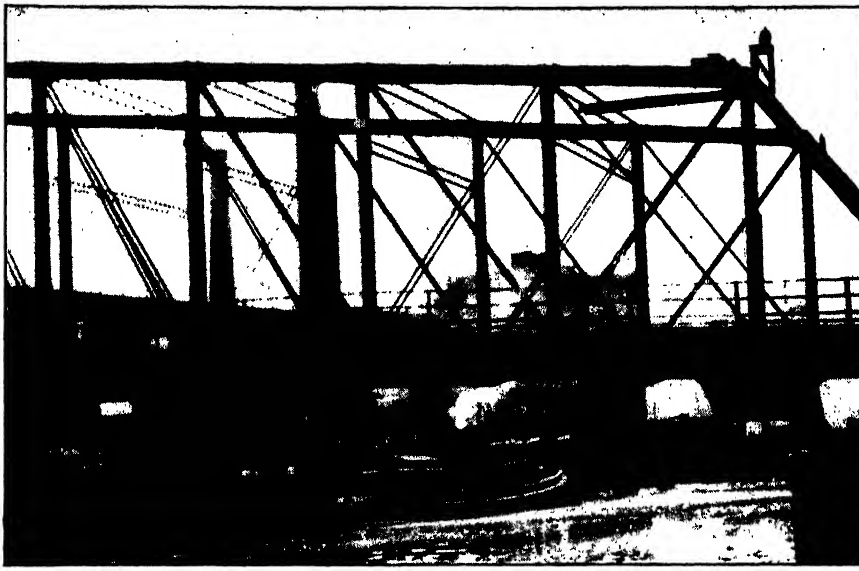


FIG. 21. PORTION OF SIMPLE TRUSS BRIDGE, SHOWING DIFFERENT TYPES OF MEMBER

A word about the directions in which these vectors point may be in order. The 80-ton load on the bridge produces a stretch or tension in the horizontal members *A* and *C* (Fig. 19). To appreciate this, imagine two sticks, hinged at the top, to be supported at the bottom on slippery plane surfaces. It will be evident that a downward force on the hinged joint will cause the two lower ends to slide outward, away from each other. They could be held together by a cord, which would then be under a tension. This is precisely the case of the bridge truss under consideration. The horizontal member *A* (Fig. 19), being under tension, is therefore exerting on the left pin, from outside, a force to the right, and it is this force that is represented by the vector *a* in Figure 20. Similarly, the oblique member *B* will be seen to be under a push from both ends, is hence under compression, and is hence acting down and to the left on the left pin, from outside. This is represented by vector *b* in Figure 20.

Since the bridge is stationary, the forces acting at all its points must be in equilibrium, which is true for those acting at the lower left pin as well as elsewhere. Hence the equilibrium condition (algebraic sum of forces must equal zero) may be applied to the two components, horizontal and vertical. For the vectors of Figure 20,

$$a - b \cos 53° = 0,$$
$$40 - b \cos 37° = 0.$$

Substituting the given values of the sine and cosine for the angles of this truss (see Fig. 19),

$$a - b\,\tfrac{3}{5} = 0,$$
$$40 - b\,\tfrac{4}{5} = 0;$$

from which the values of a and b will be found to be

$$a = 30, b = 50.$$

Hence a load of 80 tons applied at the middle produces a tension of 30 tons in the left horizontal member and a compression of 50 tons in the left oblique member. These are also the stresses in the corresponding members in the right half, by reason of the symmetry of structure and loading. The stress in the middle vertical member is easily seen to be 80 tons, which completes the analysis. For more complicated structures and asymmetrical loading more labor is involved in carrying out the computations, but the procedure is exactly the same.

This kind of calculation is involved in the design of every bridge, since the type and strength of the various members must be arranged to support the maximum load expected. Steel cables or strips will support tensions, but stiff struts are required to withstand compressions. These two types of bridge member may be seen in the photograph of Figure 21. The maximum forces which are assumed at each joint must include the weight of the parts of the bridge itself, which, in the larger structures, are necessarily heavier than the loads to be carried. But whether large or small, the stresses in each member are calculated in advance by applying the principle of static equilibrium to each joint of the bridge in turn, in the course of which the chief mathematical tool is the method of resolution of forces.

Questions for Self-Examination

1. Describe the nature of the first solution (at the hands of Isaac Newton) of the problem of force combinations.
2. Analyze the case of the sled drawn with an oblique rope, as an example of resolution of forces.
3. State the principle of equilibrium of forces and apply it either to a three-cornered tug-of-war which you will formulate for the purpose or to a triangular bridge truss.
4. Describe, in a numbered sequence, the steps involved in computing stresses in a bridge truss.

Problems on Chapter 4

1. Forces of 300 pounds and 400 pounds act at right angles to each other on the same point. Find the magnitude and the direction of the single force that would replace these two. 500 lbs at 37° with the 400 force.
2. Three ropes are attached to a freight car standing on a north-south track. Through one, extending north, a 300-pound force is exerted. Through the second, extending 15 degrees east of north, a 200-pound force is exerted. Through the third, extending 20 degrees west of north, a 150-pound force is exerted. What northward force does the car experience? 630 lbs.
3. Three cords are fastened together at a point. Two of them are stretched by respective forces of 8 and 15 pounds and make a right angle with each other. Find

the force which, when applied to the third at an angle to be calculated, will put the three in equilibrium. **17 lbs at 28° with the 15-lb force.**

4. A rope is fastened to two hooks 24 feet apart with 26 feet of it suspended between them. 100 pounds is hung at the middle. What is the tension in the rope and the angle it makes with the horizontal? **130 lbs, 23°.**

5. A boy weighing 50 pounds sits in a swing having ropes 10 feet long. He is pulled back through a horizontal distance of 6 feet. What angle do the swing-ropes make with the vertical? What horizontal force is required? Under what tension is each swing-rope? **37°, 38 lbs, 31 lbs.**

6. A sled of weight W pounds experiences a force of F pounds through the pull of a rope which makes an angle of α degrees with the horizontal. What horizontal force H in pounds is acting, and what would be the apparent weight w in pounds, of the sled if drawn across a platform balance?

W	F	α	H	w
20	10	10°	9.8	18
20	10	20°	9.4	17
20	10	30°	8.7	15
20	10	45°	7.1	13

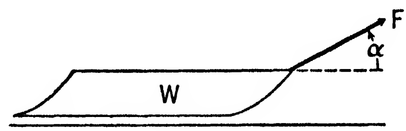


Fig. 22

7. A steamship travels v miles per hour at an angle of α degrees east of north. Find the northward N and the eastward E components of its velocity in miles per hour.

v	α	E	N
20	20°	6.8	19.
20	40°	13.	15.
20	60°	17.	10.
20	80°	20.	3.5

8. A picture whose weight is W pounds is hung from a nail, the cord making an angle of α degrees with the horizontal. What is the tension T in pounds in the cord?

W	α	T
20	20°	29
20	30°	20
20	40°	16
20	50°	13

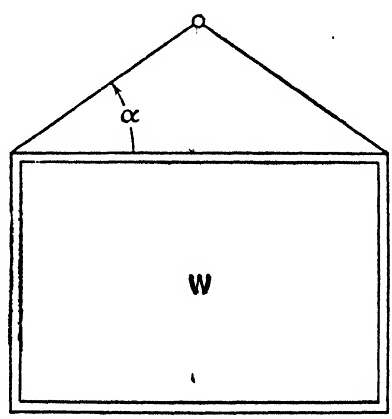


Fig. 23

9. A weight W pounds is hung from the free end of a horizontal, weightless, hinged boom. To the free end of this boom is also attached a rope which makes an angle

α degrees with the horizontal. Find the tension T in the rope and the compression C in the boom, in pounds.

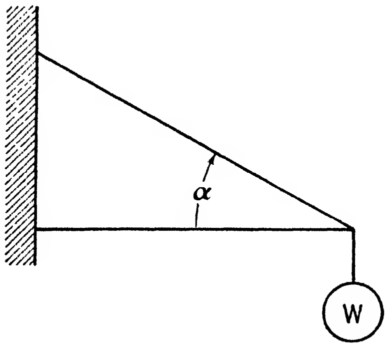


FIG. 24

W	α	T	C
100	30°	200	170
100	40°	160	120
100	50°	130	84
100	60°	120	58

10. A horizontal force of F pounds acting on a telephone pole is to be balanced by the tension G pounds of a guy wire which makes an angle of α degrees with the horizontal. Find G.

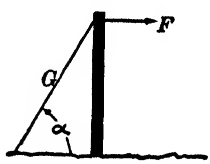


FIG. 25

F	α	G
1000	40°	1300
1000	50°	1600
1000	60°	2000
1000	70°	2900

11. The force with which a piston rod bears on its crosshead is F tons. The connecting rod makes an angle of α degrees with the horizontal. Neglecting friction and the weight of the parts, what is the compression C in tons in the connecting rod, and the upward push P in tons of the lower guide on the crosshead.

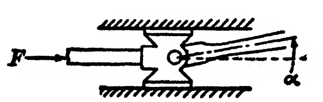


FIG. 26

F	α	C	P
100	5°	100.	8.7
100	8°	101.	14.
100	10°	101.	18.
100	12°	102.	21.

12. Air brakes are usually applied through a "toggle joint." A force of f pounds is applied through such a joint, each arm making an angle of α degrees with the horizontal. Find the force F in pounds with which each brake bears on its wheel.

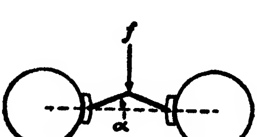


FIG. 27

α	f	F
4°	10	72
3°	10	96
2°	10	140
1°	10	290

13. At a distance a from one end of a loose cord of length $(a + b)$ is attached a weight W pounds. There is an angle of ninety degrees between the two sections of the cord. Find the tension A and B in each section, in pounds.

$a+b$	a	W	A	B
7	3	100	80.	60.
17	5	100	92.	38.
23	8	100	88.	47.
31	7	100	96.	28.

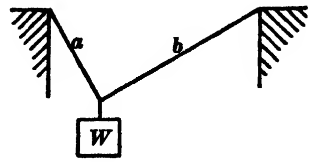


FIG. 28

14. At a distance a feet from one end of a loose cord of length $(a + b)$ feet is attached a weight W pounds. The ends of the cord are tied to supports a distance c feet apart. Find the tension A and B in each section of the cord, in pounds.

$a+b$	a	c	W	A	B
7	3	6	100	100.8	90.63
13	5	12	100	131.3	120.5
13	6	9	100	74.69	63.03
15	5	14	100	137.3	123.0

15. Find the tension (or compression) in each member of the pinned trusses shown below, when loaded as shown. The sides of every triangle are in the ratio 3–4–5. The given quantities are the load (80) and the dimensions (3–4–5). Italicized answers represent compressions. Those not italicized are tensions.

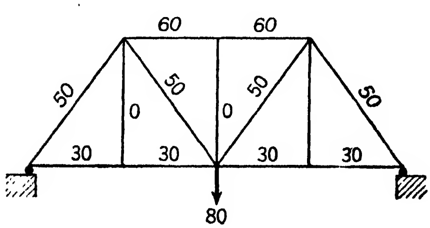


FIG. 29

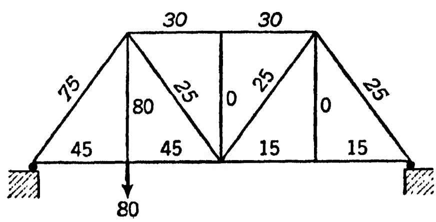


FIG. 30

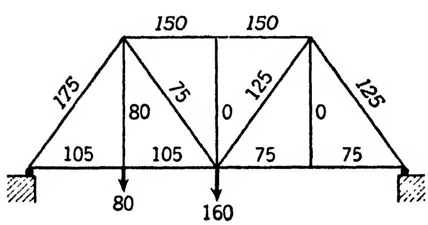


FIG. 31

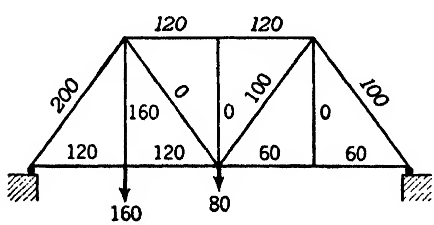


FIG. 32

CHAPTER 5

Equilibrium; the Inclined Plane and Composition of Vectors

The Inclined Plane

Another practical case which may be treated by resolution of forces is that of the inclined plane. The term *inclined plane* calls up a vision of a sloping plank such as is used in loading and unloading freight. This is, of course, a rudimentary form. In that form its utility is rather limited. But the broader significance of the inclined plane as a mechanical device will become evident by recalling that the wedge and the screw both embody it in principle. Ubaldi, in his *Mechanicorum Liber* (127), was the first to point this out. In this work he observed that the screw might be considered as a wedge wrapped around a cylinder (see Fig. 33), the wedge, in turn, being a manifestation of an inclined plane (101: 130).

The more obvious properties of an inclined plane are very obvious indeed. An object resting on such a plane will slide toward the lower end unless its motion is prevented. The more steeply is the plane inclined, the greater is the force that is necessary to prevent an object sliding along it. One is led to inquire what the ratio is between the angle of inclination and the force necessary to prevent the motion of a given object resting on it or — what is the same thing if friction may be disregarded — the force necessary to move the object up the plane.

The problem is recognizable as a simple case of resolution of forces. The

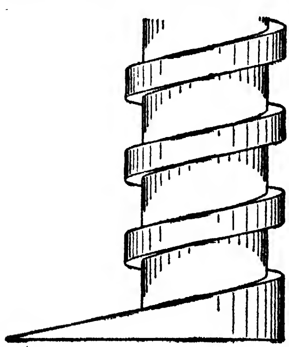


Fig. 33. The Screw as an Inclined Plane

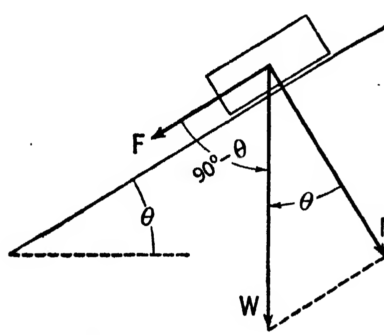


Fig. 34. Analysis of a Frictionless Inclined Plane

force necessary to hold the object in place will have the same magnitude as the component F (along the plane) of the vertical pull W of gravity (Fig. 34). By the principle of resolution, this component has the value

$$F = W \cos (90^\circ - \theta) \text{ or}$$
$$F = W \sin \theta. \tag{1}$$

Hence, the required force is proportional to the sine of the angle. Equation (1) may be checked by observing that the required force should become zero when the plane is horizontal and should become the full weight W when the plane is raised to the vertical. Since the sine of zero degrees is zero and the sine of ninety degrees is unity, substitution of these values for the sine in equation (1) will be seen to give the correct values for these known cases.

The other aspect of the problem has to do with the force with which the object bears against the plane. For a horizontal plane this will be simply the weight of the object. For a vertical plane it clearly has the value zero. Reference to Figure 34 will show that the component N of the gravitational pull W, which is perpendicular or *normal* to the plane, is given by

$$N = W \cos \theta. \tag{2}$$

Since the values of a cosine are unity and zero respectively for the zero and ninety-degree values of the angle, equation (2) will be seen to satisfy these conditions as mentioned above.

The Laws of Friction

The first thorough study of the laws of friction was made by Charles-Augustine Coulomb (1736–1806) (29), who is better known for his pioneer work in electricity. The results embodied essentially the modern view of the action of friction. This is that the force which retards the motion of one surface over another depends entirely upon two factors. These are (1) the materials and degree of polish of the surfaces involved and (2) the force pressing the surfaces together. It is independent, within certain rather wide limits, of the areas in contact, an observation first recorded by Leonardo da Vinci (68:10), and of the speeds with which they move over each other. It can, of course, be greatly reduced by lubrication, but this is, in effect, the substitution of a liquid surface of contact for the former solid surfaces and hence is included in case (1). It is found that a greater force is required to initiate the motion of one surface over another than to maintain it after it has once been started; also that the retarding force is, of course, greater for the case of sliding than for the case of rolling. Hence, it is common to distinguish between three types of friction: starting, sliding, and rolling. Except as otherwise specified, it is sliding friction that is under consideration herein.

The dependence of frictional retarding force f on the perpendicular or *normal* force N pressing the surfaces together is found to be that of direct proportion. Stated algebraically, using the conventional sign of proportionality, $\propto$,

$$f \propto N.$$

This may be stated as an equation by introducing a factor of proportionality, k, the value of which must then be determined experimentally, the equation being

$$f = kN. \tag{3}$$

Thus, by measuring the frictional force f between two surfaces pressed together by a normal force N, the value of the factor k will be determined for that case. For smooth surfaces this value of k is found to depend primarily on the materials involved and is termed the *coefficient of friction.* This coefficient will have one value for steel on steel, another for steel on brass, and so on.

The Limiting Angle of Repose

That the element of friction plays an important part in the practical theory of the inclined plane will immediately be evident. This is especially true in the case of the wedge and the screw, the principal mechanical outgrowths of the inclined plane. Not all of the force F of Figure 34, for example, will be effective in accelerating the block down the slope. From F must be subtracted the frictional force f. But friction is the normal force N multiplied by the coefficient of friction k. That is,

$$f = kW \cos \theta. \tag{4}$$

Unless the plane is tilted sufficiently so that the component F along the plane is at least equal to the frictional force f, no sliding will occur. Since, as the plane is sloped more and more, F increases and f decreases, it will be evident that for some particular angle they will become equal to each other. The angle for which this condition exists is known as the *limiting angle of repose.* Its value is derivable from a statement of the equality of the two forces, that is,

$$F = f \quad \text{or} \quad W \sin \theta = kW \cos \theta,$$

from which

$$\frac{\sin \theta}{\cos \theta} = \tan \theta = k \tag{5}$$

Hence, when the coefficient of friction between two surfaces is known, the limiting angle of repose can at once be stated by finding with the aid of a table of tangents the angle θ which corresponds to the given value of the tangent.

Thus no sliding can occur down an incline or along a screw unless the angle of the incline or pitch of the screw is greater than a certain value dependent on the coefficient of friction between the sliding surfaces. It is simple to demonstrate that an equivalent statement is that no force, however great, can cause an object (Fig. 35) to slide unless the angle θ which the force makes with the normal is at least as great as the limiting angle of repose. It is for this reason that wedges seldom slip out of place and that screws are seldom designed to be turned by pushing the nut along their length. The action of the "lock washer" depends on this principle. Only by a combination of good lubrication and large angle can wedges and screws be made thus to reverse their normal action.

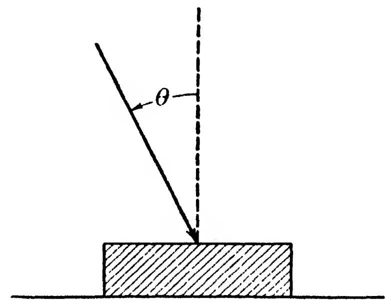


FIG. 35. WILL IT SLIDE?

Composition of Forces

The process of composing or combining forces is simply the reverse of the problem of resolving them. Instead of having one force given and the task assigned of finding two or more components which, acting jointly, will be a complete equivalent of the given single force, the task is now that of finding a single force which is the complete equivalent of two or more given forces.

The simplest case of composition of forces is clearly that in which the forces act in the same line, either in the same or in opposite directions. The magnitude of the resultant is then either the arithmetical sum or the arithmetical difference of the two components. While the case is not trivial, it is too elementary to require more than passing notice.

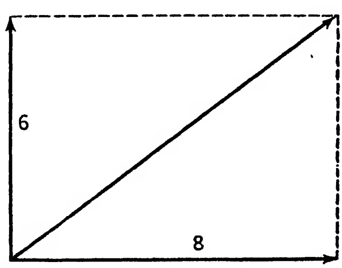


FIG. 36. RESULTANT OF TWO FORCES AT RIGHT ANGLES

Two Forces at Right Angles

The case of two mutually perpendicular components, while still simple, merits close attention, for it contains the essence of the whole problem of composition of forces. It is illustrated in Figure 36, which is simply a repetition of Figure 15 except that in this case the two component vectors, one of magnitude *8*, directed toward the right, the other of magnitude *6*, directed upward, are given instead of the resultant being given. The procedure involved in finding the resultant is simply the reverse of that involved in the process of resolution. Complete the rectangle, as indicated by the dotted lines. The resultant is then represented both in magnitude and in direction by the diagonal vector. The magnitude will, of course, be 10 and the direction 37° with the horizontal, as is shown by mere comparison with Figure 15.

As in the case of resolution of vectors, the graphical method of composition may advantageously be replaced by an arithmetical method. It will be clear that for components mutually at right angles the magnitude of the resultant will be given by the application of the Pythagorean theorem, the square root of the sum of the squares. Similarly, the angle that the resultant makes with the horizontal will have its tangent specified by the quotient of the vertical by the horizontal magnitudes. Expressed algebraically (see Fig. 37),

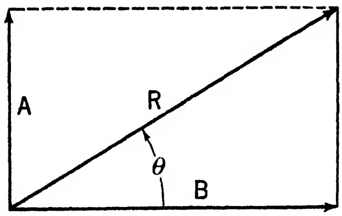


FIG. 37. THE GENERAL RECTANGULAR CASE

$$R = \sqrt{A^2 + B^2}. \qquad (6)$$

$$\tan \theta = \frac{A}{B}. \qquad (7)$$

It will be noted that the right side of each of these equations contains only the data originally given; hence, that the magnitude R and the direction θ may immediately be computed.

Oblique Forces

The case of the composition of two vectors not at right angles is, naturally, less simple; but the principle is the same. Newton stated it as follows (91:16): "A body acted on by two forces simultaneously will describe the diagonal of a parallelogram . . ." whose sides represent the given forces. The case is represented in Figure 38, in which the given vectors are A and B, making with each other the given angle ϕ. The resultant may be found graphically by completing the parallelogram and drawing the diagonal R. The magnitude of the resultant is then represented by the length of R and the direction by the angle θ which R makes with one of the given vectors, say B.

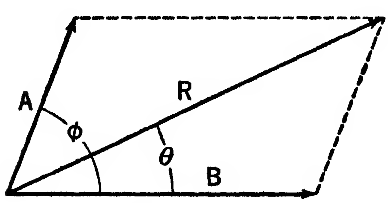


FIG. 38. COMPOSITION OF OBLIQUE FORCES

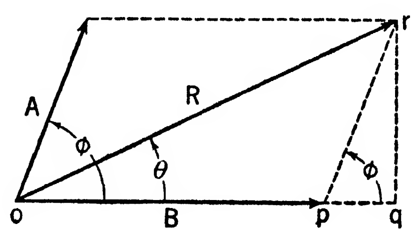


FIG. 39. THEORY OF COMPOSITION OF OBLIQUE FORCES

But again, the graphical solution, though furnishing an excellent illustration of the principle, must be supplemented by a method analogous to that embodied for the rectangular case in equations (6) and (7). The corresponding equations for the oblique case are

$$R = \sqrt{A^2 + B^2 + 2\,AB \cos \phi} \qquad (8)$$

$$\text{and} \quad \tan \theta = \frac{A \sin \phi}{B + A \cos \phi}. \qquad (9)$$

It will be noted that the right side of each of these equations contains only the data originally given; hence, that the magnitude R and the direction θ of the resultant may immediately be computed.

These equations may be derived as follows. Referring to Figure 39 which is the same as Figure 38 with the two construction lines pq and qr added, it will be evident that

$$\begin{aligned} R^2 &= \overline{qr}^2 + \overline{oq}^2 = \overline{qr}^2 + (op + pq)^2 \\ &= (A \sin \phi)^2 + (B + A \cos \phi)^2 \\ &= A^2 \sin^2 \phi + B^2 + 2AB \cos \phi + A^2 \cos^2 \phi \\ &= A^2 (\sin^2 \phi + \cos^2 \phi) + B^2 + 2AB \cos \phi \\ &= A^2 + B^2 + 2AB \cos \phi \text{ (since } \sin^2 \phi + \cos^2 \phi = 1). \\ \therefore R &= \sqrt{A^2 + B^2 + 2AB \cos \phi} \end{aligned}$$

which is equation (8).

Similarly, $$\tan \theta = \frac{qr}{oq} = \frac{qr}{op + pq} = \frac{A \sin \phi}{B + A \cos \phi}$$

which is equation (9).

Equilibrants

Equations (8) and (9) give respectively the magnitude and direction of the *resultant* of two given forces. Often it is the *equilibrant* which is desired; that is, the single force which will just balance or equilibrate the two given forces. As would naturally be surmised, this will possess the same magnitude as the resultant and be oppositely directed (Fig. 40). Hence the same computation that gives the resultant of two given vectors will also yield the equilibrant, simply by the addition or subtraction of 180° in connection with the computed angle θ.

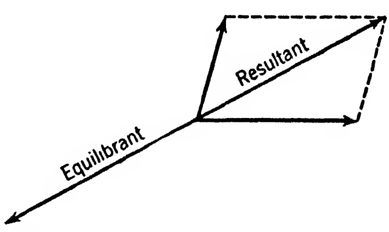


FIG. 40. THE EQUILIBRANT BALANCES THE RESULTANT

The most commonly used device for the experimental study of combinations of forces in equilibrium is the so-called force table, devised by the French mathematician, Augustin-Louis Cauchy (1789–1857) (74:47). It usually consists primarily of a large horizontal graduated circle. Three or more cords, tied to an otherwise unattached ring at the center, pass over pulleys and support weight-pans. The pulleys may be clamped to any portion of the graduated circle. The experiment consists of adjusting the weights and the angles to secure a balance, whereupon equilibrium is known to exist and may be tested by the application of equations (8) and (9), regarding any one of the forces as the equilibrant of the remaining two or more (see Fig. 41).

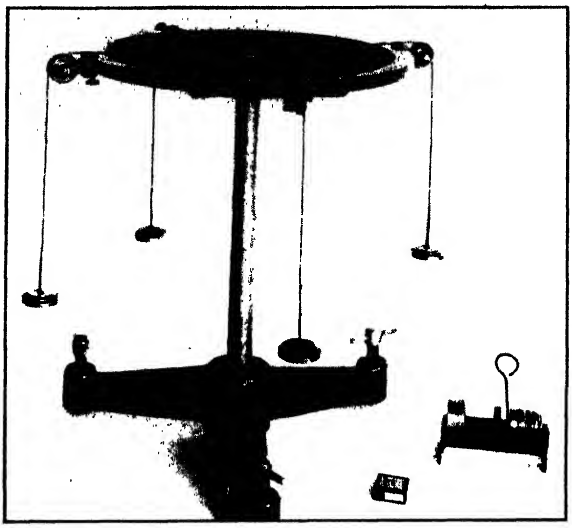


FIG. 41. FORCE TABLE

Questions for Self-Examination

1. State the principal attributes of friction.
2. What is meant by the term *limiting angle of repose*? How does it apply to the use of wedges?
3. Formulate a graphical example of composition of forces.

Problems on Chapter 5

1. Find the angle between the two ropes of a hammock when the tension on each rope is twice the weight of the person in the hammock. 30°.
2. A 260-pound barrel rests on an inclined plane 13 feet long with one end 5 feet higher than the other. Find the components of the gravitational force on the barrel both parallel and perpendicular to the plane. 100 lbs. 240 lbs.
3. Find the resultant, both in direction and magnitude, of five forces of magnitude 3, 4, 5, 6, and 7 respectively, having directions represented by the five sides of a regular pentagon taken in order. 4.3 at 234° with first side.
4. Four forces, A, B, C, and D, act at a point. B is twice as great as A and acts at right angles to it. C is equal to the sum of A and B and acts at right angles to their resultant. D is equal to the sum of A, B, and C and acts at right angles to their resultant. Show that the resultant of all four is $5A\sqrt{2}$ (reckon all right angles in the same direction).
5. A frictionless car weighing W pounds is on an incline which makes an angle of α degrees with the horizontal. What force F in pounds, parallel to the incline, is necessary to hold it in place, and with what force w in pounds does the incline support the car?

W	α	F	w
200	10°	35	200
200	20°	68	190
200	30°	100	170
200	45°	140	140

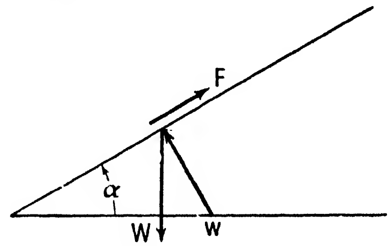


FIG. 42

6. Given two forces of magnitude A and B pounds, differing in direction by an angle ϕ degrees, find the magnitude R in pounds of the resultant, and the angle θ in degrees that it makes with the force B.

A	B	ϕ	R	θ
200	100	45°	280	30°
300	200	30°	480	18°
400	300	120°	360	73°
150	100	20°	130	79°

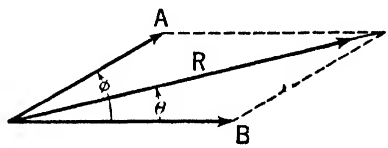


FIG. 43

7. Given three forces of magnitude A, B, and C pounds, making angles α, β, and γ degrees respectively (measured counterclockwise) with the positive direction of an x-axis, find the magnitude E in pounds and the angle θ in degrees, which the *equilibrant* makes with the same line.

A	α	B	β	C	γ	E	θ
200	0°	300	120°	400	220°	260	$-\frac{1}{2}°$
300	0°	250	110°	350	250°	130	+ 130°
535	0°	425	100°	350	200°	330	+ 250°
215	0°	165	130°	200	290°	190	+ 160°

8. A horizontal force of F pounds is found to be necessary to produce uniform velocity in a sled of weight W pounds. What is the coefficient of friction k?

W	F	k
20	2	.10
20	5	.25
20	8	.40
20	12	.60

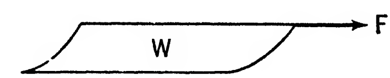


FIG. 44

9. What force of F pounds, *pulling* at an angle of α degrees with the horizontal, is necessary to move a sled of weight W pounds with uniform velocity over a horizontal surface, the coefficient of friction being k?

W	α	k	F
10	0°	.3	3.
10	15°	.3	2.9
10	40°	.3	3.1
10	60°	.3	3.9

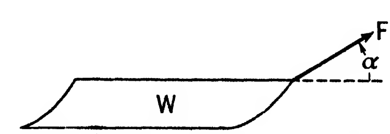


FIG. 45

10. What force of F pounds, *pushing* at an angle of α degrees with the horizontal, is

necessary to move a sled weighing W pounds with uniform velocity over a horizontal surface, the coefficient of friction being k? Interpret the negative sign of the last answer.

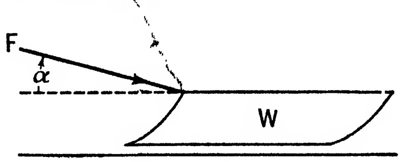


FIG. 46

W	α	k	F
10	0°	.3	3.
10	15°	.3	3.4
10	40°	.3	5.2
10	75°	.3	−97.

11. What force of F pounds, acting parallel to the incline, is necessary to move a sled of weight W pounds up an incline which makes an angle of α degrees with the horizontal, the coefficient of friction being k?

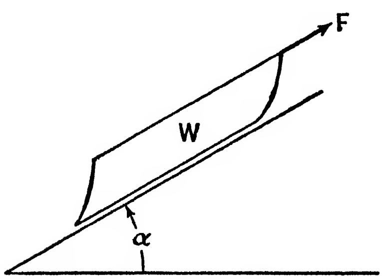


FIG. 47

W	α	k	F
10	30°	.3	7.6
10	50°	.3	9.6
10	70°	.3	10.4
10	80°	.3	10.3

12. What force is necessary to move the sled of the preceding problem *down* the incline? Interpret the negative sign of the last answer.

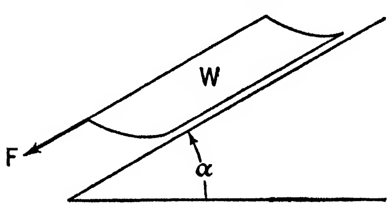


FIG. 48

W	α	k	F
10	5°	.3	2.1
10	10°	.3	1.2
10	15°	.3	.31
10	25°	.3	− 1.5

13. What force of F pounds, acting horizontally, is necessary to move a sled of weight W pounds up an incline which makes an angle of α degrees with the horizontal, the coefficient of friction being k?

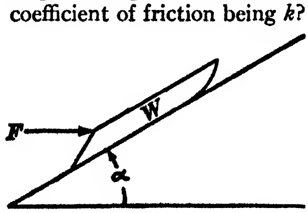


FIG. 49

W	α	k	F
10	5°	.3	4.0
10	10°	.3	5.0
10	15°	.3	6.2
10	20°	.3	7.6

14. At what angle α in degrees will a sled *just* slide down an incline without the application of a force, the coefficient of friction being k?

k	α
.1	5° 45′
.2	11° 15′
.3	16° 45′
.5	26° 30′

15. What force F in pounds is necessary to move the sled of problem 13 *down* the incline?

W	α	k	F
10	5°	.3	2.1
10	10°	.3	1.2
10	15°	.3	.30
10	20°	.3	− .58

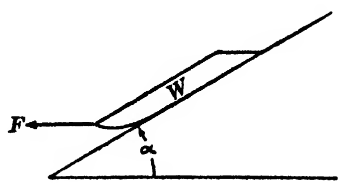


Fig. 50

16. What force of F pounds, acting at an angle β degrees with an incline, is necessary to *pull* a sled of weight W pounds up an incline which makes an angle of α degrees with the horizontal, the coefficient of friction being k?

W	β	α	k	F
10	5°	20°	.3	6.1
10	10°	20°	.3	6.0
10	15°	20°	.3	6.0
10	25°	20°	.3	6.3

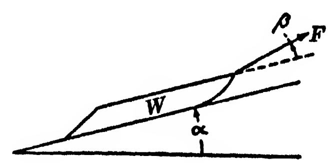


Fig. 51

17. What force is necessary to *push* the sled of the preceding problem up the incline?

W	β	α	μ	F
10	5°	20°	.3	6.4
10	10°	20°	.3	6.7
10	15°	20°	.3	7.0
10	25°	20°	.3	8.0

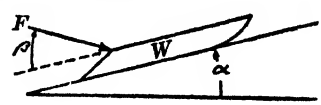


Fig. 52

18. What force is necessary to *pull* the sled of problem 17 *down* the incline?

W	β	α	μ	F
10	5°	10°	.3	1.19
10	10°	10°	.3	1.17
10	15°	10°	.3	1.17
10	25°	10°	.3	1.18

Fig. 53

19. What force is necessary to *push* the sled of problem 18 down the incline?

W	β	α	μ	F
10	5°	10°	.3	1.3
10	10°	10°	.3	1.3
10	15°	10°	.3	1.4
10	25°	10°	.3	1.6

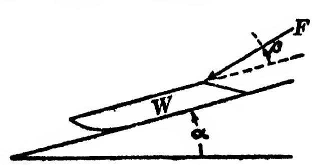


Fig. 54

CHAPTER 6

Equilibrium; Non-Concurrent Forces

Equilibrium in Rotation

Equilibrium for the case of concurrent forces, which was treated in the preceding chapters, was seen to exist if the algebraic sums of the respective vector components assumed the value zero. But if the vectors do not all meet at a common point, this condition of equilibrium is not sufficient. For this case the fact that the vector components may add up to zero does not necessarily imply that the object to which they are applied is in equilibrium. For example: the stick of Figure 55 is acted upon by two equal and opposite forces, both perpendicular to the stick, but at opposite ends. Though the algebraic sum of the forces is clearly zero, the stick is obviously not in equilibrium, but will undergo a clockwise rotational acceleration.

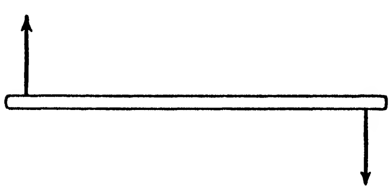


FIG. 55. THE STICK IS IN TRANSLATIONAL BUT NOT IN ROTATIONAL EQUILIBRIUM

It therefore appears that to produce equilibrium in non-concurrent forces the condition that the vectors add to zero is not sufficient. Some other circumstance in addition is required. A clue to the nature of this circumstance may be found in the fact that the lack of equilibrium in the foregoing example made itself evident in rotational acceleration. Apparently the new condition of equilibrium involves the prevention of rotational acceleration. That is to say, it involves ability to produce an equal and opposite rotational acceleration.

The idea of rotational acceleration, though not introduced heretofore, should present no difficulty, particularly in view of the merely parenthetical use that is to be made of it in this chapter. It will be more fully treated in Chapter 15. It is to rotational motion what linear acceleration is to translational motion, this having been described in Chapter 3. With translational acceleration, as will be seen in Chapter 10, the idea of *force* is closely associated. But wherever rotation is involved, an additional element enters. The opening of a heavy door is much easier when the knob is as far as possible from the line of the hinges. The force that is involved for one case is quite different, for the same effect, from that involved in the other. A very little experimentation will show that the force required depends on the distance of the point of its application from the hinged side of the door; the greater the distance, the smaller the force needed. Hence, the practice of placing the doorknob at the opposite side of the door from the hinges is a utilitarian measure, a departure from the consideration of symmetrical appearance which presumably was responsible for placing the knob at the center of the door in early English houses.

The relation between the force required to operate a door and the distance of the point of its application from the line of the hinges is the very simple one of an inverse ratio. A push in the middle must be twice as strong as one at the outer edge to produce the same effect. This suggests that with a given rotational acceleration may be associated a corresponding entity which consists of the product of a force by the corresponding distance from the axis of rotation. If in a given rotating body a steady rotational acceleration is to be maintained, any alteration of the magnitude of the force being exerted must be compensated by a reciprocal alteration of its distance from the axis of rotation, the product of the two being, for a single such force, always proportional to the rotational acceleration.

Rotational Equilibrium

The entity thus associated with rotational acceleration, in the same way that force is associated with translational acceleration, has proven of sufficient importance to be worthy of a name. The name in most common use is *torque*, from a Latin verb meaning to twist.[1]

Hence it may be said that the rotational acceleration of a given rotating body is proportional to the applied torque. The equilibrium case of zero rotational acceleration will then be seen to be associated with zero torque. This does not imply the absence of all torques, any more than translational equilibrium implies the absence of all forces. It simply means that the algebraic sum of all the torques acting must be zero. This involves a sign convention for torque and for rotational motion. The sign is commonly taken as positive for counterclockwise rotation and negative for clockwise. With this convention it is possible to state what is termed the *condition of rotational equilibrium. Rotational equilibrium is realized when the algebraic sum of all torques is zero.* This corresponds, for the case of rotation, to the condition of translational equilibrium, namely, that *translational equilibrium is realized when the vector sum of all the forces is zero* (page 36). This condition of *translational* equilibrium is sometimes referred to as the first principle of statics, and the foregoing condition of *rotational* equilibrium as the second principle of statics. The two principles taken together embody the whole science of statics, which until the time of Galileo was the sole content of mechanics.

[1] The term *moment of force*, long used for this concept, seems to be dropping out.

Application of Rotational Equilibrium

The condition of rotational equilibrium may be illustrated by using it to compute the reactions at the supports of a loaded beam. Suppose loads of 200, 500, and 600 pounds respectively to be concentrated as in Figure 56 at distances of 3, 5, and 8 feet from the left end of a 10-foot beam, the weight of which is negligible in comparison with the load. The problem is to find the share of the load borne by each end support, A and B, which will therefore be the unknowns in the equilibrium equations. As long as the beam does not break and neither of the supports gives way, a condition of equilibrium obtains and hence both principles of statics apply. Consider first the second principle.

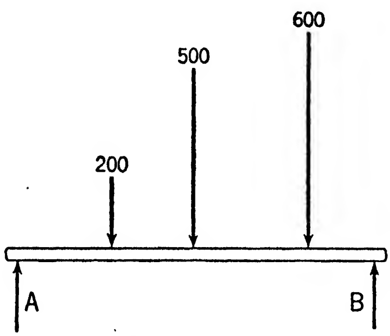


Fig. 56. Reactions at Supports

In constructing a torque equation it is necessary to fix upon some axis of rotation. For torque is expressible as *force times lever arm,* and the lever arm is the distance from the line of action of the force to the axis of rotation, *real or assumed.* Since in this case there is no rotation at all, an axis must be assumed, and this assumption can be entirely arbitrary. Since there is no rotation about any axis whatever, there is bound to be no rotation about any particular assumed axis that may be selected. Any indecision that may be felt about where to choose such an axis may be relieved by the observation that there is some advantage in causing it to pass through the line of action of one of the unknown forces. The torque corresponding to that force thereupon becomes zero, since its lever arm is zero, thus simplifying the equation. In this way, taking the axis at the right end of the beam, perpendicular to the plane of the diagram, the following torque equation may be constructed:

$$-A \cdot 10 + 200 \cdot 7 + 500 \cdot 5 + 600 \cdot 2 = 0,$$

the signs being chosen in accordance with the convention stated on page 53, namely, positive for counterclockwise rotation and negative for clockwise.

The solution of the above equation may readily be verified as $A = 510$, which thus constitutes the value of one of the two unknowns. The second may be found by constructing another torque equation, this time preferably (though not at all necessarily) assuming an axis of rotation through the left end of the beam. This equation is

$$-200 \cdot 3 - 500 \cdot 5 - 600 \cdot 8 + B \cdot 10 = 0,$$

whence $B = 790$ pounds, thus evaluating the second unknown. The correctness of the solution may be checked by using the first condition of equilibrium, which is known to apply because the beam, by hypothesis, has zero translational acceleration as well as zero rotational acceleration. This condition yields the force equation

$$510 + 790 - 200 - 500 - 600 = 0,$$

which is seen to be true.

Meaning of the Term "Lever Arm"

The identification of the lever arm is not always as simple a matter as the foregoing illustration would indicate. Suppose, for example, that one of the forces, say the *600* force, were applied obliquely, making an angle of 30° with the horizontal, as in (a) of Figure 57. Would its lever arm about the right end still be two feet? A little consideration will show that an affirmative answer to this question would lead to serious ambiguities. To be sure, the point of application is still two feet from the assumed axis of rotation; but suppose a boss or extension to have been a part of the structure, as in (b) of Figure 57, and the force applied to that. The distance from the axis of rotation to the point of application is now different, yet it will be evident that the situation has really not been changed at all.

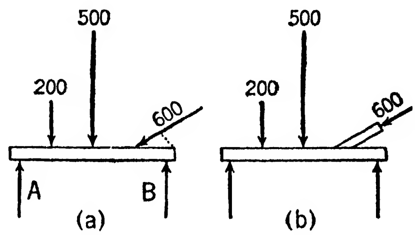


FIG. 57. EFFECT OF AN OBLIQUE FORCE

The difficulty disappears by focusing attention on *the distance from the axis of rotation to the line of action of the force;* indicated by the dotted line in Figure 57. This distance is commonly called the *lever arm.* It was first recorded in a book by Giovanni Battista Benedetti (1530–90). On page 143 of his *De Mechanicis* (17), he pointed out that as far as rotation about the point *o* is concerned (Fig. 58) the oblique force *c* applied at *a* could be replaced by a vertical force of the same magnitude applied at *i* where *oi* had the same length as *ot*, the *perpendicular distance from the axis to the line of action of the force.* The depth of Benedetti's perception lay

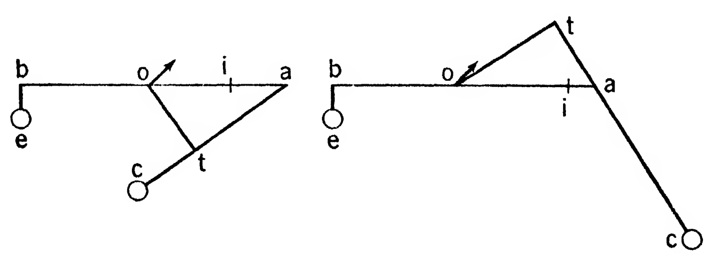


FIG. 58. THE FIRST REALIZATION OF THE REAL SIGNIFICANCE OF TORQUE
(From the *De Mechanicis* of Benedetti, 1599.)

in this divorcement of the idea of the lever arm from the structural details of the object under consideration, and the identification of it as the perpendicular distance from the axis to the line of action of the force. This he saw to be the case even in the complete absence from that region of any part of the structure of the object or mechanism associated with the applied force.

Applying this definition of lever arm to the inclined force of Figure 57, it will be evident that the lever arm of that force about the right-hand support is given by the dotted line the length of which may easily be computed to be one foot. It applies equally to both cases illustrated in Figure 57, the former ambiguity having disappeared. By applying anew the conditions of equilibrium it may be found that the *vertical* reactions at the supports for the new case will be $A = 450$ pounds and $B = 550$ pounds, and that a horizontal force of about 520 pounds toward the right must be applied somewhere along the length of the beam to produce translational equilibrium in that direction.

The Idea of Center of Gravity

The condition of rotational equilibrium may be applied in a slightly different way which will be found to lead to two major mechanical concepts, that of the center of gravity and that of the balance.

Suppose that the loaded beam of Figure 56, instead of being supported at the ends, were to be balanced on a single support as in Figure 59. In this case, not merely the magnitude of the reaction at that support is required, but also the location of the point of application. There are thus two unknowns, and these the two torque equations can be made to yield as before. Let C be the magnitude of the equilibrant and x its unknown distance from the left end (Fig. 59). Taking each end in turn as the assumed axis and neglecting the weight of the beam as before, the torque equations are

$$-200\cdot 3 - 500\cdot 5 + C\cdot x - 600\cdot 8 = 0$$
$$200\cdot 7 + 500\cdot 5 - C\cdot(10 - x) + 600\cdot 2 = 0.$$

The solution is $C = 1300$, $x = 6\frac{1}{13}$. The condition of translational equilibrium may be used as a check as before.

It thus appears that a single upward force of 1300 pounds, applied

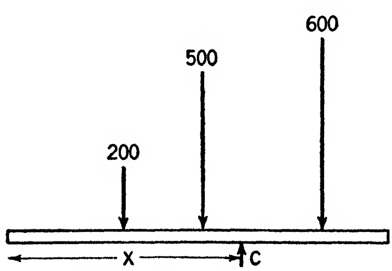


FIG. 59. THE CENTER OF "GRAVITY"

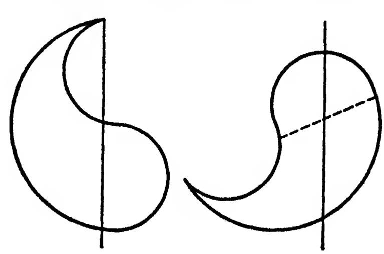


FIG. 60. IDENTIFYING THE CENTER OF GRAVITY OF A PLANE OBJECT

slightly more than six feet from the left end of the beam, will equilibrate the three downward forces on the beam. But such an upward force would also equilibrate a single downward force of 1300 pounds applied at the same point; whence it follows that the three downward forces distributed along the beam are in a certain sense the equivalent of a single downward force concentrated at a point $6\frac{1}{13}$ feet from the left end. The preceding chapter developed the fact that two or more forces applied at a single point could be replaced by a single force called the *resultant.* It is now evident that the same concept can be extended to the case of non-current forces, by the correct selection of a point of application. If the three downward forces, 200, 500, and 600, were produced by weights applied at the given points on the beam, the resultant vector would pass through a point called the *center of gravity* of the three weights jointly. Thus, the center of gravity may be defined as *the point at which the gravitational attraction on a system of distributed weights may be considered to be concentrated.* In identifying it for a given system of such weights, use may be made of the condition of rotational equilibrium.

Identification of Centers of Gravity

The last sentence may elicit some dissent. Surely we know that the center of gravity of a uniform rod, for example, is at the mid-point of its length, or that the center of gravity of a circular plate or a rectangle or a square is at its respective geometrical center; and we know this without going through a process of setting up and solving torque equations. The answer to this objection is simply that the common knowledge that the center of gravity of a uniform rod is at its midpoint proceeds from the numerous observations that have been previously made to the effect that it "balances" at that point. This is simply a manual process of solving torque equations. It is less convenient for plane objects and is so exceedingly inconvenient for solids that for them the process of actual calculation with the aid of torque equations is almost mandatory. The details of such calculation involve some mathematical processes not at the command of the average underclassman, and the calculation will not be attempted here. The physical concepts are, however, no different from those already developed. In every case of identification of centers of gravity, the condition of rotational equilibrium will be seen to be the foundational process.

A common practical way to locate the center of gravity of a plane figure is to suspend it from two points. The center of gravity will then be the intersection of the vertical lines through the points of suspension. This is simply another case of manually solving two equations of rotational equilibrium. The suspended object can come to rest only when the algebraic sum of all torques upon it is zero. This, in turn, can only occur for a position in which the lever arm has a value zero, for the weight cannot become zero. The lever arm will have the value zero when the center of gravity is in a vertical line with the point of support. Hence, the unknown center of gravity of a freely suspended plane object will be found somewhere on a vertical line through the point of support, a fact first deduced by Archimedes in the third century B.C. (5:xxxvii). The intersection of two lines thus determined gives its location.

Stability

This is equally true in principle when the object is so turned that the center of gravity is above, instead of below, the point of support. But it is next to impossible to secure a balance for such a position, for the force called into play by any disturbance accentuates the disturbance. The equilibrium is said to be unstable for this case of the precariously balanced object, and stable for the case of the suspended object. In the latter case, the center of gravity is below, instead of above, the point of support, hence any displacement from the position of equilibrium calls into play forces which restore, instead of destroy, the *status quo ante*. For this case it is apparent that at stable equilibrium the center of gravity is at the lowest possible level consistent with the constraints imposed upon the body, whereas in the precariously balanced case the center of gravity is at the highest possible point. The altitude of the center of gravity thus becomes a measure of instability. A piece of lumber balanced on one end is in a very unstable state of equilibrium, its center of gravity being high above the supporting level. The equilibrium is most stable when the board lies flat, the center of gravity then being very close to the supporting level; and the stability has an intermediate value when the board is "on edge," with the center of gravity at some level between the two extremes. The first study of stability in equilibrium was made by Simon Stevin, the man who anticipated Galileo in experimentation on falling bodies.

The Principle of the Beam Balance

The influence of the relative positions of center of gravity and point of support is especially prominent in the design of the beam balances used for weighing. The beam of a so-called *equal-armed balance* is so designed that when its center of gravity is directly below the point of support, the beam, after oscillation about a horizontal position, ultimately comes to rest horizontally. Figure 61 shows the working principle of this device, *a* representing the center of gravity and *b* the point of support. If the center of gravity is very far below the point of support, the balance becomes very insensitive. That is, it requires a relatively large unbalancing torque to produce a given displacement. This is because the length of the lever arm, which is greater in proportion as the distance *ba* is greater, creates a large restoring torque

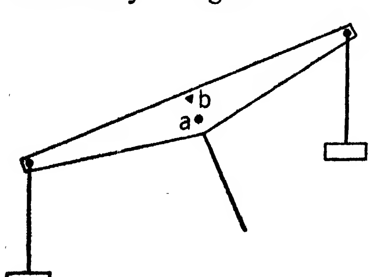


FIG. 61. MODEL BALANCE

for even small angular displacements from the horizontal. While this insensitivity may be a disadvantage, such a balance is rapid in its action and likely to be more rugged than more sensitive models. But if sensitivity is greatly desired, it may be secured by decreasing the distance between the center of gravity and the point of support. Much smaller unbalancing torques will now produce the same disturbance of equilibrium as greater ones did before. More time and greater care will then be required to operate the balance than before. An infinite sensitivity may theoretically be obtained by causing the point of suspension to coincide with the center of gravity. Such a weighing device, in spite of its infinite sensitivity, would be quite useless, since, when in balance, it would remain in any position. Hence, a compromise is necessary between sensitivity and usability. In actual balances there is a wide range of sensitivities to meet the varying requirements of actual practice.

The primary effect of shifting the center of gravity of a balance arm up and down is thus seen to be the alteration of the sensitivity of the balance. An additional effect is on the period of oscillation. The further the center of gravity is removed from the point of suspension, the more rapidly the balance oscillates. The speed with which the weighing process is performed, therefore, increases as the sensitivity is decreased.

Only the equal-armed balance has been discussed. The possibility of balances having unequal arms did not seem to meet with general recognition until the time of the Romans, though the work of Archimedes laid the foundation for such devices. Their great advantage lies, of course, in the fact that with them heavy objects can be counterbalanced by small weights. The element of convenience and portability thereby introduced is in practice a major consideration. Prior to the day of Archimedes the unequal-armed balance was not understood, though we learn from Aristotle (8: Cap. I) that dishonest tradesmen were known to shift the "center" of their balances toward the pan in which the weights lay when selling their products (48:238). Systematic short weight is evidently not exclusively a modern practice!

Questions for Self-Examination

1. State and discuss the two principles of statics, showing that one without the other is insufficient to produce equilibrium.
2. Give an account of Benedetti's discovery of the general concept of the *lever arm.*
3. How are the locations of centers of gravity determined in theory and in practice?
4. What characterizes stable and unstable equilibrium? Give two examples of each.
5. What brings a balance beam to the horizontal position?
6. What determines the sensitivity of a balance and why?

Problems on Chapter 6

1. A beam is carried by three men, one at one end and the other two supporting it between them on a cross-piece so placed that the load is equally divided between

the three men. Find where the cross-piece is placed.

$\frac{1}{4}$ of the distance from the free end.

2. Upward forces of 50 and 75 pounds are applied to a 12-foot beam weighing 100 pounds at 3 and 10 feet from the left end. Downward forces of 80 and 120 pounds are concentrated at the left and right ends respectively. Find the magnitude and the point of application of the single additional force that will support and balance the loaded beam. 180 lbs, 6.5 ft from the left end.

3. The foot of a uniform ladder weighing 50 pounds rests at a distance 5 feet from a vertical frictionless wall. The upper end rests against the wall 12 feet from the ground. Find the force with which the ground supports the ladder and the angle which this force makes with the horizontal. Compare this with the angle that the ladder makes with the horizontal. 51 lbs at 78°.

4. Solve the preceding problem for the case of a 150-pound man two-thirds the way up the ladder. 210 lbs at 75°.

5. Is the ladder of the preceding problems more likely to slip when the man is at the bottom or when he is at the top? Why?

6. A beam of length l feet, whose weight is negligible, is supported at the ends. A concentrated load of A pounds is placed a feet from one end, B pounds b feet, and C pounds c feet from the same end. Find the reactions at the supports, in pounds.

l	A	a	B	b	C	c	D	E
12	100	3	500	7	200	10	320	480
12	300	4	800	8	100	11	480	730
10	200	3	500	5	600	8	510	790
16	500	4	300	9	200	14	530	470

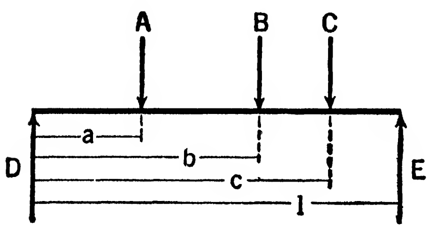


FIG. 62

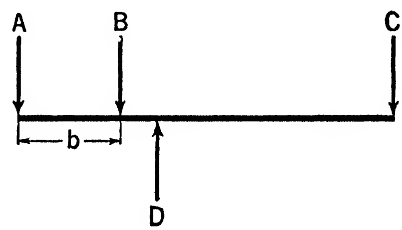


FIG. 63

7. A beam of length l feet, whose weight is negligible, is to be balanced on a single support. Concentrated loads of A and C pounds are placed at the ends, and another of B pounds placed at a distance b from A. Find the distance d from A in feet at which the support must be placed.

l	A	B	C	b	d
10	100	200	300	4	6.3
12	300	100	500	5	7.2
7	200	700	300	3	3.5
9	400	300	200	3	3.

8. Find the horizontal force F in pounds applied at the axle, necessary to raise a wheel of radius r inches and weight W pounds over an obstacle of height h inches.

	r	W	h	F
8.	5	10	2	13
	17	10	9	19
	13	10	8	24
	25	10	18	34

	l	B	A	a	α	F
9.	6	50	50	2	45	59
	8	60	40	3	30	90
	10	70	30	4	25	110
	12	80	20	5	20	140

9. A trapdoor of length l feet weighs B pounds. A load of A pounds is concentrated

at a feet from the hinged end. Find the force F which, acting at an angle α with the horizontal, is necessary to raise the door.

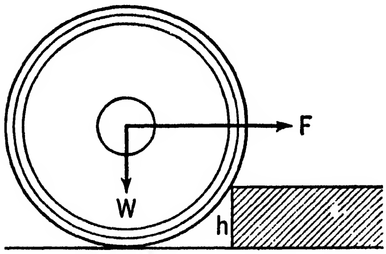


FIG. 64

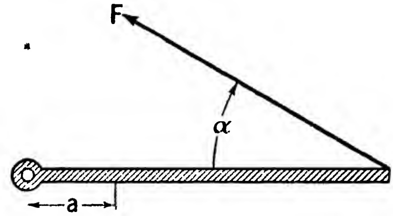


FIG. 65

10. A gate of weight W pounds, whose dimensions are a feet horizontally and b feet vertically, is supported by two hinges. The lower hinge can support only a horizontal thrust. The center of gravity is at the center of the gate. Find the magnitudes A and B of the reactions at the hinges, and the angle α which the reaction at the upper hinge makes with the horizontal.

W	a	b	A	B	α
50	4	4	25	56	63°
50	4	5	20	54	68°
50	3	7	11	51	78°
50	5	8	16	52	73°

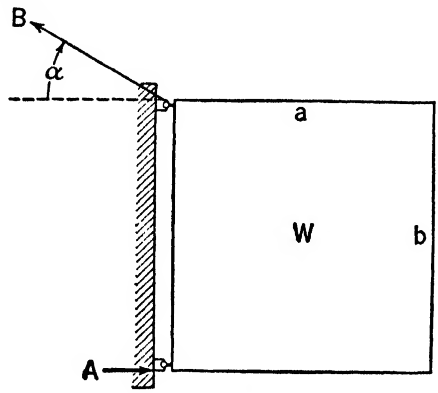


FIG. 66

11. To the upper edge of a box of weight W pounds, height b feet, and breadth a feet, a force is applied at an angle α degrees with the horizontal. What must be the magnitude F in pounds of the force in order to overturn the box?

	W	α	a	b	F
11.	200	10°	2	3	61
	200	30°	2	3	56
	200	50°	2	3	58
	200	70°	2	3	69

	W	l	a	k	α	P
12.	10	36	9	.2	22°	11
	10	36	9	.3	31°	12
	10	36	9	.4	39°	13
	10	36	9	.5	45°	14

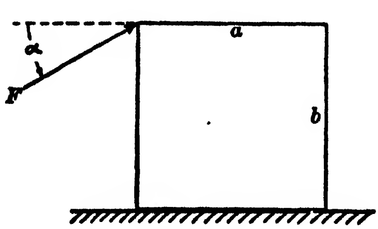


FIG. 67

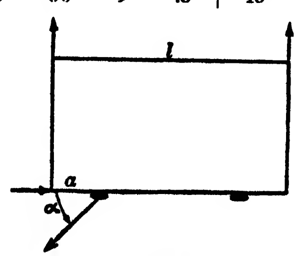


FIG. 68

12. A drawer of weight W pounds and width l inches slides upon its edges. It also experiences friction at the side if pulled obliquely. The coefficient of friction is k. The knobs are at a distance a inches from each edge. If the drawer is pulled out by a single knob, at what angle α in degrees with the front of the drawer must the force be exerted in order that the drawer may not "bind"? What force P in pounds must be exerted?

13. The foot of a uniform ladder of weight W pounds rests in a hollow, at a distance a feet from a vertical frictionless wall. The upper end rests against the wall at a distance b feet from the ground. Find the force B in pounds with which the wall supports the upper end, and the magnitude A in pounds and angle α in degrees (with the horizontal) with which the support at the lower end pushes up on the ladder.

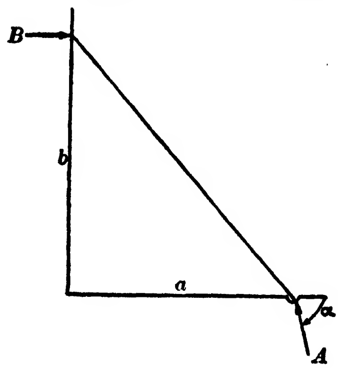


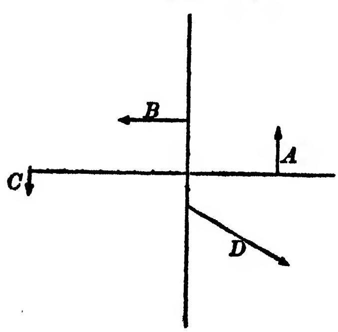
FIG. 69

W	a	b	A	B	α
100	12	5	160	120	40°
100	5	12	100	21	78°
100	6	8	110	38	69°
100	8	6	120	67	56°

14. A ladder of weight W pounds rests against a vertical wall. The foot is pulled out until the ladder is on the point of sliding. The coefficients of friction are k_1 and k_2 at the upper and lower ends respectively. What angle β does the ladder make with the horizontal when it is about to slide. What are the magnitudes F_1 and F_2 in pounds of the reactions at each end of the ladder?

W	k_1	k_2	β	F_1	F_2
50	.2	.6	30°	27	52
50	.3	.4	48°	21	48
50	.4	.3	56°	14	47
50	.6	.2	66°	10	46

15. Forces A, B, C, and D, applied at distances a from the center, act in the directions α upon the arms of a windlass as shown in the following diagrams. Find in each case the magnitude and lever arm of the *equilibrant* E, and the angle (measured counterclockwise) which the equilibrant makes with the arm to which the force A is applied.



	A	B	C	D	E
F	100	150	50	250	100
α	90°	180°	270°	330°	130°
a	5	3	9	2	18

FIG. 70

	A	*B*	*C*	*D*	*E*
F	60	80	40	50	120
α	270°	180°	90°	143° 8′	360°
a	4	6	7	5	2.0

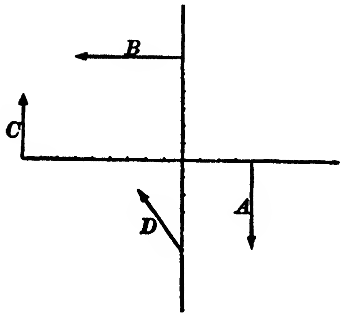


FIG. 71

	A	*B*	*C*	*D*	*E*
F	150	100	300	200	220
α	45°	150°	270°	20°	160°
a	5	7	3	8	16

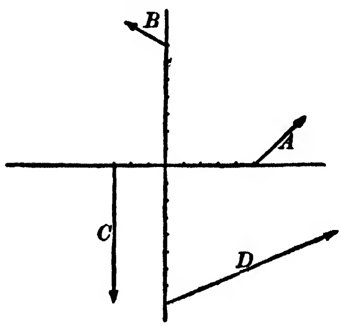


FIG. 72

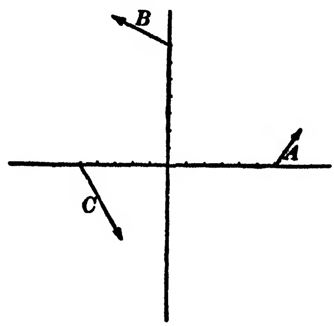


FIG. 73

	A	*B*	*C*	*E*
F	100	130	170	28
α	53° 8′	157° 23′	298° 4′	140°
a	6	7	5	73

16. In the crane of the accompanying diagram, you are given the weight W, the length b, and weight m of the uniform boom, the height h of the upright, the distance a from the ground at which the weight cable crosses the upright, and the length l of the cable which supports the boom. Required the angles α and γ, the tension T in the cable l, and the compression C in the boom.

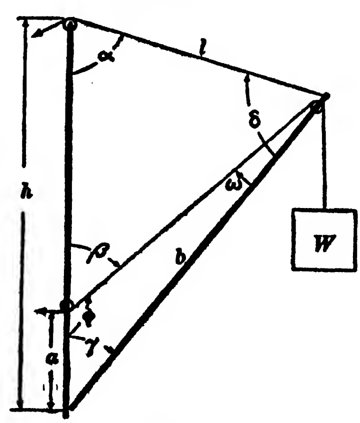


FIG. 74

w	*b*	*m*	*h*	*a*	*l*	*α*	*γ*	*T*	*C*
10	30	2	20	5	15	120°	26°	6.8	23
10	30	2	20	5	20	97°	41°	9.1	24
10	30	2	20	5	25	83°	56°	11.0	25
10	30	2	20	5	30	71°	71°	14.0	24

CHAPTER 7

Equilibrium: The Strength and Elasticity of Materials

Stress and Strain

When forces are applied to a body in such a way as to produce distortion, they are opposed by forces within the body itself, mainly cohesive and elastic. If, as distortion increases, the increasing internal forces ultimately come into equilibrium with the applied forces, the body will not rupture. To the extent that the internal forces succeed in bringing the body back to its original size or shape upon removal of the external forces, the body is said to be *elastic.* Though no substance is either perfectly elastic or perfectly inelastic, steel and rubber approximate to the former and putty to the latter.

If, on the other hand, the internal forces do not come into equilibrium with external forces effecting distortion of a body, rupture will eventuate. Until the rupture is on the point of occurring, the internal forces are conventionally taken as being in equilibrium with the external forces.

The internal force per unit of area across which the force is acting is called the *stress.* Thus, if a cast-iron column of cross-section 10 square inches supports a load of five tons, the stress is

$$\frac{\text{force}}{\text{area}} = \frac{10{,}000 \text{ lbs}}{10 \text{ sq in}}$$

$$= 1000 \text{ lbs per sq in.}$$

The relative distortion (change of length per unit original length; change of volume per unit original volume) is called the *strain.* Thus, if the foregoing column were 14 feet long and were shortened $\frac{1}{100}$ inch by the load which it supported, the strain would be

$$\frac{\text{change of length}}{\text{original length}} = \frac{.01}{168} = .00006 \text{ approximately.}$$

This use of the term *strain* to describe distortion should be noted. It is a technical term. Unfortunately, the same word in ordinary parlance involves the idea of force or exertion. It has no such connotation as used here. Though distortions or strains are produced by stresses, the word *strain* itself is *the relation of two distances* (or two volumes) and is entirely distinct from the forces or stresses that bring about the strains.

Kinds of Stress

At first glance there seems to be a limitless variety of combinations of forces that can be applied to a body. Almost all of them, however, fall into one of three classifications. They are (*a*) tensile and compressive stresses, (*b*) shearing stresses, (*c*) hydrostatic stresses (pressures).

A body is under tensile or compressive stress when the principal result is a change of length. The column cited in the previous section was, of course, under a compressive stress. A vertical wire supporting a weight is under tensile stress. The measure of any stress being force per unit area, the area in this type of stress is measured in a plane *perpendicular* to the line of the stress.

A body is under shearing stress when the principal result is a change of shape. A simple case is illustrated in Figure 75. The strain is characterized by the sliding of the layers of the material over one another. The name *shear* to describe this strain produces the imagery of the action of a pair of shears, which is entirely appropriate. Indeed, a common way of representing shear is by a pair of parallel semi-barbed arrows, as in Figure 76, precisely the action of shears on cloth, paper, or sheet metal.

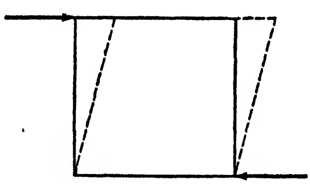


FIG. 75. A SIMPLE CASE OF SHEAR

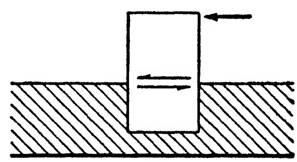


FIG. 76. CONVENTIONAL REPRESENTATION OF SHEAR

The sign of a shearing stress is conventionally determined by the direction of the *left* arrow: positive, if the left arrow points up; negative, if it points down, as in the figure. The area to be measured in computing the magnitude of a shearing stress is, unlike the case of tensile and compressive stresses, *taken parallel* to the line of the stress.

It is principally shearing stresses which rivets have to withstand. The total force acting on the riveted joint, divided by the total cross-sectional area of the rivets holding the joint together, gives a measure of the shearing stress involved. When such a joint gives way, it may be by shearing of the rivets or by the tearing of the plates between them, depending on whether the allowable shearing stress of the rivets or the allowable tensile stress of the plates is exceeded first.

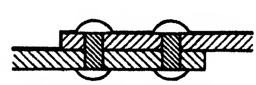


FIG. 77. A RIVETED JOINT

A body is under hydrostatic stress when the pressure upon it is the same from all sides. The accompanying strain is relative change of volume, as mentioned above in connection with the definition of strain. As the name of this type of stress indicates, it is encountered principally in connection with liquids and gases. Elastic stresses in liquids and gases are indeed limited to changes of volume. Neither of the preceding types of stress is at all applicable. Areas involved in computing hydrostatic stress must, however, be taken perpendicular to the line of the stress, as with tension and compression.

Combined Stresses

Stresses called into operation in everyday practice are of such manifold form that difficulty is often encountered in recognizing their simple components. Even the "simple components" are sometimes not exactly simple. For example, the shear illustrated in Figure 75, simple though the case is, involves an increase of length in the direction of the long diagonal of the distorted specimen and a decrease in length along the short diagonal. So both tension and compression are involved in even a simple case of shear.

Torsion, as in the action of a screwdriver, is simply a type of shearing stress in which successive layers of the material turn on one another instead of sliding on one another. When torsion is effected by means of a crank or a wrench, it is usually accompanied by a bending stress.

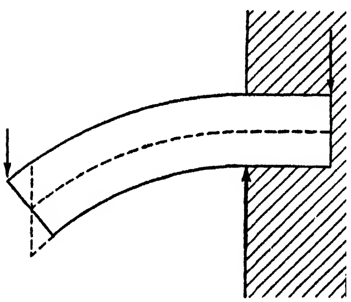


FIG. 78. TENSION AND COMPRESSION IN BENDING

Bending is so common that it merits special attention. Rather curiously, it is really only a special case of tension and compression. When a beam is bent, as in Figure 78, the upper surface is longer and the lower surface shorter than their lengths when the beam is undistorted. There will, however, be a portion of the beam at or near the center line which possesses its original length. This is termed the ***neutral axis*** characterized by zero stress in bending. The greatest stresses occur in the upper and lower surfaces of the beam. To meet this circumstance, steel beams designed to withstand bending are given a cross-section which may be described by the letter *I*, and are known, quite naturally, as ***I-beams*** (Fig. 79).

Shear in Bending

Usually bending stresses involve also some shear in addition to the tension and compression described above. Examination of the three vectors representing the forces which are in equilibrium in Figure 76 will make this clear, and Figure 80 will emphasize it still further. Even if there were no bending of the beam under the indicated load, there would still be in theory a shearing distortion, consisting of sliding of adjacent cross-sectional layers over one another. The total deflection of the beam is the sum of the shear and the bending. Actually, for beams ordinarily used, the deflection due to shear is negligibly small in comparison to that due to bending. Only in the case of extremely short projections do the two types of deflection assume the same order of magnitude.

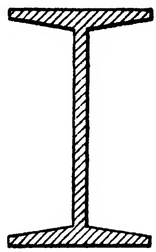


FIG. 79. CROSS-SECTION OF AN I-BEAM

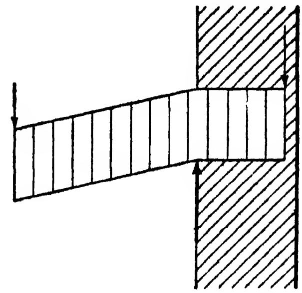


FIG. 80. SHEAR DISPLACEMENT OF A BEAM

Lest the reader assume that the foregoing remarks apply only to the type of beam illustrated in Figures 78 and 80, termed a *cantilever* beam, Figure 81 shows a beam supported at both ends and loaded in the middle. There is clearly shear in this case too, in addition to the bending. It is positive for the left half and negative for the right.

Though shear in beams is seldom of great importance taken by itself, it is useful as a means of identifying the so-called *dangerous sections* of loaded beams; that is, the points where failure of beams will occur in case of overload and where, in any case, the tensile and compressive stresses are a maximum.

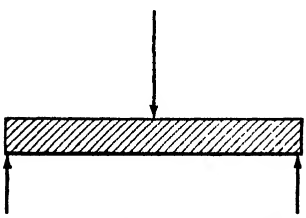


FIG. 81. WEIGHTLESS BEAM SUPPORTED AT THE ENDS

Rather curiously, it is at these sections of a beam where the shear has zero value that the maximum tensions and compressions occur in bending, and therefore where a beam fails if loaded beyond the breaking strength of the material. The method of identifying such sections is as follows.

Shear Diagrams

If the magnitude of the shearing stresses be plotted as ordinates against positions along a beam as abscissas, the resulting curve is called a *shear diagram*. The case of Figure 78 is a simple one — almost too simple to be illustrative. Disregarding the weight of the beam, the measure of the shear along the entire exposed length is simply the load at the end.[1] Thus, the shear diagram for this case is simply a horizontal straight line up to the point where the beam enters the wall. There another force, directed upward, comes

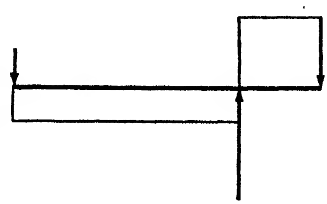


FIG. 82. SHEAR DIAGRAM FOR THE BEAM OF FIG. 78, ASSUMED WEIGHTLESS

[1] Shear being represented by a pair of equal and oppositely directed vectors (see Figure 76), it is customary to take the left-hand vector as the measure of its magnitude.

into play. This is greater than the force at the end, since, by the principle of equilibrium of forces, it must equal the sum of the two downward forces, one at each end. The resulting shear diagram is shown in Figure 82. Ordinary mechanical experience tells one that the place where such a beam would break would be at the point where it enters the wall. It is to be noted that this is the point of zero shear.

Failure of the beam of Figure 81 would clearly occur at the middle. Figure 83 shows the shear diagram for this case, and again the dangerous section is seen to be associated with the position of zero shear.

The shear diagram presents a different appearance if a beam is uniformly loaded instead of having a concentrated load of such magnitude that the weight of the beam may be disregarded in comparison with it. Thus, for the uniformly loaded cantilever, the shear diagram would be as in Figure 84, and for the uniformly loaded beam supported at the ends, the shear diagram would be as in Figure 85. The dangerous sections have still the same location.

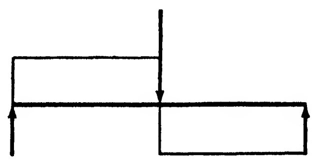


FIG. 83. SHEAR DIAGRAM FOR THE BEAM OF FIG. 81

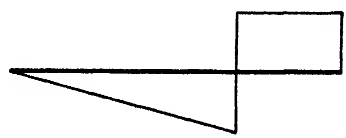


FIG. 84. SHEAR DIAGRAM FOR UNIFORMLY LOADED CANTILEVER

The combination of a concentrated load combined with one uniformly distributed is illustrated in Figure 86. This time one's intuition is not as reliable as in the two previous cases in indicating where the dangerous section is. The point of zero shear, however, gives the desired information.

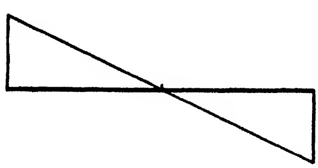


FIG. 85. SHEAR DIAGRAM OF UNIFORMLY LOADED BEAM, SUPPORTED AT THE ENDS

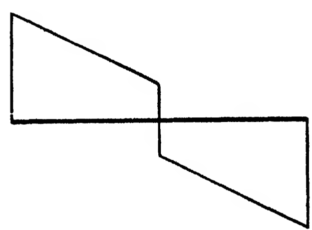


FIG. 86. SHEAR DIAGRAM OF BEAM UNIFORMLY LOADED PLUS CONCENTRATED LOAD AT CENTER, SUPPORTED AT ENDS

The case of distributed but non-uniform loads can be similarly treated. As before, the shear for each point is given by the total load to the left of that point. When constructing the shear diagram, downward forces are negative and upward forces (supports) positive.

Hooke's Law

If an elastic specimen is placed under stress, it will, of course, be deformed, and in general the greater the stress, the greater the resulting strain. Nearly three centuries ago Robert Hooke (1635–1703) surmised that the strain would be found proportional to the stress and verified the surmise by experiment. He announced his discovery, to quote:

> *Ut tensio, sic vis;* that is, the power of any spring is in the same proportion with the extension thereof; that is, if one power stretch or bend it one space, two will bend it two, and three will bend it three, and so forward. Now, as the theory is very short, so the way of trying it is very easie. [56.]

Hooke's law applies to any elastic body. More properly speaking, a body is said to be elastic only within the limits that it obeys Hooke's law. If a specimen is stretched or compressed or twisted or bent beyond the point where there is a strict proportionality between strain and the stress producing it, the specimen is said to have been deformed beyond the *elastic limit.* A permanent "set" results. The strains produced in most structural members and mechanical devices are, however, within the elastic limits of the materials of which they are made. Hence, proportionality between strain and stress is the rule. That is, Hooke's law governs most distortions of materials used in everyday experience.

The Modulus of Elasticity

Whenever a proportion is encountered, the constant of proportionality is likely to be found of considerable importance. Thus, in the case of uniform motion, where distance traveled was proportional to time elapsed, the constant of proportionality was the velocity. That is,

$$s \propto t \quad \text{or} \quad s = vt. \tag{1}$$

Similarly, for free fall, where velocity attained was proportional to time elapsed, the constant of proportionality was the acceleration of gravity. That is,

$$v \propto t \quad \text{or} \quad v = gt. \tag{2}$$

Hooke's law also states a proportionality; namely, stress is proportional to strain; always, of course, within the elastic limit. That is,

$$\text{stress} \propto \text{strain} = \text{some constant} \times \text{strain}. \tag{3}$$

This constant of proportionality is called the *modulus of elasticity* of the material involved. Equation (3) is really the defining equation for the modulus of elasticity of any material. Thus,

$$\text{modulus} = \frac{\text{stress}}{\text{strain}}. \tag{4}$$

The modulus of elasticity assumes a different aspect in each of the three types of stress described on page 65. That associated with tensile and compressive stresses is generally called *Young's modulus*, after Thomas Young (134), who coined the term *modulus* as used in this connection. That associated with shear is termed the *rigidity modulus* and that associated with hydrostatic stress is termed *volume modulus*.

Young's modulus for cast iron may be deduced from the data on the cast-iron column given on page 64:

$$\text{modulus} = \frac{\text{stress}}{\text{strain}} = \frac{10{,}000/10}{.01/168} = 1.68 \cdot 10^7 \, \frac{\text{lbs}}{\text{in}^2}.$$

Conversely, given Young's modulus, one could compute the compression or extension of a given specimen when the load and the dimensions were known. Similarly, shearing strains and volume changes could be computed with the aid of tables giving rigidity and volume moduli respectively. (See Appendix.)

Elasticity of Gases

The most resilient materials known are gases. Between the elastic behavior of gases and that of solids and liquids there are points of similarity as well as points of difference that will repay examination.

One of the central features of the behavior of gases is formulated under the name *Boyle's law*, formulated by Robert Boyle (1627–91), a contemporary of Hooke. In a book published in 1660, which labored under the cumbersome title, *A Defense of the Doctrine Touching the Spring and Weight of the Air*, Boyle set forth the relation between the pressure and volume of air. Everyone who has had experience pumping up a tire is well aware that the application of pressure reduces the volume of air. Boyle's law states the relation between pressure and volume of a given mass of air as that of inverse proportionality, that is,

$$p \propto \frac{1}{v}. \tag{5}$$

Introducing a constant of proportionality, c, to substitute an equality for the proportionality

$$p = \frac{c}{v}, \text{ whence } pv = c. \tag{6}$$

To verify this hypothesis "that supposes the pressures and expansions to be in reciprocal proportion," Boyle made use of a manometer, one end of which was closed, mercury being poured into the open end. The air entrapped in the closed side was compressed by the weight of mercury. Boyle (77:85)

> continued this pouring in of quicksilver till the air in the shorter leg was by condensation reduced to take up by half the space it possessed (I say pos-

sessed, not filled) before; we cast our eyes upon the longer leg of the glass on which was likewise pasted a list of paper, carefully divided into inches and parts, and we observed, not without delight and satisfaction that the quicksilver in that longer part of the tube was 29 inches higher than the other. For . . . the air was (previously) able to counterbalance and resist the pressure of a mercurial cylinder of about 29 inches, as we are taught by Torricellian experiment; so here the same air being brought to a degree of density about twice as great as it had before, obtains a spring (of 29 additional inches) twice as strong as formerly.

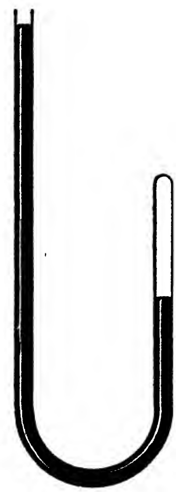


FIG. 87. MANOMETER OF THE TYPE USED BY BOYLE

A Stricture on Boyle's Law

A quarter of a century after Boyle stated his law, Edme Mariotte (1620–84) made one qualification which is important, and which Boyle, though he must have known and allowed for it in his experimentation, had not explicitly mentioned. Mariotte remarked (77:88), rather incidentally, that

> The air also dilates very easily by heat and condenses by cold, as we can notice any day by experiment.

Thus, the pressure of an enclosed gas will fluctuate with temperature, besides the changes in pressure that may attend any changes of volume that are brought about. Unless changes of temperature are guarded against, the inverse proportionality between pressure and volume in a gas discovered by Boyle and by Mariotte will not be accurately registered. In other words, *Boyle's law obtains only for constant temperatures.* The actual study of the effect of changing temperature on gas was, rather surprisingly, not made for more than a century.

It is common to represent Boyle's law graphically. If values of p in equation (6) be represented as ordinates with v as abscissas for a given value of c, a curve similar to Figure 88 results. This curve has the form of an hyperbola. Boyle's law was first represented in 1686 in this way by Edmund Halley, the friend who published Newton's *Principia* at his own expense.

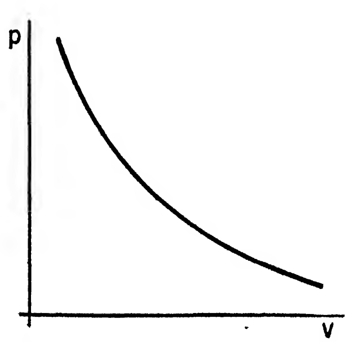


FIG. 88. GRAPHICAL REPRESENTATION OF BOYLE'S LAW

Other Limitations on Boyle's Law

It is appropriate to inquire into the limits of the validity of Boyle's law. Is pressure inversely proportional to volume for all

FIG. 89. ROBERT BOYLE

gases, through all ranges of pressure, and at all temperatures? A part of the answer is to be found in data which Boyle himself took. After his first manometer had been broken by accident, he made another, the open side of which was more than eight feet in length. The pressures in this tube were found to mount more rapidly than the inverse proportion allowed for. Subsequent observers have verified this fact. The discrepancy vindicates Boyle's law, however, rather than invalidating it; this for two reasons. First, the effective, unoccupied volume of a container is not its entire interior, but is less than the entire interior by the actual volume occupied by the molecules of the gas. The free space which the molecules have for their motion is thus smaller than the actual volume of the container. Second, the external pressure is not the only agency holding the gas molecules together. There is an attraction between the molecules themselves which aids the external pressure in this respect. When allowance is made for these two effects, as is done when the so-called Van der Waals' equation is substituted for Boyle's law, the discrepancies shown by the data of Boyle and his successors largely disappear. Boyle's law is sometimes said to apply rigorously only to "perfect" gases; that is, gases whose molecules have no volume and which do not exert any forces on each other except during the instants of actual impact.

These two disturbing factors become prominent in a gas as the temperature approaches the point of liquefaction. In fact the liquid state obtains when the forces of attraction between molecules are great enough so that no external pressure is required to prevent expansion. Hence, Boyle's law applies the most closely to gases whose "boiling points" — that is, temperatures of transition from the liquid to the gaseous state — are very low, such as hydrogen, nitrogen, oxygen, and the so-called noble gases, helium, argon, etc., and applies much less closely to steam, to the gases used in mechanical refrigerators, or to any gas not at a temperature far above the point of liquefaction.

Why Are Gases Compressible?

There is no serious attempt on the part of any of the earlier writers to account for the compressibility peculiar to gases. Aside from some poorly conceived speculations to the effect that the atoms of gases experienced mutual repulsions, a view of pressure as a purely static phenomenon, the first attempt to explain the behavior of gases on a simple mechanical basis was made by Daniel Bernoulli (1700–82). Bernoulli's approach was along an entirely new line, in which gas pressure was viewed as the result of myriads of atomic impacts against the walls of the container. This point of view regarded pressure as a phenomenon of dynamic equilibrium rather than of static equilibrium as the repulsion theory had done. The idea was so novel that it was not taken seriously for more than a century. Ultimately it developed into one of the major chapters of modern physics, the kinetic theory of gases.

Bernoulli's approach is significant enough to warrant a closer view. In his *Hydrodynamics*, published in 1738, is found the following presentation (30:220; 77:248):

> Imagine a vertical cylindrical vessel (Fig. 90) in which fits a movable piston *EF* upon which there is placed a weight *P*. Let the enclosure contain some very small particles moving hither and thither with extreme rapidity. It is these corpuscles which impinge upon the piston and which by their continually repeated impulses sustain it that constitute the elastic fluid; a fluid which expands whenever the weight *P* is removed or diminished (and) which is compressed whenever the weight is increased.

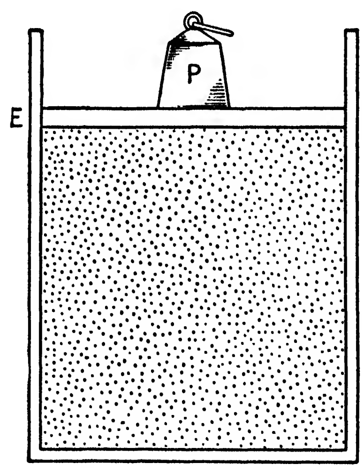


FIG. 90. BERNOULLI'S THEORY OF GAS PRESSURE

Bernoulli formulated algebraically the relation between the action of these flying particles and the external steady pressure which would be produced by them and deduced that

> the compressing weights are almost in the inverse ratio of the spaces which air occupies when compressed by different amounts.... And the temperature of the air while it is being compressed must be carefully kept constant

> ... since it is admitted that heat may be considered as increasing internal motion of the particles.

This is an almost complete prevision of a stage of physics which was not actually reached for one hundred and ten years.

Volume Modulus of a Gas

It will be worth while to compare the elastic properties of gases, as thus established by Boyle and elaborated by Bernoulli, with those of solids and liquids. There is a certain similarity which is indeed more than superficial, as was pointed out on page 66. But there is also an important difference. Briefly it is that, though elastic moduli of solids and liquids are constant within the elastic limits of the materials involved, the volume modulus of a gas, on the contrary, changes with pressure. Indeed, within the limits of applicability of Boyle's law, the volume modulus *is* the pressure.

The foregoing statement requires some elucidation. Suppose, for example, that one pound added to the load on a steel wire is found to extend the wire .01 inch. The next pound will produce the same extension, and so on indefinitely until the elastic limit is reached. Contrast this with the behavior of a gas. Increasing the pressure on one cubic foot of air, initially at 14.7 pounds per square inch (atmospheric pressure), by another 14.7 pounds per square inch will diminish the volume by one half a cubic foot in accordance with Boyle's law. Increasing the pressure by another 14.7 pounds per square inch will not produce the same diminution of volume, as would result if the modulus were a constant, but will instead reduce the volume to one third the original volume in accordance with Boyle's law. In this way, beginning with the first measurement of pressure, the successive equal pressure increases produce volume diminutions of one half, one sixth, one twelfth, one twentieth, etc., of the original volume. This means that the volume modulus of the air is steadily increasing. That the successive values of the modulus are proportional to, and indeed numerically equal to, the pressure may be shown as follows. Suppose that a change of pressure Δp [1] produces a change of volume Δv. Then the volume modulus is, by equation (4),

$$\text{modulus} = \frac{\text{stress}}{\text{strain}} = \frac{\Delta p}{\Delta v/v}. \tag{7}$$

But from Boyle's law, equation (6),

$$pv = c, \text{ and also}$$
$$(p + \overline{\Delta p})\ (v - \overline{\Delta v}) = c. \tag{8}$$

[1] Small increments of a quantity are often represented by the Greek capital letter Δ (delta). In this case Δp means not the product of the two quantities Δ and p as ordinary algebraic notation would require, but is to be interpreted, "a small change of pressure." The Δ and the p together constitute a single symbol and can no more be separated than a script d can be cut into two parts, one of them being an a, by lopping off the upper projection.

Expand equation (8)

$$pv - p\,\overline{\Delta v} + v\,\overline{\Delta p} - \overline{\Delta p}\,\overline{\Delta v} = c.$$

Subtract $pv = c$ and disregard the product $\overline{\Delta p}\ \overline{\Delta v}$ as being negligibly small in comparison with the other terms in the equation. Whence,

$$-\,p\,\Delta v + v\,\Delta p = 0;$$

or

$$\frac{\Delta p}{\Delta v/v} = p. \tag{9}$$

But by equation (7), the expression $\dfrac{\Delta p}{\Delta v/v}$ is the expression for the volume modulus of the gas. Hence the volume modulus of a gas *is* the pressure of that gas. It consequently is not a constant as is the case for solids and liquids, but varies even within the elastic limit. This identity of volume modulus with pressure for a gas will be of considerable significance when the velocity of sound in a gas comes under scrutiny.

Questions for Self-Examination

1. Distinguish between stress and strain.
2. Describe three types of stress and give an example of each.
3. What is the "dangerous section" of a loaded beam?
4. Set up a loaded beam and draw its shear diagram.
5. State Hooke's law and its limitations.
6. Classify moduli of elasticity and give the general definition of the term.
7. State Boyle's law and its limitations.
8. Discuss the volume modulus of gases.

Problems on Chapter 7

1. A rod of circular cross-section, diameter d inches, supports W tons. What is the tensile stress T in pounds per square inch? Consult the Appendix to determine the cases in which the rod, if of steel, would not hold the load.

	d	W	T		a	b	W	S
1.	.50	10	100,000	2.	2	4	4	1000
	1.0	20	51,000		4	2	4	1000
	2.0	30	20,000		2	10	8	800
	3.0	40	11,000		10	2	8	800

2. A rectangular beam, with cross-sectional dimensions a by b inches, rests upon two supports and carries W tons midway between them. Find the shearing stress S in pounds per square inch.

3. A wrought-iron bolt d inches in diameter is tested to destruction by applying W tons as shown. The head is t inches thick. What are the two ways in which the bolt can fail? Calculate the shearing and tensile loads (S and T) which the bolt can withstand. Which type of failure will occur? Take the tensile strength

of wrought iron as 60,000 and shearing strength as 40,000 pounds per square inch.

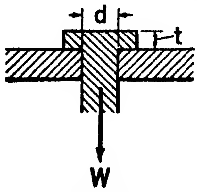


FIG. 91

d	*W*	*t*	*S*	*T*
.50	20	.5	31,000	47,000
.75	50	1.	94,000	110,000
1.00	100	1.75	220,000	190,000
1.25	160	2.25	350,000	290,000

4. A circular shaft d inches in diameter is twisted. Two sections l feet apart suffer a relative displacement of θ degrees. What is the shearing strain S at the surface of the shaft?

d	*l*	θ	*S*
2.	10	4°	.00055
2 5	12	3°	.00045
3.	14	2°	.00031
3.5	16	1°	.00016

5. These struts all have a cross-section A inches square. The inclined members make θ degrees with the horizontal. What is the tensile stress T in the horizontal member when a load W pounds is applied at the top? If the wood has a maximum allowable shearing stress of S lbs/in² parallel to the grain, what is the closest distance d that the notches may come to the end of the horizontal member?

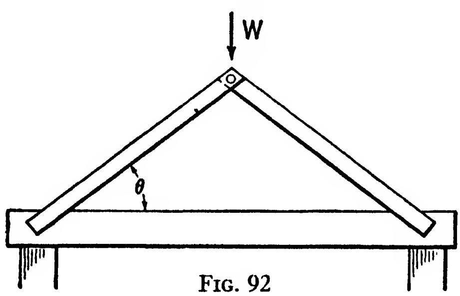


FIG. 92

A	θ	*W*	*S*	*T*	*d*
5	60°	1200	400	350	.17
4	50°	1000	600	420	.18
3	40°	800	800	480	.20
2	30°	600	1000	520	.43

6. A riveted joint of the type shown in Figure 93 joins two iron plates t inches thick. The holes and rivets are d inches in diameter. The distance between centers of rivets is p inches. The allowable tensile stress in the plates is 60,000 lbs/in² and the allowable shearing stress in the rivets is 40,000 lbs/in². Will the joint fail by shearing of the rivets or by tearing the plates between the rivets? What force per linear inch will produce such failure?

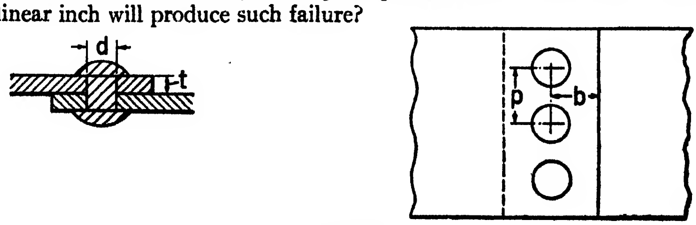


FIG. 93

	t	*d*	*p*	*F*		*t*	*d*	*p*
6.	.250	.500	1.10	7,100	7.	.250	.5	1.0
	.375	.625	1.20	10,000		.375	.625	1.2
	.500	.750	1.30	13,000		.500	.7500	1.3
	.625	.875	1.40	14,000		.625	.875	1.5

7. In the riveted joint of problem 6, what must be the spacing of the rivets (between centers) if the joint is equally likely to fail by shearing of rivets and tearing of plates between rivets?

8. In problem 6, what is the closest allowable distance *b* of centers of rivets to edge of plate if the joint is just not to fail by shearing of the plate? Figure 94 is a drawing of a plate that has failed this way. Take the allowable shearing stress of the plate as 40,000 lbs/in².

t	*p*	*F*	*b*
.250	1.10	7,140	.39
.375	1.20	10,225	.41
.500	1.30	12,680	.41
.625	1.40	14,050	.49

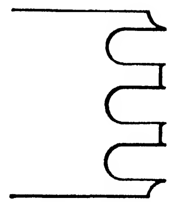


Fig. 94

9. In problem 1 what will be the elongation *e* in inches of the rod if it is initially *l* feet long? The material is steel with Young's modulus 3.2×10^7 pounds per square inch.

d	*W*	*l*	*e*
.50	10	5	—
1.0	20	10	.20
2.0	30	15	.11
3.0	40	20	.085

10. A rectangular beam, with cross-sectional dimensions *a* by *b* inches, rests upon two supports at its ends and carries *W* pounds at the middle. What will be the deflection d_s of the beam due to shear if it is made of oak and is *l* inches long? Compare with the deflection d_b due to bending, computed from the relation

$$d_b = \frac{Wl^3}{4\,Yab^3},$$

where *Y* is Young's modulus in pounds per square inch and *b* is the vertical dimension in inches. For oak, Young's modulus is 1.5×10^6 and the rigidity modulus is $.7 \times 10^6$, both in pounds per square inch.

a	*b*	*W*	*l*	d_s	d_b
2	4	1000	60	.011	.28
2	4	1000	30	.0054	.035
2	4	1000	12	.0021	.0022
4	2	1000	30	.0054	.14

11. The formula of problem 10 is much used in the determination of Young's modulus. Calculate Young's modulus in pounds per square inch for the following materials from the tabulated observations on rectangular beams made of the respective materials.

material	*a*	*b*	*l*	*W*	*d*	*Y*
Oak............	2.	4.	5	1000	.280	1.5×10^6
Copper.........	.5	.5	1	500	.232	15×10^6
Cast iron.......	2.	1.	2	1000	.102	17×10^6
Steel...........	1.	1.	3	1000	.389	30×10^6

12. The intake valves of one of the cylinders of a gas compressor working isothermally are set to open at p pounds per square inch. The exhaust valves open at P pounds per square inch. What fraction f of the stroke has been completed when the exhaust valves open?

	p	P	f
12.	14.7	75	.80
	75.	350	.79
	350.	1200	.71
	1200.	3000	.60

	d	s	l
13.	100	1.00014	.007
	1,000	1.0014	.720
	10,560	1.015	80.
	36,960	1.053	980.

13. The volume modulus of water is 3×10^5 pounds per square inch. What is the specific gravity s of water at depth d feet under the compression of the weight of the water above? For a body of water having an average depth d feet, how much lower l is the surface in feet than it would have been if water had been completely incompressible? Assume as a first approximation that the density of water does not change with depth. You can then correct the preliminary values of s for the error of this assumption. (*Note:* One may take the average depth of all the oceans as two miles.)

14. Compute the reactions at the supports, C and D, for each of the beams shown in Figure 95. Then construct shear diagrams and determine how many feet x from the left end the dangerous sections are located. All beams are 12 feet long, weigh 100 pounds per foot, and have concentrated loads A and B distant a and b feet respectively from the left end. The supports are located respectively c and d feet from the left end.

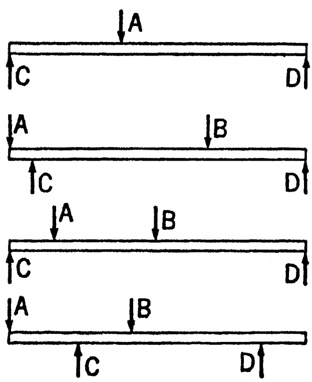


FIG. 95

A	a	B	b	c	d	C	D	x_1	x_2	x_3
400	3			0	12	900	700	5.		
1000	0	300	8	1	12	1900	650	1.	8.5	
1500	2	200	5	0	12	2000	930	4.7		
1000	0	500	7	3	10	2300	370	3.	8.3	10

15. Calculate the value E of the volume modulus of air for each set of data in problem 12, thereby showing that it is numerically equal to the pressure. (Assume a one-pound change in pressure of an arbitrarily chosen volume at each given pressure, calculate the corresponding change in volume and apply equation (9).) Why do the first two pairs of values agree less well than the others?

p	E
14.7	16.
75.	76.
350.	350.
1200.	1200.
3000.	3000.

CHAPTER 8

Equilibrium in Fluids

Archimedes' Principle

The utility of the concept of equilibrium is not confined to solids. It applies at least as effectively to fluids, a term which includes both liquids and gases. In fact it was applied to liquids nearly two thousand years before it took definite shape in the mechanics of solids. The first formulation of the principle of equilibrium in liquids was made by Archimedes. The tradition is that it burst upon him as he was submerging himself in a full bathtub and that in the ecstasy of his discovery he raced nude to his study, crying, "Eureka (I have found it)."

What is known as *Archimedes' principle* is contained in the following paraphrase of his own statement of it (5:257–58):

> A body immersed in a fluid is buoyed up with a force equal to the weight of fluid displaced.

The principle is sometimes verified by a very elementary experiment, using a device first attributed to Al Biruni (973–1048) (25:19). A vessel with a spout near the top to facilitate the catching of overflow, is just filled with some liquid, and a solid object is lowered into it. The overflow is then caught and weighed. If the solid object floats, the weight of the overflow will equal the weight of the object. If it sinks, the weight of the overflow will equal the loss of weight which the object experiences when lowered into the liquid. In either case, the object is buoyed up with a force equal to the weight of the liquid displaced, which verifies Archimedes' principle.

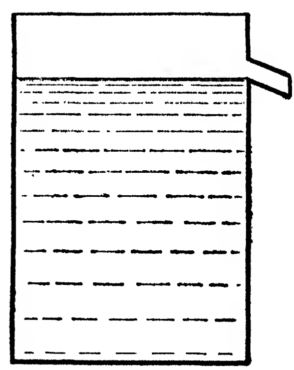


FIG. 96. AL BIRUNI'S SPECIFIC GRAVITY VESSEL

The Measurement of Density

One of the modern uses of Archimedes' principle is in measuring the volumes of irregular solids. When such a solid is immersed in a liquid contained in a graduated vessel, the rise in the level of the surface gives a measure of the volume of the submerged body. This becomes particularly

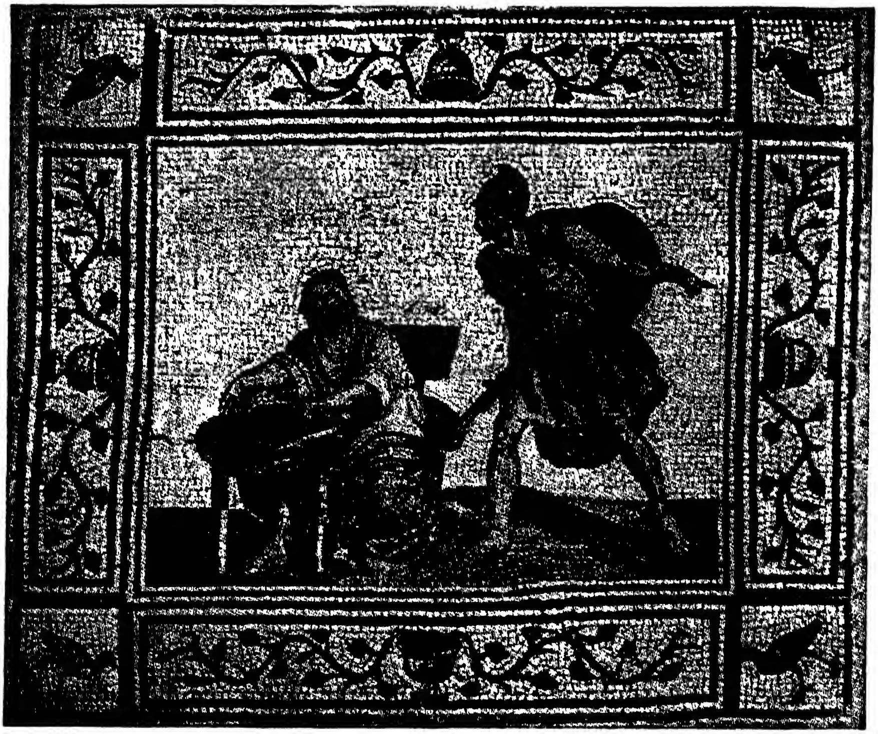


FIG. 97. THE DEATH OF ARCHIMEDES IN 212 B.C.
A mosaic found in the ruins of Pompeii, and which must therefore have been constructed before the year 70. (Courtesy of *Scripta Mathematica.*)

useful in determinations of ***densities*** of such objects as jewels and precious stones, the volumes of which cannot readily be measured otherwise.

The concept of density is a very simple one. The fundamental nature of the intuition of density is shown in the ease with which one may be trapped by the old catch question, "Which is heavier, a pound of lead or a pound of feathers?" Like all intuitions, however, it requires careful definition before it can be used as a scientific concept. The modern definition of density is ***mass per unit volume***, or, stated algebraically,

$$d = \frac{m}{v}. \tag{1}$$

The utility of Archimedes' principle in measurements of density of solids is thus evident in the very definition. Archimedes' principle is even more useful in this connection than is evident at first sight. This is because refinements in the balance have made precision measurements of weight[1] more satisfactory than precision measurements of volume. The common

[1] For present purposes weight may be taken as a measure of mass. The distinction between the two will be developed in Chapter 9.

procedure, therefore, when the density of an irregular solid is in question, is to find its weight m in air (strictly in a vacuum, but the difference, not ordinarily important, can be corrected for if necessary) and then its weight w in some liquid of known density D. The difference in weights is, by Archimedes' principle, the weight of the liquid displaced by the solid; in other words, the weight of liquid equal in volume to the solid. The volume of this liquid, and hence the volume of the solid, is therefore

$$v = \frac{m - w}{D}. \tag{2}$$

The density of the solid then becomes calculable by equation (1).

Specific Gravity

The units in which density is specified will depend on circumstances. It is common to say that the density of water is 62.4 pounds per cubic foot. It is also 1000 kilograms (one metric ton) per cubic meter, from the original definition of the gram, as the mass of one cubic centimeter of water. Similarly, the density of iron, for example, is 7860 kgm/m^3, of ice 916 kgm/m^3, of ethyl alcohol 793 kgm/m^3, and of air at normal temperature and pressure 1.29 kgm/m^3.[1] It is more common, however, to specify density in terms of the density of water. Thus, a cubic meter of iron weighs 7.86 times the same volume of water. The cumbersome term *specific gravity* was coined in the twelfth century to describe the ratio of the density of any substance to the density of water, and has persisted. Since it expresses a ratio, no units are involved. The specific gravity of ice is thus .916; of alcohol, .793; and of air, .00129. The utility of Archimedes' principle in the determination of specific gravities of solids and liquids should be even more evident than in the determination of densities, since it involves directly the ratio of weights in air (or rather in a vacuum) and in water or other liquid.

Pascal's Principle

Another characteristic of fluids, now known as Pascal's principle, apparently eluded Archimedes, though it was implied in his principle. It is commonly stated as follows:

> Pressure exerted at any place on a fluid in a closed vessel is transmitted undiminished throughout the fluid and acts at right angles to all surfaces.

Though these are not the words in which the principle was first stated, they constitute a reasonable modern version of the original phraseology. Pascal

[1] Strictly speaking, the numerical values of densities in the metric system will be on this scale. In this book, however, densities will usually be stated in kilograms per liter instead of kilograms per cubic meter. See, for example, the table of densities in the Appendix. This possesses the minor advantage that such familiar quantities as the C.G.S. unit of density of water (1 g/ml) has the same numerical value when stated in kg/l. But for actual calculations requiring absolute values, as in dynamical relations, the strict M.K.S. unit of density (kg/m^3) must, of course, be used.

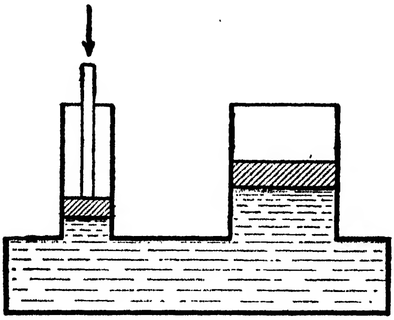


FIG. 98. PASCAL'S HYDRAULIC PRESS (77:76)

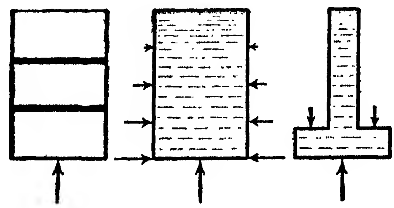


FIG. 99. THE HYDROSTATIC PARADOX

enunciated the principle, in a somewhat verbose way, in connection with his description of a device which later became an important industrial tool, the hydraulic press (patented in 1795 by Joseph Bramah).

Pascal's principle is, as he put it, merely a corollary of the "continuity and fluidity of the water." It is, indeed, the principal distinction between fluids and solids that solids possess elasticity of both size and shape, whereas fluids exhibit elasticity of shape, but not of size. Fluids resist any effort to compress them, but not to change merely their shape. From this follows the transmission of pressure in all directions; from this follows also the inability to resist or transmit tangential forces at the boundaries. Consequently fluid pressure must be perpendicular to the surface.

Pascal purposely disregarded the effect of the weight of the water. It will be illuminating to consider it, notwithstanding. If bricks were piled one above the other to a total weight of 100 pounds, a 100-pound vertical force at the bottom would be all that would be required to support the pile. This force would normally be distributed over the area of the bottom brick, requiring an average pressure, say, of about three pounds per square inch. But if 100 pounds of water were contained in a tank having the same area at the bottom, the contents would experience other forces besides those applied at the bottom. The sides of the tank would also have to withstand horizontal pressures, varying uniformly from three pounds per square inch at portions adjacent to the bottom to zero at the top. But if the tank, while still possessing the same area at the bottom, were shaped like an inverted T, so that it contained a much smaller weight of water, and the bottom were a movable, watertight piston, the pressure at the bottom, being three pounds per square inch, would still require a total upward force of 100 pounds to support it, as the hydraulic-press analysis of Pascal makes clear. The generalization of this apparently anomalous circumstance is commonly known as the *hydrostatic paradox*. If the vessels should be placed in turn on a balance, the weight of the liquid contents would, of course, be quite different in the two cases, even though the force exerted by the liquid on the bottom of each container were the same. Therein lies the "paradox," which, however, can be resolved by a little thought. The paradox is sometimes illustrated by a set of so-called *Pascal's vases*, glass containers of a variety of shapes and sizes, communicating with each other at the bottom. The equality of the heights of the liquid in all of these vessels gives evidence of the equality of the pressures at the bottoms.

The Open Manometer

The equality of pressures at the bottoms of communicating vessels of liquid is sometimes utilized in the determination of the relative densities of two liquids. Suppose, for example, that a glass U-tube (in effect a pair of "Pascal's vases") contained water in one side and mercury in the other. The two liquids would not stand at the same height, but, by virtue of the equality of the pressures at the bottom, would stand at heights inversely proportional to their densities. Thus, a column of water would stand 13.6 times as high as a column of mercury (Fig. 100). In the same way it is possible to compare the densities of any two liquids which do not mix. The method depends on the pressure equilibrium at the bottom between the two columns of liquid in the U-tube, or *manometer* as such a device is termed. The manometer is in fact primarily a pressure-measuring instrument. It is used quite commonly in comparing pressures of two bodies of gas, especially when they are at relatively low pressures, as the name indicates. (The first syllable, *man-*, stems from a Greek root meaning *rare* or *thin.*)

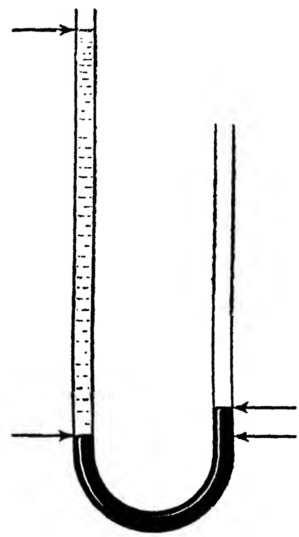


FIG. 100. RELATIVE HEIGHTS OF WATER AND MERCURY IN EQUILIBRIUM

The Mercurial Barometer

Figure 101 shows how a manometer may be used to measure the pressure exerted by the atmosphere. Mercury is usually used for this purpose. When the air is removed from one side of the manometer by a vacuum pump, the pressure of the atmosphere on the other side forces the mercury up in the evacuated side until it is balanced by the height of this column of mercury. Such a balance occurs when the difference of the heights of the mercury on the two sides is between 72 and 76 centimeters under conditions usually obtaining (weather and altitude). When used in this way, the manometer is termed a *barometer*, meaning pressure gauge.

The first barometer on record was made in 1644 by Evangelista Torricelli, who had been a pupil of Galileo. Its importance lies not so much in the device itself as in the fact that it was an outgrowth of the first correct

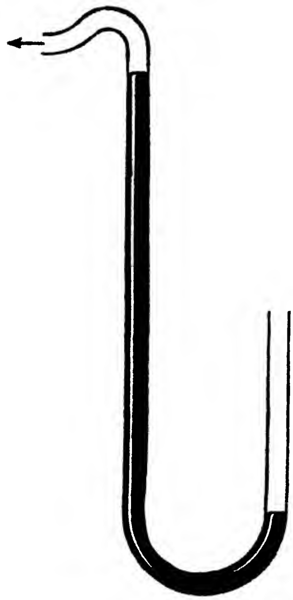


FIG. 101. USING A MANOMETER TO MEASURE ATMOSPHERIC PRESSURE

FIG. 102. AN EARLY ILLUSTRATION OF THE CONSTRUCTION AND USE OF A TORRICELLI TUBE
(Courtesy of Taylor Instrument Companies.)

view of the nature of atmospheric pressure and of a vacuum. This was a recognition of the fact that

> We live immersed at the bottom of a sea of elemental air, which by experiment undoubtedly has weight, and so much weight that the densest air in the neighborhood of the surface of the earth weighs about one four-hundredth part of the weight of water.

This is a rather badly stated (as well as inaccurate) comparison of the densities of air and water. But it was nevertheless a new and very productive idea. The principle of the modern barometer was involved in Torricelli's early devices, two of which are pictured in Figure 103.

Pascal's Study of the Vacuum

In 1647, Pascal took up the investigation where Torricelli had left it. It had, in the meantime, become a highly controversial subject. The controversy originated in the insistence of the Scholastics on the Aristotelian doctrine of the practical and logical impossibility of a vacuum. The rallying cry of this school of thought was "*Natura abhorret vacuum.*" Even Galileo had been only mildly ironical about this Aristotelian doctrine. Having observed, prior to the Torricellian experiment, that (46:16)

> it was not possible, either by a pump or by any other machine working on

> the principle of attraction to lift water a hair's breadth above eighteen cubits,

he is said to have remarked that Nature's abhorrence of a vacuum seemed to be limited to eighteen cubits of water, but to have gone little farther with the problem.

Having become convinced of the correctness of Torricelli's conclusions, Pascal leaped into the controversy, invoking a series of striking experimental demonstrations to support his point of view. He was fortunate in living in Rouen, the site of one of the best glassworks in Europe, because he was thus enabled to conduct experiments which would otherwise have been impossible. He supplemented these with a number of other ingenious and sensational demonstrations, including the use of a plunger-type syringe as a vacuum pump. This was the first time a mechanical pump had ever been used to produce a vacuum. It furnished Otto von Guericke with a tool which he used most effectively a quarter of a century later. Pascal capped his contributions to the controversy by his famous experiment on an adjacent mountain. The project was to carry a barometer to the top of Puy-de-Dôme and observe any attendant fluctuations in the height of the mercury column.

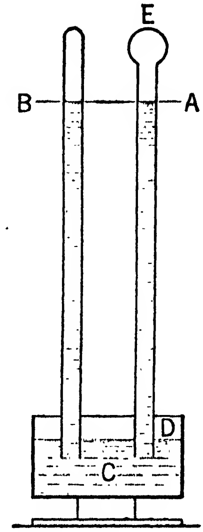


FIG. 103. TORRICELLI'S BAROMETERS (77:71)

This anticipated a common use made of the barometer today, namely, its use in measuring altitude. He foresaw this, and commented on the possibility. In addition, he observed the correlation between fluctuations of barometer height and subsequent weather conditions and remarked (20:671):

> This knowledge can be very useful to farmers, travelers, etc., to learn the present state of the weather, and that which is to follow immediately.

Questions for Self-Examination

1. State Archimedes' principle, define density, and use the two to solve this problem: A piece of metal weighs 200 pounds in air, 100 pounds in water and 105 pounds in oil. Find the density of metal and oil in lbs/cu ft. Take for the density of water the approximate value 60 lbs/cu ft.
2. Discuss the stability of floating bodies.
3. What is the "hydrostatic paradox"?
4. Mention some episodes in the evolution of the barometer.
5. What was the main point at issue between Pascal and his detractors?
6. Summarize the contributions to hydrostatics of Torricelli, Pascal, Galileo.

Problems on Chapter 8

1. Ice floats in fresh water with .08 of its volume above water. What is its specific gravity? If an iceberg floats with .1 of its volume out of water, what is the specific gravity of the sea water? .92. 1.02.

2. A piece of metal weighs 1.6 kilograms in air and 1.4 kilograms in water. A piece of wood weighs .8 kilogram in air and the two fastened together weigh .6 kilogram in water. Find the specific gravity of each specimen. 8. .5.

3. A diving helmet rests on the shoulders of a diver. It is of the type used by amateurs for a few minutes' submersion only and has no air supply. If the diver's nostrils are one third the distance from the open bottom of the helmet to the top, what is the greatest depth attainable? 17 ft.

4. Air is forced into a tank by an air pump having a maximum inside volume half that of the tank. How many strokes will be required to pump it up to a pressure of three atmospheres, assuming that Boyle's law applies and that the process begins with air in the tank at atmospheric pressure? 4.

5. How much lower should the barometer read at an elevation of 500 meters than at sea level, taking the average density of air as 1.2 kilograms per cubic meter? 4.4 cms.

6. A metal ball weighs M grams in air, m grams when immersed in water, and w grams when immersed in oil. What is the specific gravity D of the ball and d of the oil?

M	m	w	D	d
200	100	105	2	.95
500	400	410	5	.90
800	700	715	8	.85
1100	1000	1020	11	.80

7. A U-tube is partly filled with water. Oil is poured into one side until it stands h centimeters above the water level on the other side, which has meantime risen w centimeters. What is the specific gravity d of the oil?

h	w	d
5	47.5	.95
10	45.	.90
15	42.5	.85
20	40.	.80

8. What pull P_1 in kilograms does a balloon of volume v cubic meters and weight m kilograms exert on its guy ropes when inflated with hydrogen of specific gravity .0695 (referred to air)? What pull P_2 does it exert if filled with helium of specific gravity .1370? Take the density of air as 1.2 kilograms per cubic meter.

v	m	P_1	P_2
500	250	310	270
1000	350	770	690
1500	450	1200	1100
2000	550	1700	1500

9. What is the error e in grams due to neglecting the buoyancy of the air in weighing an object of mass m grams and specific gravity d on a beam balance? (Assume the density of the brass weights to be 8 grams per cubic centimeter, and that of the air to be .0012.)

m	d	e
20	.2	−.12
200	2.	−.09
800	8.	0
2000	20.	+.18

10. How many meters h would a liquid of specific gravity d rise in an evacuated tube, atmospheric pressure being p grams per square centimeter?

d	p	h
1.	1000	10.
5.	1000	2.
10.	1000	1.
13.6	1000	.74

11. What must be the height h in feet of the level of water in a tank, to produce a pressure of p pounds per square inch on its base? (Assume the density of water to be 62.4 pounds per cubic foot.)

p	h
40	92
80	180
100	230
150	350

CHAPTER 9

Weight and Mass

Weight is Variable

In the experiments described on page 48 the observers were relieved of the inconvenience of maintaining the forces under study through their own muscular exertions by enlisting gravity in their aid. The downward pull of the earth on some metal disks termed *weights* furnished the required forces. Let us suppose, now, that the force table were set above a very deep well, far deeper than even the deepest mine shaft, and that into this well one of the disks were lowered. It would presently be observed, as the disk was lowered to progressively greater depths, that the former condition of equilibrium on the force table no longer obtained.[1] The effect would be as though weight were slowly being removed from the descending pan.[2]

The difference in gravitational attraction on an object as it is raised or lowered to different altitudes is far more than the hair-splitting matter that it is sometimes felt to be. The change in the weight of a one-pound object as it is raised from floor to ceiling of an ordinary room would be detectable with a good-quality analytical balance such as is in common use in all physics and chemistry laboratories. Moreover, extensive use is made of fluctuations in weight with change of location in the process of geophysical prospecting for oil and other minerals. For this purpose a Hungarian inventor named Eötvos designed a special type of balance which bears his name; though exceedingly delicate, it is nevertheless capable of being used in the field by prospecting parties.

The reason that gravitational attraction on an object fluctuates with changes in altitude is not far to seek. Consider the extreme case of the imaginary disk under observation being lowered to the center of the earth. For this purpose the well may be supposed to have been sunk to that depth. There would now be no downward gravitational pull on it whatever, for the attraction due to the body of the earth, which now surrounds the disk on all sides, is uniformly distributed in all directions; hence, there is no component directed toward the center of the earth as before.

[1] An experiment of this kind was actually carried out in 1662 and a report made to the Royal Society of London in December of that year (19:*1*:133).

[2] An effect due to the non-homogeneity of the interior of the earth would produce a slight increase in gravitational attraction for a short distance below the surface. This effect, experimentally demonstrated in 1885 by Sir George Airy, is here disregarded.

The First Observation of Variability of Weight

Let us return to the disk which had lost its weight in consequence of its descent to the center of the earth. It may be noted that a similar diminution of gravitational attraction would have been observed if the disk had been raised to great heights above the earth. Indeed, small fluctuations of the sort may be observed without leaving the earth's surface. On account of the equatorial bulge, with consequent variation in the distance of various portions of the surface of the earth from the center, there exist variations similar to the foregoing with mere changes in latitude. The effect was observed first in 1671, when Jean Richer went from Paris to Cayenne in French Guiana to make astronomical observations. He found that his pendulum clock, a new invention by Christiaan Huygens, which in Paris had been adjusted to keep correct time, ran slow in the new locality by two and one half minutes a day. He was, of course, entirely at a loss to account for it until, after his return, Huygens pointed out that the rate of a pendulum clock would be affected by alterations in the gravitational attraction of the earth on the pendulum.

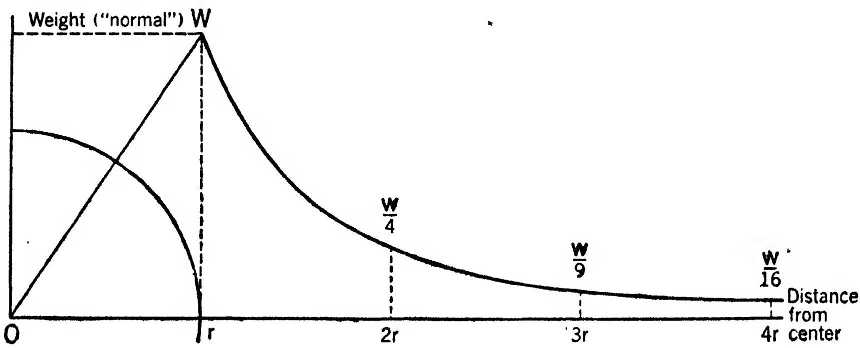


FIG. 104. HOW WEIGHT DIMINISHES ABOVE AND BELOW THE SURFACE OF THE EARTH

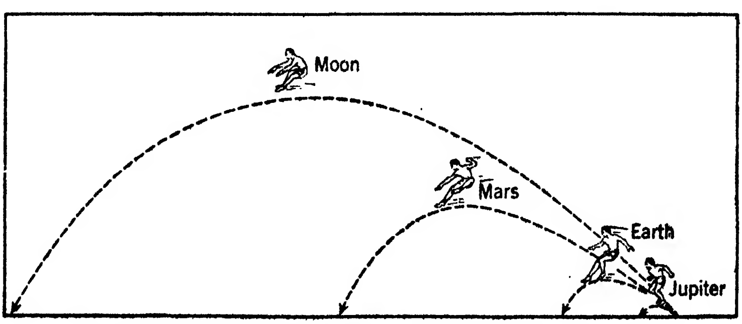


FIG. 105. DISTANCES OF JUMP ON DIFFERENT MEMBERS OF THE SOLAR SYSTEM

The important thing to observe in all this is that weight, which is the name given to the mutual gravitational force between the earth and any object under consideration, does not have the fixed, invariable value sometimes ascribed to it by the unthinking. An object weighing one pound at the surface of the earth will weigh zero at the center and substantially zero if taken to the far reaches of astronomical space. On the contrary, at the surface of a dense star, such as the small companion of Sirius, it would weigh several tons. Any statement that an object weighs so many pounds always carries the qualification, usually not stated, but present nevertheless, limiting its validity to the surface of the earth, and even then is somewhat uncertain until latitude or altitude is specified.

The Concept of Mass

Notwithstanding these unquestionable fluctuations in weight, it is impossible for us to disregard a robust intuition that there are, after all, no corresponding fluctuations in what has been somewhat ineptly termed the *quantity of matter.* The hundred-gram metal disk did not lose its substance when it was at the center of the earth merely because its weight became zero. It could neither have shrunk into nothingness nor faded into some kind of incorporeal ghost of its former self. In a way which is at first hard to formulate, we are confident that something about that disk besides its volume remains unchanged through all fluctuations of weight. Such intuitions merit careful study. They are often misconceptions which require careful revision. But occasionally, as in the present case, they are fundamental verities which, after clarification, may be made to do good scientific service. In either case they cannot be accepted at their face value, but must be painstakingly analyzed.

Galileo struggled with this seeming paradox in a somewhat confused manner, but never really solved it. Descartes tried his hand at it without success. Not until Isaac Newton (1642–1726) formulated his famous laws of motion was the problem clarified. This seems quite amazing in view of the fact that the solution is one which the majority of individuals with normal powers of observation carry around vaguely as an intuition from childhood.

Newton gave the name *quantity of matter* to the entity which is conceived of as remaining unchanged through the fluctuations of weight described above. The name has proven unfortunate and his definition of the concept still more so, as will presently be seen. Indeed, he himself introduced the term *inertia* to replace it in the later portions of his *Principia.* Nowadays the term *mass* is most commonly used to express this idea.

The distinction between mass and weight is unquestionably the most important distinction in the early history of physics. That the terms should frequently be confused by the layman is not surprising, especially in view of the common use of the same units, such as ounces, pounds, and tons, to express the two, an exceedingly unfortunate practice. Moreover, clear as the difference now appears to the initiated, there must be an element of subtlety in any important physical concept which eluded the minds of all the great scientific thinkers up to the time of Kepler. Nevertheless, the distinction between mass and weight is entirely straightforward and logically simple. It is essentially contained in the statement that, whereas weight is the force of gravitational attraction upon an object, mass is the inertia or sluggishness which an object, when frictionlessly mounted, exhibits in response to any effort made to start it or to stop it or to change in any other way its state of motion.

If, for example, the suspended disk which was used to illustrate fluctuations of weight had been tapped horizontally with a hammer when at the top of the well where its weight was normal, and again in a direction parallel to the first blow when at the center of the earth where its weight was zero, provided that the speed with which the hammer struck was the same in both cases, what would be the ratio of the second resulting speed of the disk to the first? The experiment can, of course, never be carried out in the form described, but the entire accumulation of mechanical knowledge necessitates the belief that the resulting speed of the disk would be the same in the two cases, notwithstanding the fact that the weights were different, being what would be termed normal in one case and zero in the other.

The Measurement of Weight and Mass

Weight is simply the name given to the gravitational force exerted by the earth on an object in question. Measurements of weight are thus simply measurements of force. Forces may be compared with the aid of balances. The spring variety is useful but inaccurate. The beam balance is more commonly used to compare weights. It has the great advantage that it automatically corrects itself for fluctuations in gravitational attraction, since these act alike on counterbalancing weights and the objects being weighed.

The real justification for the assumption that comparison of weights will give accurate measures of comparisons of mass must, of course, be sought in experiment. The assumption has, in fact, been fully vindicated in that way, subject to the condition that it is only in the locality where the weight measurement was actually made that the mass may be assumed proportional thereto.

However convenient the proportionality of weight and mass may be from the standpoint of practical convenience, from the standpoint of facilitating clarity of thought on the principles of mechanics it has been a catastrophe. Most of the confusion between the concepts of weight and mass, concepts which are, in fact, logically entirely disparate, has been occasioned by the accident of their numerical proportionality. This has been a stumbling block from the time that man first began to give rational consideration to the physical phenomena of nature. While mass is proportional to weight for any given locality, it must be remembered that a mass has the same value at all places,[1] whereas a weight may vary through wide limits, depending upon locality.

The real distinction between weight and mass came with the work of Isaac Newton, culminating in his three famous Laws of Motion. So basic are these laws that it is commonly considered that one of them, the second, constitutes the definition of mass, if force be considered the more fundamental, or of force, if mass be considered the more fundamental. The usual choice between these alternatives will be seen (p. 96) to be to regard mass as the fundamental quantity and to use the relation between mass and acceleration, which constitutes Newton's second law of motion, to define the unit of force.

Questions for Self-Examination

1. State some circumstances, real and imaginary, under which weight changes without any change in mass.
2. Recount the experience of Jean Richer in 1671 with one of the first pendulum clocks.
3. Distinguish between mass and weight and give examples of your distinction.

[1] Implications here and elsewhere of the invariability of mass take no account of the variability postulated in the theory of relativity. That and the associated concepts are scarcely appropriate at this stage.

CHAPTER 10

Mass and Acceleration

A Classical Experiment on Force and Motion

Three concepts, among others, have received attention in the foregoing chapters; namely, acceleration, force, and mass. Very little has been said up to this point about possible relations that they might bear to each other, though it has been noted that Galileo's work involved an implication that the force of gravitational attraction was responsible for the observed uniform acceleration of falling bodies.

Attention may now be given to the general problem of the kinds of motion produced by forces, the magnitudes of which are varying in certain ways. The first and simplest of such problems is that of the motion produced by a constant force. The most direct way to approach this is through experiment, using a machine first devised by George Atwood, of Cambridge University, prior to 1780 (7b).

The principle of Atwood's machine is illustrated in Figure 106. Two equal masses are hung over a pulley which is as near massless and frictionless as it is possible to make it. For the unavoidable small elements of mass and friction there are simple ways of compensating. As long as the two masses are evenly balanced, gravitational attraction will be neutralized as far as the production of any motion is concerned. If an initial motion, say upward, is imparted to one mass, it will continue to rise in the absence of friction and the other to descend with undiminished speed until the end of the course of travel is reached. This is in accordance with the principle which Galileo applied to the horizontal part of the motion of a projectile: namely, that such motion would remain uniform in the absence of any force opposing it. If, however, the equilibrium of the two masses were destroyed by adding or removing weight on one side, the motion which would, of course, ensue would be of whatever character is to be expected from the action of a constant force. It is this type, motion under constant force, which will now be examined.

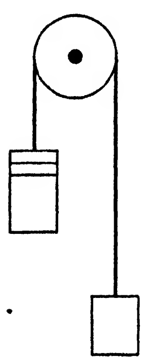


Fig. 106. The Principle of Atwood's Machine

The study of this motion is greatly facilitated by a rapid-fire recording device indicating the distance traveled by the masses in successive equal intervals of time. A modern way of effecting this is to cause accurately timed electric sparks to jump from one of the masses to the frame of the machine, making perforations in a strip of paper placed there for the purpose. The distances between successive perforations will then be the desired distances of travel.

Motion under Constant Force is Uniformly Accelerated

Suppose, for example, that each of the two masses weighs one kilogram and that one tenth of a kilogram is removed from one of the masses and placed on the other. Suppose also that the time-recorder produces five sparks per second and that the distances, measured from the beginning, traveled in the intervals one fifth, two fifths, three fifths, etc., of a second by the descending mass, are subsequently measured and found to be as follows:

time (in fifths of a second)	0	1	2	3	4	5	6
distance (in centimeters)	0	1.0^{-}	3.9	8.8	15.7	24.5	35.3
distance (relative)	0	1	4	9	16	25	36

The figures of the bottom row are intended to facilitate comparison of distances. Inspection of the table shows that distances are in proportion to the squares of the elapsed time-intervals, which identifies this as the same type of motion studied by Galileo, namely, uniformly accelerated motion. Hence, the first deduction to be made from this experiment is that ***motion under a constant force is uniformly accelerated motion.***

This gives an enormous amount of information. For with the discovery that this is uniformly accelerated motion goes the knowledge that equations (1) to (6) of Chapter 3 are applicable, provided only that for g (the acceleration characteristic of free fall under gravity) there be substituted another term, such as a, which will be free of any implication that it possesses the value 9.8 m/sec, or that it is necessarily directed downward. Since y carries an implication of vertical motion, it would perhaps be well to change that term also to s, a common notation for space or distance traveled, without any commitment as to direction. For although, in the present experiment, the acceleration happens to be in a vertical direction (even, indeed, vertically downward as in free fall), this is due to some purely incidental features of the machine with which the experiment is performed. A corresponding experiment in which masses move only horizontally is, in fact, often performed instead of this one. The six equations applying to the present case, corresponding to equations (1) to (6) of Chapter 3, may therefore be written as follows:

$v = at$	(1)	$v = v_0 + at$	(4)
$s = \frac{1}{2}\, at^2$	(2)	$s = v_0t + \frac{1}{2}\, at^2$	(5)
$v = \sqrt{2\, as}$	(3)	$v = \sqrt{v^2_0 + 2\, as}$	(6)

Acceleration is Proportional to Force

But even with this seven-league step, the principal objective of the present experiment is far from realized. Though equations (1) to (6) describe thoroughly the motion which is characteristic of uniform acceleration, they give no indication of how this motion is related to the force which produces it. Some information about this relation is already at hand, however. Motion characterized by uniform acceleration has been found to result from the application of a constant force. The next point on which to inquire is what happens when the experiment is repeated, again with a constant force, but of different magnitude than before. Suppose that this time two tenths of a kilogram, twice as much weight as before, is transferred from one mass to the other, both masses having been originally in equilibrium as in the first experiment. The resulting observations will now be as follows:

time (in fifths of a second)	0	1	2	3	4	5
distance (in centimeters)	0	2.0⁻	7.8	17.6	31.4	49.0
relative distance	0	2	8	18	32	50
relative distance (previous expt.)	0	1	4	9	16	25

The bottom row gives the relative distances from the previous experiment for comparison. Inspection of the table shows that in this case, as before, the acceleration is uniform, though of different magnitude, and that each distance in the present experiment is twice the corresponding distance in the previous experiment. This suggests that the acceleration of the masses in this experiment has twice the value of that in the former experiment. If this is true, then the following important conclusion may be formulated: *The acceleration is proportional to the applied force.*

The establishment of the fact that the response of a movable object to the application of force is in the form of *acceleration* possesses an historical and scientific significance which it would be hard to overrate. Prior to the scientific era, nobody seems to have imagined such a thing, and its discovery was symbolic, perhaps more than any other single circumstance, of the transition in the study of motion from the days of arid qualitative speculation to those of mathematical and scientific development. Intuitively it often seems more natural to associate force with *speed* than with acceleration. Most individuals who are without training in the physical sciences will identify "motion" with force, and if the point be pressed, will express the opinion that the greater the force, the more rapid the motion. There is, indeed, much in the accumulation of unanalyzed daily experience by the average individual to suggest such an impression, but it is one of the most insidious misapprehensions in the history of science. The change in emphasis which Newton effected, from *speed* to *acceleration* as a consequence of the application of force, inaugurated an entirely new and highly productive line of thought. In fact, it was one of the principal manifestations of the onset of the scientific era.

Acceleration is Inversely Proportional to Mass

But to return to Atwood's machine. The information which it can be made to yield is as yet by no means exhausted. In fact, it has still to make its most important contribution. Thus far it has disclosed that motion under a constant force is motion with uniform acceleration and that *for a given mass* this acceleration is proportional to the force. It remains to discover the effect of changing the mass. Some prevision of what the result of the experiment is likely to be may be attained by recalling the characterization of mass in the previous chapter as "the inertia or sluggishness which an object exhibits in response to any effort to change its state of motion." Hence, if the force remains unchanged, there is reason to suspect that acceleration will increase as mass diminishes, and *vice versa*. This suspicion is promptly verified. If half-kilograms are substituted for the one-kilogram masses and then one tenth of a kilogram is transferred from one side to the other as in the first experiment, the following measurements result.

time (in fifths of a second)	0	1	2	3	4	5
distance (in centimeters)	0	2.0⁻	7.8	17.6	31.4	49.0
distance (first expt.)	0	1.0⁻	3.9	8.8	15.7	24.5

Each of the distances is seen to be double the corresponding one of the first experiment. Hence the conclusion that the *acceleration is inversely proportional to the mass.*

The Constant of Proportionality

All of the foregoing conclusions may be summarized in the statement that motion under a constant force consists of a uniform acceleration which is directly proportional to the force and inversely proportional to the mass. Formulated in algebraic terms this is

$$a \propto \frac{f}{m}. \tag{7}$$

Clear of fractions and convert this into an equation instead of a proportionality by introducing a proportionality factor k. Then

$$f = kma, \tag{8}$$

k being an unknown constant, the value of which is yet to be determined.

The value of k will obviously depend upon the units in which the measurements are made. Usually the units are so chosen as to give k the value unity; whence equation (8) becomes

$$f = ma. \tag{9}$$

This is effected by selecting for f units which are quite different from the pound or the kilogram heretofore used. Their values may be deduced by performing, in imagination, a simple experiment.

Confining attention at first to the metric system, consider a mass of one kilogram to be accelerating at the rate of 1 meter per second^2. Substituting the value unity for m in equation (8), (1 kg) and for a (1 m/sec^2) and setting $k = 1$ as specified above, the value of f turns out to be unity also. That is, the unit of force selected in the metric system is *that force required to impart to one kilogram an acceleration of one meter per second per second.* This is a new unit, which has been named the *newton.*[1] The corresponding English unit is *that force required to impart to one pound an acceleration of one foot per second per second.* It is called the *poundal.*

The relation of the newton to the kilogram (regarded as a unit of force in Chapter 4 and subsequently) may be seen by considering a case of free fall. A mass of one kilogram falling freely under gravity will have an acceleration of 9.8; whence from equation (9), $f = 1 \cdot 9.8 = 9.8$ newtons, the force of gravity on it. But the force of gravity on a mass of one kilogram is commonly said to be one kilogram. Hence, the magnitude of one newton is 1/9.8 times the force of gravity on a one-kilogram mass. Similarly, in the English system, the magnitude of one poundal is 1/32.2 times the force of gravity on a one-pound mass.

Newton's Second Law of Motion

Equation (9) is an algebraic formulation of what is commonly known as Newton's second law of motion. It formed one of the foundation stones

[1] The use of the meter, kilogram, and newton as units of length, mass, and force respectively, follows a practice which is gaining ground in scientific circles. It is a departure from the older practice of using the centimeter, gram, and dyne — the dyne being defined as the force required to impart an acceleration of 1 cm/sec^2 to a mass of 1 gram. The two objections to the older system are (1) that the practical units of length and mass (the centimeter and gram respectively) are derivatives of the fundamental standards established by international agreement, whereas it would be better practice to make actual use of the fundamental standards; and (2) that the older system creates an exceedingly awkward threefold system of electrical units which can be avoided only by abandoning the mechanical units which give rise to it. The system used here is free from both of these objections, and its sole disadvantage is that it has not yet come into common use. The International Committee on Weights and Measures decreed in 1935 that this system should go into effect in January, 1940, and it is being very generally put into effect in spite of the second World War.

of his famous *Principia.* A free translation of it, expressed in modern scientific terminology, is:

> Rate of change of momentum is proportional to the unbalanced force and is in the direction of that force.

Momentum (more explicitly presented in Chapter 13) is the product of mass by velocity. Rate of change of momentum is then the product of mass by rate of change of velocity. But rate of change of velocity has already received the name acceleration. Hence "rate of change of momentum" is simply the product *ma* appearing in equations (8) and (9). These equations are, therefore, algebraic statements of Newton's second law of motion.

Newton's First Law of Motion

The foregoing *second* law of motion of Newton was preceded by his *first* law, which, in translation, is as follows:

> *Law I.* 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.

It was this law of motion which Galileo anticipated in his treatment of the horizontal component of the motion of projectiles. As has been noted, it contravened the mechanical notions of all his predecessors, not to mention the preconceptions with which all of us, even today, seem to approach the study of mechanics. Galileo apparently did not realize how basic it really was, or he would have developed the concept more fully. This, however, was left for others to do. Newton was not the first to state the principle explicitly and to sense its importance, but he *was* the first to incorporate it into a comprehensive set of laws of motion, the essential validity of which is amply confirmed by the scientific structure which has been reared upon them.

Newton's First Law as a Special Case of the Second

When Newton's first law is viewed against the background of the second, it will at once become evident that the first is simply a special case of the second. That is, if the force is zero, the acceleration is also necessarily zero, and the speed is therefore constant. It is perhaps not so evident that the motion will necessarily be in a straight line also. It has, however, already been observed (p. 14) that a given velocity, as distinct from speed, is not completely specified unless the direction as well as the magnitude is named. Development of the idea that change of direction is just as much a change of velocity as is change of magnitude of speed and that acceleration and force are involved in it will occur in a later chapter (p. 183). In the meantime, reliance will have to be placed tentatively upon the mere plausibility of the statement that any alteration of motion whatever, whether of speed or of direction, requires the application of a force, and that, hence, in the absence of any force, neither kind of alteration will occur.

Since Newton's first law is thus merely a special case of the second, the question may be raised as to why it is separately stated at all. Is it not merely a redundancy? The answer to this question is clearly in the affirmative. The second law having once been stated, the first is definitely a redundancy and is not logically required. Nevertheless, it is almost always retained and explicitly stated as one of the separate laws of motion. Aside from its being an inescapable historical fact that it was in this form that Newton stated his laws, an added reason for the retention of the first law could justifiably be that it possesses some value to the beginner in science. Though it describes a special case of the second law, the case is *so* special and requires so much consideration, and the consideration of it is withal so profitable, that there is distinct advantage in isolating and studying it in its own right. However, the first law could have been stated more logically as a corollary to the second law, which it is, than as an independent law, which it most certainly is not.

Newton's Third Law of Motion

Newton's statement of his third law reads thus in his *Principia* (91:13):

> *Law III.* To every action there is always opposed an equal reaction; or the mutual actions of two bodies upon each other are always equal and directed to contrary parts.

His use of the new and undefined terms *action* and *reaction* need cause no difficulty. For the present they will be construed as referring to forces, which demonstrably act always in pairs.

A homely illustration may help to establish some preliminary insight into Newton's third law of motion. Almost every child has spent uncounted hours with what is somewhat whimsically termed an express wagon. A favorite pursuit with this toy is to imagine that this wagon is some self-propelled vehicle such as an automobile or a locomotive. The illusion of self-propulsion is fostered by the fact that the child is actually in the moving wagon, though the motion is maintained by the contact of a projecting leg or stick with the ground. Though the motion is forward, the muscular effort producing it is directed toward the rear. If this be termed the action, the desired forward force which accompanies it may be classified as the reaction.

The Oppositeness of the Reaction to the Action

This illustration is chosen, in spite of its apparent triviality, because it is the prototype of all engines of transportation, without exception. If anyone doubts that a force to the rear is exerted by the engines of, say, an ocean liner through the screws, let him consider the motion of the water in the wake; or let him stand in the blast of air behind an airplane which is on the point of taking off; or let him observe the shower of mud from the spinning wheels of an automobile which is trying to get out of a mud-hole. Even the act of walking consists of successive backward pushes by the feet on the ground, the reactions to which produce forward motion of the pedestrian. The elementary fact of a rearward aspect to the forces applied in all such cases is sometimes concealed by the very complication of modern engines of transportation, but to uncover it does not require profound powers of discrimination.

The phenomenon of action and reaction in such cases as the foregoing is sometimes illustrated by mounting the track of a toy electric or spring-driven locomotive so that the track, as well as the locomotive, is free to move. It will be seen that the track takes up motion in the direction opposite to that taken up by the locomotive. The "kick" of a gun, with its larger-scale counterpart in the recoil of heavy cannon — to absorb the effects of which it is necessary to make definite mechanical provision — is a further illustration of the same phenomenon.

The Equality of the Reaction to the Action

It is to be noted that Newton's third law states both the oppositeness and the equality of the two aspects of a force, the action and the reaction. The only conclusion that it is permissible to draw from the foregoing illustrations, however, at least until they have been more carefully analyzed, is that action and reaction are merely opposite in direction. For evidence as to their equality in magnitude further search must be made. In a limited way, the experiments of Chapters 4 and 5 on equilibrium of forces furnished an illustration of the equality of action and reaction for the case in which change of motion is not involved; that is, in which acceleration is zero. It was found in the laboratory that in every case tested the resultant of two or more forces was not only opposite in direction to the force required to balance the system, but also that its magnitude was equal to the balancing force within the limits of experimental error.

The important aspect of Newton's third law is the case in which the forces are statically unbalanced, the state of unbalance being necessarily accompanied by an acceleration not present in the case described in the preceding paragraph. Any assertion of the equality of action and reaction for this case can be made only on a foundation of experiment. Atwood's machine can be made to furnish such a foundation. Consider the tension in the cord of such a machine (Fig. 107). Unless there is friction at the wheel, the tension of the cord must be the same on both sides, one to be regarded as the action, the other as the reaction. At first sight, however, the action seems to be the weight of the larger mass and the reaction that of the smaller. How then, it may be objected, is it possible to assert the equality of the action and the reaction? And yet the tension in the cord must be the same at the point where it is attached to the large mass as at the point where it is attached to the small mass.

The dilemma may be resolved by observing that the reaction is not merely the weight of the small mass, as erroneously assumed, but that to this weight, expressed in force units, must be added the product of the small mass by its acceleration; so the reaction is seen to be somewhat larger than the mere weight of the small mass. It is the sum of these two forces which becomes equal to the tension in the cord. Unless this tension had been greater than the weight of the small mass, no acceleration would have been produced. Similarly, on the other side of the machine, the action is not the weight of the large mass, but is less than this weight by the product of the large mass by its acceleration. Evidence of this lies in the experience of anyone who has felt the temporary decrease in his own weight whenever an elevator in which he was a passenger started its downward trip. In the case of Atwood's machine, the tension in the cord is therefore less than that necessary to support the weight of the large mass, which accordingly possesses a downward acceleration. This tension may therefore be computed, either by subtracting the product of the large mass by the acceleration, which is the same for both sides, from the weight of the large mass, or by adding the product of the small mass by the acceleration to the weight of the small mass, and in the absence of friction the result will be the same in either case. That this is true may be verified by computations from the data of the preceding chapter. In each case the equality of values of $mg + ma$ with corresponding values of $Mg - Ma$ will be noted, hence, the equality of action and reaction.

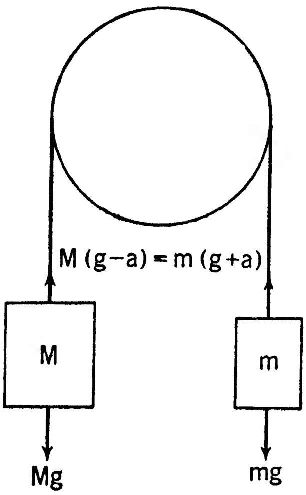


FIG. 107. THE EQUALITY OF ACTION AND REACTION IN ATWOOD'S MACHINE

Universal Applicability of Newton's Laws

In such ways as this the validity of Newton's laws of motion may always be brought to the test of experiment. Perhaps the most striking type of experimental proof, however, is not terrestrial experimentation, such as in the work with Atwood's machine, with all the considerable degree of unavoidable error and the necessary restriction of conclusions based upon them to terrestrial affairs, but rather the successive issues of such publications as the *Nautical Almanac*. This volume of six hundred pages, published four years in advance, contains on every page hundreds of predictions of the positions of the sun, moon, and planets among the "fixed stars"; the exact times, durations, places of commencement, path and conclusion of eclipses — all worked out to the degree of accuracy required by modern navigation. Every one of the data is based on Newton's laws of motion, assumed to be correct for astronomical bodies, though originally verified only on a terrestrial scale. Their accuracy has been vindicated so completely that any deviations from them would suggest only that errors in computation had been made, or, if that explanation were shown untenable, that such phenomena as the disturbance from an undiscovered planet were involved. Of this nature was the origin of the discovery of Neptune by Adams and Leverrier independently in 1845–46, and of Pluto in 1930 by Tombaugh. Not until Einstein's general theory of relativity was propounded in 1916 was an astronomical observation interpreted as giving evidence of any deviation from Newton's laws of motion; and that evidence of deviation was only to an extent of less than one minute of arc per century for the course of the planet Mercury in its orbit. To an extent which is possibly without parallel in the history of science, Newton's laws of motion, originating in observations on terrestrial bodies, have been found applicable throughout the physical universe.

Questions for Self-Examination

1. How can data from Atwood's machine be made to support the assertion that motion under constant force is uniformly accelerated?
2. Similarly, that acceleration is proportional to the applied force?
3. Similarly, that acceleration is inversely proportional to mass?
4. State Newton's three laws of motion and show how the first is really a special case of the second, not a separate law.
5. List some examples of *oppositeness* of forces in situations involving Newton's third law of motion, stating briefly the way in which each example illustrates the law.
6. List some examples of *equality* of action and reaction, stating briefly the way in which each example illustrates the law.

Problems on Chapter 10

English units have been used up to this point. Metric units are usually preferable. To aid in making the transition from English to metric units, every problem in this section is stated in duplicate, once using English units, and again using metric units. The former are the odd-numbered, and the latter the even-numbered problems. Nearly all the problems in subsequent sections are stated in metric units.

1. If the engine of an m-ton automobile truck which is running on a level road acts with a constant force[1] of f pounds above that required to overcome friction, what is the acceleration a in miles per hour per second?[2]

m	f	a
2	500	2.7
3	600	2.2
4	700	1.9
5	800	1.8

2. If the engine of an m metric ton[3] automobile truck which is running on a level road acts with a constant force[1] of f kilograms above that required to overcome friction, what is the acceleration a in kilometers per hour per second?[4]

m	f	a
2	250	4.4
3	300	3.5
4	350	3.1
5	400	2.8

[1] Self-propelled vehicles approximate more nearly to the laws of constant *power* than to those of constant *force*. See problem 12 of Chapter 14.

[2] Take $g = 32.2$. [3] A metric ton is 1000 kilograms. [4] Take $g = 9.80$.

3. If the truck of problem 1 encounters a hill of slope s degrees, what is its acceleration a in miles per hour per second?

	m	f	s	a		m	f	s	a
3.	2	500	1	2.4	4.	2	250	1	3.8
	2	500	3	1.6		2	250	3	2.6
	2	500	5	.83		2	250	5	1.3
	2	500	7	.07		2	250	7	.11

4. If the truck of problem 2 encounters a hill of slope s degrees, what is its acceleration a in kilometers per hour per second?

5. An M-ton train acquires a speed of v miles per hour in t seconds on a level track. The drag due to friction is f tons. Find the average force F in tons exerted by the engine.

M	v	t	f	F
500	25	100	5	11
500	20	64	5	12
500	15	36	5	15
500	10	16	5	19

6. A train of mass M metric tons acquires a speed of v kilometers per hour in t seconds on a level track. The drag due to friction is f metric tons. Find the average force F in metric tons exerted by the engine.

M	v	t	f	F
500	50	100	5	12
500	40	64	5	14
500	30	36	5	17
500	20	16	5	23

7. The coefficient of friction between the rails and wheels of an m-ton engine is k. What is the maximum acceleration a in miles per hour per second which the engine can produce on a level track, when coupled to a train of mass M tons, the frictional drag being f tons?

m	k	M	f	a
100	.2	200	5	1.1
100	.3	200	5	1.8
100	.4	200	5	2.6
100	.5	200	5	3.3

8. The coefficient of friction between rails and wheels of an engine of mass m metric tons is k. What is the maximum acceleration a in kilometers per hour per second which the engine can produce on a level track, when coupled to a train of mass M metric tons, the frictional drag being f metric tons?

m	k	M	f	a
100	.2	200	5	1.8
100	.3	200	5	2.9
100	.4	200	5	4.1
100	.5	200	5	5.3

9. A block of mass M pounds slides without friction along a level tabletop. A cord from this block passes over a pulley and supports another block of mass m pounds. What acceleration a in feet per second per second is produced?

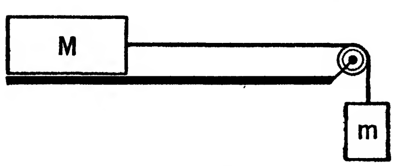


FIG. 108

M	m	a
10	1	2.9
8	3	8.8
5	6	18.
1	10	29.

10. A block of mass M kilograms slides without friction along a level tabletop. A cord from this block passes over a pulley and supports another block of mass m kilograms. What acceleration a in meters per second per second is produced?

M	m	a
10	1	.89
8	3	2.7
5	6	5.3
1	10	8.9

11. Solve problem 9, allowing for the existence of friction, the coefficient of friction being k. Interpret the negative sign of the last result.

M	m	k	a
1	10	.3	28.
5	6	.3	13.
8	3	.3	1.8
10	1	.3	−5.8

12. Solve problem 10, allowing for the existence of friction, the coefficient of friction being k. Interpret the negative sign of the last result.

M	m	k	a
1	10	.3	8.6
5	6	.3	4.
8	3	.3	.53
10	1	.3	−1.8

13. A weight hung on a spring balance produces a reading of M pounds. The weight and balance are suspended from an elevator which starts upward with an acceleration of a feet per second per second. What is the new reading m?

M	a	m
10	4	11
10	6	12
10	8	12
10	10	13

14. A weight hung on a spring balance produces a reading of M kilograms. The weight and balance are suspended from an elevator which starts upward with an acceleration of a meters per second per second. What is the new reading m?

M	a	m
10	1.20	11
10	1.80	12
10	2.40	12
10	3.	13

15. Upon approaching the top floor, the elevator of problem 13 slows down with an acceleration of a feet per second per second. What is the reading m of the spring balance?

M	a	m
10	4	8.8
10	6	8.1
10	8	7.5
10	10	6.9

16. Upon approaching the top floor, the elevator of problem 14 slows down with an acceleration of a meters per second per second. What is the reading m of the spring balance?

M	a	m
10	1.20	8.8
10	1.80	8.2
10	2.40	7.6
10	3.	6.9

17. A force of F pounds acts at an angle of α degrees on a sled of mass W pounds. What is the acceleration a in feet per second per second, and the velocity v in feet per second after the lapse of t seconds? Neglect friction.

F	α	W	t	a	v
5	5°	10	5	16	80
5	10°	10	5	16	79
5	15°	10	5	16	78
5	25°	10	5	15	73

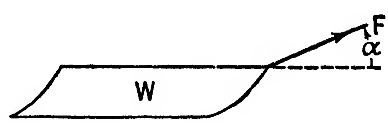


FIG. 109

18. A force of F kilograms acts at an angle of α degrees on a sled of mass W kilograms. What is the acceleration a in meters per second per second, and the velocity v in meters per second after the lapse of t seconds? Neglect friction.

F	α	W	t	a	v
5	5°	10	5	4.9	24
5	10°	10	5	4.8	24
5	15°	10	5	4.7	24
5	25°	10	5	4.4	22

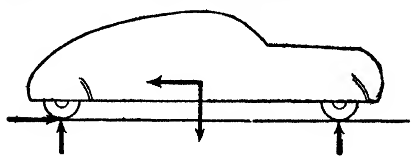
19. Solve problem 17 allowing for friction, the coefficient of friction being k.

F	α	W	t	k	a	v
5	5°	10	5	.3	6.8	34
5	10°	10	5	.3	7.	35
5	15°	10	5	.3	7.1	36
5	25°	10	5	.3	7.	35

20. Solve problem 18 allowing for friction, the coefficient of friction being k.

F	α	W	t	k	a	v
5	5°	10	5	.3	2.1	10.3
5	10°	10	5	.3	2.1	10.7
5	15°	10	5	.3	2.2	10.9
5	25°	10	5	.3	2.1	10.6

21. An automobile has a mass of *m* tons. Its center of gravity is *h* feet above the ground and midway between the front and rear wheels. The wheel base is *l* feet. With how many pounds of force *F* do the front wheels bear on the ground when the automobile is accelerating at the rate of *a* miles per hour per second? Neglect friction and air resistance.



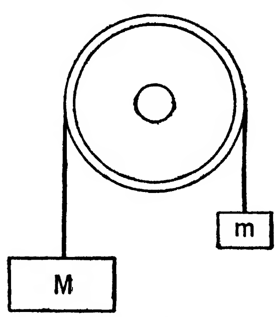
m	*h*	*a*	*l*	*F*
2	2.	5	10	1820
2	2.5	5	10	1770
2	3.	5	10	1730
2	3.5	5	10	1680

FIG. 110

22. An automobile has a mass of *m* metric tons. Its center of gravity is *h* meters above the ground and midway between the front and rear wheels. The wheel base is *l* meters. With how many kilograms of force *F* do the front wheels bear on the ground when the automobile is accelerating at the rate of *a* kilometers per hour per second? Neglect friction and air resistance.

m	*h*	*a*	*l*	*F*
2	.60	10	3	890
2	.75	10	3	860
2	.90	10	3	830
2	1.05	10	3	800

23. Two weights of *m* and *M* pounds respectively are hung by a weightless cord over a pulley (Atwood's machine). Neglecting friction and the inertia of the wheel, find the acceleration *a* of the two weights in feet per second per second, the tension *T* on the cord in pounds, and the total downward force *F* in pounds which the cord exerts on the pulley.



M	*m*	*a*	*T*	*F*
9	1	26	1.8	3.6
8	2	19	3.2	6.4
7	3	13	4.2	8.4
5	5	0	5.	10.

FIG. 111

24. Two weights of *m* and *M* kilograms respectively are hung by a weightless cord over a pulley (Atwood's machine). Neglecting friction and the inertia of the wheel, find the acceleration *a* of the two weights in meters per second per second, the tension *T* on the cord in kilograms, and the total downward force *F* in kilograms which the cord exerts on the pulley.

M	*m*	*a*	*T*	*F*
.900	.100	7.8	.18	.36
.800	.200	5.9	.32	.64
.700	.300	4.	.42	.84
.500	.500	0.	.50	1.

25. Solve problem 23, assuming that the string is retarded at the wheel by a frictional force of *f* pounds. Interpret the negative sign of *a* in the last answer.

M	*m*	*f*	*a*	T_1	T_2	*F*
9	1	.5	24.	2.3	1.7	4.
8	2	.5	18.	3.6	3.1	6.7
7	3	.5	11.	4.5	4.1	8.6
5	5	.5	− 1.6	5.2	4.8	10.

26. Solve problem 24, assuming that the string is retarded at the wheel by a frictional force of f kilograms. Interpret the negative sign of a in the last answer.

M	m	f	a	T_1	T_2	F
.900	.100	.050	7.4	.23	.17	.40
.800	.200	.050	5.4	.36	.31	.67
.700	.300	.050	3.4	.46	.40	.86
.500	.500	.050	− .5	.53	.47	1.

27. The distance required for a car with good brakes on dry pavement to come to a stop from a speed of v miles per hour after a given signal is observed to be s feet. This includes distance traveled during the three-quarters second allowed for the driver's reaction time. Show that the coefficient of friction, k, is just under .5 for each case.

v	s	k
20	50	.48
30	95	.49
40	153	.49
50	227	.49
60	314	.49

28. The distance required for a car with good brakes on dry pavement to come to a stop from a speed of v kilometers per hour after a given signal is observed to be s meters. This includes distance traveled during the three-quarters second allowed for the driver's reaction time. Show that the coefficient of friction, k, is just under .5 for each case.

v	s	k
32.19	15.24	.48
48.28	28.95	.49
64.37	46.64	.49
80.47	69.20	.49
96.56	95.71	.49

29. A man of weight M pounds stands on a suspended platform of weight m pounds. The rope holding the platform passes over a frictionless pulley and thence into the hands of the man. What force F in pounds must he exert to produce an upward acceleration of a feet per second per second in himself and the platform? With how many pounds w does he bear on the floor while producing this acceleration?

	M	m	a	F	w
29.	150	75	1	120	40
	150	75	3	120	41
	150	75	6	130	43
	150	75	10	150	49

	M	m	a	F	w
30.	75	30	30	54	23
	75	30	90	57	25
	75	30	180	62	27
	75	30	300	69	29

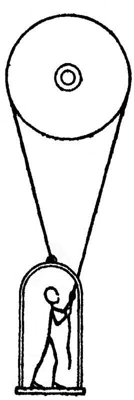


FIG. 112

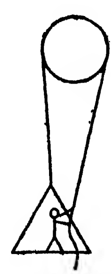


FIG. 113

30. A man of weight M kilograms stands on a suspended platform of weight m kilograms. The rope holding the platform passes over a frictionless pulley and thence into the hands of the man. What force F in kilograms must he exert to produce an upward acceleration of a centimeters per second per second of himself and the platform? With how many kilograms w does he bear on the floor while producing this acceleration?

CHAPTER 11

Universal Gravitation

Newton and His Law

Great as was Newton's accomplishment in formulating the laws of motion, the discovery for which he is best known is his law of universal gravitation. Like his laws of motion, this one also grew out of terrestrial observations (the famous fall-of-the-apple story), and again like them was extended to cover the entire physical universe. Laplace once remarked that Newton was doubly fortunate — first, in that he possessed the ability thus to discover the foundations of the physical universe, and second, in that he could never have a rival because there was only one universe to be discovered.

The law of gravitation is usually phrased somewhat as follows:

> Every particle in the universe attracts every other particle with a force which is directly proportional to the product of the masses of the particles and inversely proportional to the squares of their distances apart.

Newton states his approach to the problem and a partial formulation of his solution as follows (14:7):

> And the same year (1666) I began to think of gravity extending to ye orb of the Moon.... I deduced that the forces which keep the planets in their Orbs must [be] reciprocally as the squares of their distances from the centers about wch they revolve: and thereby compared the force requisite to keep the Moon in her Orb with the force of gravity at the surface of the earth and found them answer pretty nearly.

Newton's Line of Reasoning

Newton says in the above quotation that he "compared the force requisite to keep the Moon in her Orb with the force of gravity at the surface of the earth and found them answer pretty nearly." It will be instructive to follow Newton's train of thought. This can readily be done, for he left a record of it in detail (91:408). The force of gravity at the surface of the earth is such as to cause objects to fall sixteen feet in the first second starting from rest. The corresponding distance of fall would be much less as far away as the moon's orbit. It would, in fact, be inversely as the square of the ratio of the radius of the moon's orbit to the radius of the earth, assuming as a fairly close approximation that the moon's orbit is circular. Since the ratio of these radii is very close to 60, the inverse square would be

$\frac{1}{3600}$. In consequence of the coincidence that there are sixty seconds in a minute, and recalling Galileo's observation that distance of fall is proportional to the square of the time elapsed, this means that at the distance of the moon, objects would require just one minute to fall the same distance as they fall in one second at the surface of the earth. Hence, the moon itself should be pulled away from the tangent to its orbit by a distance of sixteen feet during any single minute that it is observed.[1]

The great question was: is this the way that the moon really moves? With the size of the earth given and also of the orbit of the moon, along with the time required for the moon to travel once around its orbit, the test of this point was a mere matter of elementary geometry. Geodetic measurements had given the circumference of the earth as 126,720,000 English feet [2] (58:68), and, hence, the circumference of the moon's orbit was 60 × 126,720,000 feet or 7,603,200,000 feet. The lunar month was 27 days, 7 hours, 43 minutes or 39,343 minutes. Hence, the orbital speed of the moon was 7,603,200,000 ÷ 39,343 = 193,250 feet per minute. Let MM' (Fig. 114) represent the distance thus traveled by the moon in a minute. Now the actual angle MEM' subtended at the earth by the distance traveled by the moon in one minute is enormously exaggerated in this diagram. On this scale the actual distance traveled by the moon would scarcely be the width of one of the lines. NM' is quite evidently the distance that the moon departs from the tangent to its orbit during the interval.

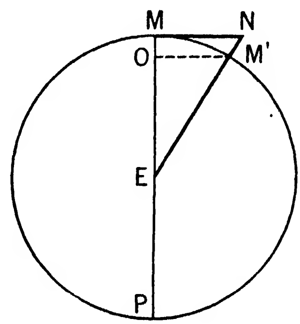


FIG. 114. FREE FALL IN EARTH'S GRAVITATIONAL FIELD AT THE DISTANCE OF THE MOON

If the angle MEM' had not been so exaggerated, this distance, NM' would be almost identical with MO, MM' would be almost identical with OM', and OP would be almost identical with MP. These approximations will be found useful in the course of the following reasoning. OM', being a perpendicular erected on the diameter of a circle, is, by a theorem of plane geometry, the mean proportional between the two portions of the diameter thus separated. That is,

$$\frac{MO}{OM'} = \frac{OM'}{OP}$$

or

$$\overline{OM'}^2 = OP \cdot MO.$$

[1] To say, as is frequently done, that the moon falls *toward the earth* at this rate, though not incorrect, is somewhat confusing to beginners, since at the end of the minute the moon is no nearer the earth than it was at the beginning. It falls in the direction of the earth away from the tangent to its orbit, which itself recedes from the earth 16 feet in that same minute, 48 additional feet in the next minute (or 16 feet referred to a new tangent), and so forth. So it is by very virtue of this "fall" that the moon remains at the accustomed distance from the earth.

[2] There is much doubt as to the value that Newton actually used for this calculation.

Using the approximations mentioned above, this may be changed to

$$\overline{MM'}^2 = MP \cdot NM'$$

or

$$NM' = \overline{MM'}^2 \div MP.$$

But MM', the distance traveled by the moon in its orbit in one minute, has been found to have the value 193,250 feet, and MP is the diameter of the orbit or $7{,}603{,}200{,}000 \div \pi$ feet. Hence, NM', the distance by which the moon departs from the tangent of its orbit in one minute, comes to be 15.43 feet. Comparing this with the value 16.1, which Newton had calculated for this distance if the moon responds to the gravitational pull of the earth, it is no wonder that he could say that he "found them answer pretty nearly."

It would be hard to overstate the significance of this discovery. It was the first clear evidence that had ever been adduced that the familiar behavior of falling bodies and the majestic sweep of the moon in its orbit were all part of the same great scheme. Nor did Newton forbear from suggesting the obvious extensions of his conclusion. Hence, we find him saying (91:410):

> The force which retains the celestial bodies in their orbits has been hitherto called centripetal force; but it being now made plain that it can be no other than a gravitating force, we shall hereafter call it gravity. For the cause of that centripetal force which retains the moon in its orbit will extend itself to all planets.

The Inverse-Square Law

But it was not sufficient to demonstrate that the sun was exerting *an* attractive force upon the planets. There was every reason to suppose that such a force would decrease as the planet withdrew from the sun, and the next question then was how the magnitude of the force was related to the distance. Newton's calculations on the moon in 1666, in which he assumed an inverse-square law for this relation, show that even at that time he must have felt that he had some foundation for such an assumption, though it seems certain that he then considered that it was probably only an approximation (15:344). Newton was not alone in contemplating the inverse-square law. Indeed, twenty years before this, Bullialdus (23:23) had asserted that the force by which the sun "prehendit et harpagat" (takes hold of and grapples) the planets must be as the inverse square of the distance. By the time that Newton returned to the study of gravitation in 1684 the inverse-square law was in the air, though it had merely the status of a quite probable surmise, which still lacked rigorous proof. Newton's contemporaries — Sir Christopher Wren, who later became the famous architect, Edmund Halley, who was in subsequent years the Astronomer Royal, and Robert Hooke, now known principally for his association with Hooke's law, and who assisted Wren in the architectural designs of the London which was rebuilt after the great fire of 1666 — all were convinced that the inverse square would be found to be the true law of gravitational attraction. However, though they discussed it much among themselves and felt that it was implied in Kepler's second and third laws, they were unable to prove to their own satisfaction that the elliptical motion called for by Kepler's second law was a necessary consequence of the inverse-square relation, though they had proved it to hold for circular orbits.

To prove this for the general case of elliptic orbits, instead of merely circular ones, or to solve the inverse problem of showing that bodies subject to an inverse-square law of attraction would describe elliptic orbits, was too much for even these very capable men of science. In August, 1684, Halley went to Cambridge to consult Newton about it, and to Halley's inexpressible gratification Newton replied that he had solved the problem some five years previously and would send the solution to Halley as soon as he could find the papers. This he did the following November, though, not being able to find his earlier papers, he had reworked the problem in the meantime. Stimulated by Halley's interest, as evidenced by this and succeeding visits, and under the momentum of this renewal of his attention to the subject of gravitation, he gathered together the material at hand and began to put it in shape for publication. Halley urged the Royal Society to undertake its publication, but was unsuccessful until he offered to finance the venture himself. The offer was accepted, and the final result was the publication in 1687 of the first edition of Newton's *Principia Mathematica* (91), the classic of mathematical physics.

Acceleration in Uniform Circular Motion

Acceleration was defined in Chapter 10 as *rate of change of velocity.* At that point the only type considered was acceleration along the line of the velocity, whether positive or negative. That is, acceleration was there considered only as it affected the *magnitude* of the velocity. But velocity, being a vector quantity, can also change in *direction* leaving the magnitude unaffected. The rate of such change is another manifestation of acceleration.

When a stone, swung in a sling to get up speed, is released, it moves off tangentially to the circle which it has been describing. In the absence of any force, it moves in a straight line with uniform velocity in accordance with Newton's first law. But before the stone left the sling, it was subject to a force applied through the string and acting toward the center of the circle. In accordance with Newton's second law, the resulting acceleration is also directed toward the center of the circle. The force necessary to hold the stone in its circular path, and hence also the acceleration, will naturally depend upon the speed of the stone and upon the radius of the circle that it describes. Christiaan Huygens was the first to state, in 1673 (59:68 ff.), the relation between acceleration, velocity, and radius for the case of uniform circular motion. It is

$$a = \frac{v^2}{r}. \qquad (1)$$

Suppose that the uniform speed in a circular path is v. This is therefore the velocity at any moment along the tangent to the circle. As described in the preceding paragraph a force toward the center of the circle is required to keep the body from flying off at a tangent. This force is unbalanced except as the resulting acceleration toward the center provides the inertial reaction. The resulting acceleration acts during the time t that the body moves from P to P' (Fig. 115), and, hence, imparts a velocity toward the center, of value at. This, superposed upon the original velocity by composition of vectors, produces the velocity at the end of the given interval. But by consideration of the angles involved,

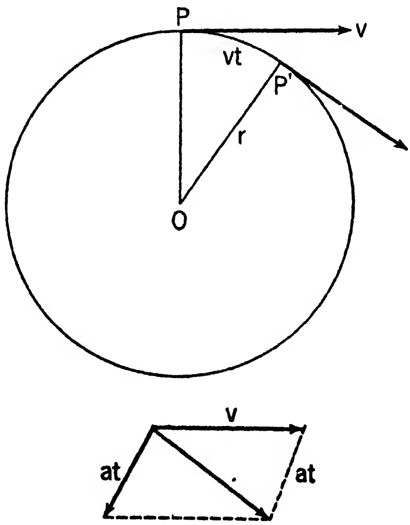


FIG. 115. ACCELERATION IN UNIFORM CIRCULAR MOTION

$$\frac{vt}{r} = \frac{at}{v}. \tag{2}$$

Hence,
$$a = \frac{v^2}{r}. \tag{3}$$

The approximation involved in equation (2) of equality between the ratios arc/radius for a sector of a circle and $\frac{\text{side opposite the angle}}{\text{side adjacent to the angle}}$ for the corresponding triangle may be readily justified. Imagine the angle made progressively smaller until, as it approaches zero, the approximation approaches an equality. This is entirely legitimate, since the circular motion is continuous, and we may, therefore, consider as small an interval of time as we please. Indeed, strictly speaking it would be necessary to consider the limit of the ratio as the time interval approached zero. But since angular velocity ω is v/r by definition, equation (3) may be written:

$$a = \frac{v^2}{r} = \left(\frac{v^2}{r^2}\right) r = \omega^2 r. \tag{4}$$

The force toward the center which deflects a moving object from a straight line into a circular path is termed a *centripetal* force ("seeking the center"). The inertial reaction of the object has the value ma and is of course directed away from the center. It is termed the *centrifugal* force ("fleeing from the center"). It is a measure of the reluctance of the object to depart from a straight-line course just as the corresponding inertial drag of an object that is being speeded up is opposite in direction to the force speeding it up and is a measure of the object's reluctance to depart from its state of uniform velocity.

The Constant of Universal Gravitation

At the beginning of the chapter (page 106) it was stated that gravitational attractions were not only proportional to the inverse squares of the distance but also to the product of the masses of the attracting bodies. The relation may be stated algebraically, as follows:

$$f = G\frac{mM}{r^2}. \tag{5}$$

It will be seen from this that to know the gravitational force between two attracting masses, it is not sufficient to know their masses, m and M, and distance, r, apart, but in addition the constant of proportionality, G, must also be known. This can only be determined by experiment, measuring the gravitational force between two known spherical masses, the distance between whose centers is known. After the value of G has thus been determined, it becomes possible, thenceforth, given two spherical masses and their distance apart, to state immediately the value of their gravitational attraction for each other, whether they be two marbles or two planets. The constant G is accordingly termed the *constant of universal gravitation.*

Moreover, with the value of G known, the mass of the earth can at once be computed and, hence, its mean density, a point of fundamental importance to geologists and astronomers. The computation of the mass of the earth is performed with the aid of equation (5). The force f of the earth's attraction on, say, a one-kilogram mass ($m = 1$) at the surface of the earth is known to be 9.80 newtons. The radius r of the earth is 6,380,000 meters. The value of G, from the most recent determination,[1] is such that two spheres, each weighing one kilogram, with their centers separated by a distance of one meter, would attract each other with a force of 6.670×10^{-11} newtons. It is, therefore, this value which is to be taken for G. Thus the value of every term in equation (5) is known, with the single exception of M, the mass of the earth, which may, therefore, be computed. It comes out to be approximately 6×10^{24} kgms. Hence, experiments having for their object a determination of the constant G of universal gravitation are sometimes given the name "weighing the earth."

A simple computation will show that the mass of the earth, as thus determined, is more than $5\frac{1}{2}$ times as great as it would be if the earth were composed entirely of water; that is to say, the mean density of the earth is over $5\frac{1}{2}$ times the density of water. It is partly on account of this high value of the density of the earth that geologists consider that the earth's interior must be composed of the heavier minerals.

Thus a great deal hinges on the experimental determination of the value of G, the constant of universal gravitation. Newton recognized this, and though he never undertook such an experiment himself, he described (91:569–70) the only two types of experiment which have since yielded

[1] P. R. Heyl, *Bureau of Standards Journal of Research*, *5*, 1243 (1930).

reliable values of G. It will be worth while to observe the method which has proven the most fruitful. It was devised in 1768 by the Reverend John Michell, who constructed apparatus with which to perform the experiment. His death interrupted the work, but it was taken up by Henry Cavendish (1731–1810), who remodeled the apparatus during the years 1797–98 and carried out what has since come to be called the Cavendish experiment. The method is briefly as follows:

In Figure 116 the reader is supposed to be looking down on the apparatus. A light rod, carrying a small metal sphere at each end, was suspended by a delicate fiber. When two heavy lead spheres were brought into the positions shown in the figure, the gravitational forces thereby called into play rotated the rod slightly out of position. The torque necessary to twist the fiber in the opposite direction sufficiently to bring the small masses back to their original positions, gave a measure of the magnitude of this gravitational attraction. The masses and distances being known, all terms in equation (5) were accounted for except G. Its value could, therefore, be calculated.

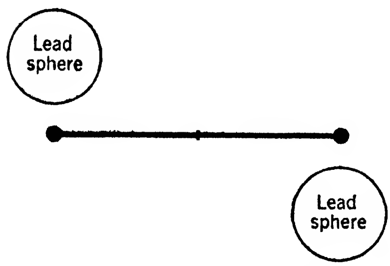


FIG. 116. WEIGHING THE EARTH

The forces involved in this experiment were almost inconceivably small, and the errors accordingly very difficult to eliminate. During the next century and a quarter other observers progressively improved upon the results which Cavendish secured, the improvements culminating in the measurements by P. R. Heyl, already quoted (page 111).

It is perhaps worth noting in conclusion that, in spite of all the effort that has been expended upon it, gravitation remains one of the greatest riddles of science. We have taken its numerical measure, but that is all. Gravitation is one of the outstanding examples of the fact that some of the deepest scientific problems are found in the fields which lie at the center of daily experience. Familiarity with a phenomenon is one thing; understanding seems to be quite another. This is the more tantalizing in the present instance in that there is reason to think that a knowledge of the nature of gravitation would prove a rosetta stone which would enable us to decipher many another of nature's puzzles. As Heyl says:

> It undoubtedly has some intimate connection, not yet understood, with the ultimate structure of the universe.

Outgrowths of Newton's Gravitational Theory

Newton's discovery of the law of universal gravitation opened up a whole new world of scientific adventure. For two hundred years following the publication of the *Principia* the principal pursuits in physics and astronomy consisted of developing in detail the implications of that monumental work. The beautiful way in which many astronomical and terrestrial problems, before insoluble, yielded to the new mode of approach was a constant delight to the entire scientific world. Some of the simpler applications of this law which were made immediately will repay a brief glance.

The mass of any astronomical body having an observable satellite became capable of calculation. This result followed by equating the centripetal force of gravitation between the satellite and its parent to the product of its mass and acceleration; for example,

$$G\frac{mM}{r^2} = \frac{mv^2}{r}. \tag{6}$$

The mass m of the satellite is seen to cancel out. If the speed v of the satellite in its orbit is observed, the only unknown is GM, a measure of the mass of the parent. Even before the value of G became known, GM was sufficient to give the relative masses of planets, but as soon as the important constant G was determined, the masses in terms of customary units such as pounds, kilograms, or tons became determinable. Thus, knowledge of the length of a year and of the distance from the earth to the sun tells the mass of the sun; knowledge of the length of a lunar month and of the distance from the moon to the earth tells the mass of the earth. In the same way the mass of Jupiter becomes known and even the masses of certain stars which have observable satellites.

Through a somewhat tardy realization that comets were subject to gravitational forces, Newton identified these theretofore mysterious apparitions as members of the solar system. His friend Halley, from study of the record of a comet of 1662, predicted its return in 1759. Though he did not live to see it again, it arrived on schedule and has ever since been known as Halley's comet.

The concept of centrifugal force, for which Newton made due acknowledgment to the priority of Huygens, though he had independently developed it himself, played a large part in applications of the theory of gravitation. Applying this concept to the earth regarded as a rotating body he accounted for the equatorial bulge and announced what its magnitude should be, stating also the variations in the acceleration of gravity which should be found at various portions of the earth in consequence thereof. He applied the same principle in reverse form to the planets, deducing the times of their rotation from their observed departure from sphericity.

The earth's equatorial bulge, being acted upon gravitationally by the sun and moon, was demonstrated to produce the precession of the equinoxes, a phenomenon theretofore unexplained.

The tides were shown to result also from the gravitational action of the sun and moon, the oceans being pulled toward the sun and moon on one side and whirled away from them on the other. Newton worked these out in detail, accounting for spring and neap tides, for the effects of coastline and of varying depths of the seas. Never before had the origin of tides been understood.

Such were a few of the many outgrowths, at Newton's own hands, of his law of gravitation taken in conjunction with his laws of motion. It is little wonder that, after the first natural period of incredulity, the scientific world and even the workaday world should have embraced the new doctrine and come to regard its founder almost as a demi-god. Newton is one of the minority of great men whose accomplishments, in spite of detractors, have been appreciated within their own times.

Questions for Self-Examination

1. Show that an object at the distance of the moon would fall from rest toward the earth a distance of 16 feet in one minute, taking the radius of the earth as 4000 miles and the distance to the moon as 240,000.
2. State Newton's law of universal gravitation and outline the course of reasoning which led Newton to formulate it.
3. How was the discovery of the inverse-square aspect of the law of gravitation connected with the publication of Newton's famous *Principia*?
4. How is the value of G, the constant of universal gravitation, experimentally determined?
5. List some of the by-products of the discovery of Newton's law of gravitation.

Problems on Chapter 11

1. Knowing that a one-pound mass experiences 9.802 newtons of force at the surface of the earth, the radius of the earth being 6380 kilometers, calculate the mass of the earth in metric tons. 6.0×10^{21} metric tons.
2. From the fact that the period of the moon around the earth is 27.32 days, and that the distance from earth to moon is 383,000 kilometers, calculate the mass of the earth in metric tons. 6.0×10^{21} metric tons.
3. From the fact that the period of the earth around the sun is 365.25 days and that the distance from sun to earth is 149,500,000 kilometers, calculate the mass of the sun in metric tons. 2.0×10^{27} metric tons.
4. If the sun and the moon (masses 1.990×10^{27} and 7.366×10^{19} metric tons respectively) were directly at the zenith at the same time, how much smaller would be the value of the acceleration of gravity at that point on the earth than it would be 12 hours later? .012 newton.
5. The radius of the earth is 20.92 kilometers greater at the equator than it is at the poles. What is the consequent diminution in the acceleration of gravity at the equator in comparison with that at the poles? Take the average radius as 6380 kilometers. Does the oblateness of the earth's outline make this answer too large or too small? .065 newton.
6. What is the further diminution of the acceleration of gravity at the equator due to the rotation of the earth? .034 newton.
7. What would be the length of a day if the earth rotated so fast that objects at the equator had no weight? Take the radius of the earth as 6380 kilometers. 1.4 hours.
8. Find the tension, T, in kilograms in a cord by which m kilograms are whirled in a circle of r meters radius, making n revolutions per second. Neglect gravity.

m	r	n	T
1	1	1	4
1	1	2	16
1	1	3	36
1	1	4	64

9. A locomotive of mass M metric tons rounds a curve at a speed of v kilometers per hour. The radius of curvature of the track is r meters. Find the centrifugal force F in metric tons.

M	v	r	F
100	50	50	39
100	50	75	26
100	50	100	20
100	50	125	16

10. A cylindrical can d meters in diameter, containing water, is rotated about its own axis at a rate of n revolutions per second. After sufficient time has been given for the water to take up the rate of rotation of the can, what angle α with the horizontal will its surface make at the circumference?

d	n	α
.15	1	17°
.15	2	50°
.15	3	70°
.15	4	78°

11. The curved portion of an automobile race track has a radius of r meters. If the average speed on the track is v kilometers per hour, at what angle α with the horizontal must the track be banked?

r	v	α
100	75	24°
100	100	38°
100	125	51°
100	150	61°

12. For each part of problem 9, how many centimeters h higher must the outer rail be than the inner to cause the resultant force to be perpendicular to the plane of the rail? The rails are d centimeters apart.

d	F	h
120	39.37	44
120	26.25	30
120	19.68	23
120	15.75	19

13. A motorcycle weighing (with its rider) M kilograms rounds a curve of radius R meters at v kilometers per hour. What force F in kilograms acts to upset the machine as a consequence of centrifugal force, and at what angle α with the vertical must the machine be inclined to counteract this effect?

M	R	v	F	α
120	15	25	39	18°
120	20	30	43	20°
120	25	35	46	21°
120	30	40	50	23°

14. A pail tied to the end of a rope l meters long is swung in a vertical circle. What must be its angular velocity ω at the top in revolutions per second in order that the water may not spill out?

l	ω
.64	.62
1.	.5[illegible]
1.44	.42
1.96	.36

15. Find the centrifugal force in tons due to the revolution of the moon around the earth, from the following data:

Mass of moon $= 7.366 \cdot 10^{19}$ metric tons
Earth to moon $= 383{,}000$ km
Period of moon $= 27.32$ days $\qquad$ $2.0 \cdot 10^{16}$ metric tons.

16. What would be the diameter of the smallest steel cable that would hold earth and moon together if the gravitational attraction between them should cease? Take the breaking strength of steel as 35,000 metric tons per square meter. 860 km.

17. Find the centrifugal force in tons due to the revolution of the earth about the sun, from the following data:

Mass of earth $= 5.982 \cdot 10^{21}$ metric tons
Earth to sun $= 149{,}500{,}000$ km
Period of earth $= 365.25$ days $\qquad$ $3.7 \cdot 10^{18}$ metric tons.

18. What would be the diameter of the smallest steel cable that would hold earth and sun together if the gravitational attraction between them should cease? Take the breaking strength of steel as 35,000 metric tons per square meter. 12,000 km.

CHAPTER 12

Harmonic Motion

Harmonic Motion as the Basic Type of Oscillation

Oscillation is fundamental in the study of physics. Not only are examples of oscillation encountered in mechanical systems, such as pendulums and all bodies moving under the action of springs and the like, but the dynamical principles of oscillation are inherently involved in all types of wave motion. This includes such varied phenomena as surface waves in liquids, sound waves, alternating current phenomena in electricity, the production and reception of radio waves, and some of the central concepts in the theory of light. From early application to the phenomena of sound, the term *harmonic motion* came to be commonly used in describing oscillatory phenomena.

While some oscillations are of a simple and easily described nature, others of them, especially those encountered in the study of sound, are extremely complicated. It was not possible to make much progress in the study of the more complicated forms of oscillation until early in the nineteenth century. J. B. J. Fourier (1768–1830), in a work on the conduction of heat, curiously enough, showed how to analyze all such complicated oscillations into component simple oscillations, which could thereupon be dealt with singly. The special type of oscillation thus established as fundamental is commonly termed *simple harmonic motion*. Since the term is cumbersome, the expression *harmonic motion*, which still lacks desirable brevity, will be used in this book wherever no possible ambiguity would result.

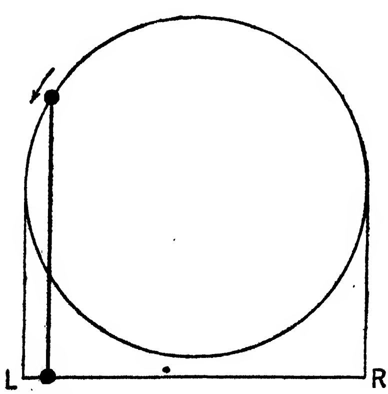


FIG. 117. THE PROJECTION OF CIRCULAR MOTION

There are several ways of defining harmonic motion. The one that will be most useful for the present purpose is as follows. Imagine (Fig. 117) a point traveling on a circle, which will be termed the *circle of reference*, with uniform speed. Suppose that a second point moves along the horizontal line *LR* below the circle in such a way as always to lie directly below the point which is traveling in the circle. The second point will be describing harmonic motion. To be more succinct, ***harmonic motion is the projection of uniform circular motion upon any line in its plane.*** The line along which the harmonic motion is made to occur is frequently taken as the diameter of the circle of reference as in Figure 119 and succeeding figures.

Properties of Harmonic Motion

Some of the properties of harmonic motion are deducible immediately from this definition. For example, in Figure 118, by considering the position of the two points at successive regular intervals, it is evident that the speed of the harmonic motion is most rapid at the center, diminishes with displacement in either direction, and becomes zero at each extreme of motion. The actual relation of this periodically fluctuating speed to the ***displacement*** — as distance of the oscillating particle from the center is called — will be deduced presently. The largest possible value of the displacement, termed the ***amplitude***, obviously occurs when the oscillating point is at either extreme of its travel, and is equal to the radius of the circle of reference. The ***period*** is the time required for one complete oscillation (two successive single oscillations) to and fro across the diameter. The point traveling the circle of reference goes once around it during this time, and, hence, describes an angle of 2π radians. If the period is T seconds, then the angular speed ω, in radians per second, is

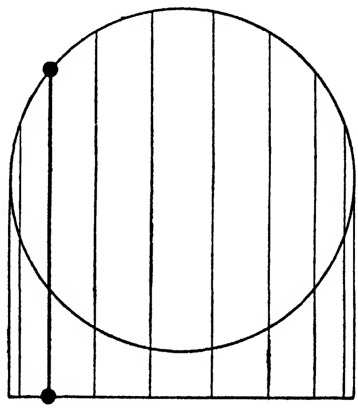


FIG. 118. SPEED IN HARMONIC MOTION

$$\omega = \frac{2\pi}{T}. \qquad (1)$$

The ***phase*** (usually expressed in radians) of an harmonic motion is simply the angle subtended at the center of the circle of reference by the motion of the point along that circle, measured from an arbitrarily established reference line. In Figure 119 the reference line is ON, the motion of the point along the circle which brings the oscillating point to a is NP, and, hence, θ, the angle subtended by this motion, is the phase of the harmonic motion corresponding to the displacement Oa. It will be clear that in this figure zero phase corresponds to maximum displacement to the right, a phase of π radians corresponds to maximum displacement to the left, and phases of $\pi/2$ or $3\pi/2$ radians correspond to zero displacement. If the point describing the circle moves with an angular speed of ω radians per second, then

$$\theta = \omega t, \tag{2}$$

the time t being usually measured in seconds.

Displacement in Harmonic Motion

The most fundamental of all mathematical relations describing harmonic motion is that relating the displacement x to the amplitude A and phase ωt (Fig. 120). By applying the definition of a cosine to the right-angled triangle of the figure, such a relation results, that is,

$$\frac{x}{A} = \cos \omega t, \text{ whence } x = A \cos \omega t, \tag{3}$$

a relation which holds for all values of ωt, and which therefore gives the value of the displacement for any desired instant in the course of the harmonic motion.

By substituting the value of ω from equation (1), this may be stated,

$$x = A \cos 2\pi \frac{t}{T}. \tag{4}$$

It will be well to pause long enough to sense the significance of this equation. Given the amplitude A and the period T of an harmonic motion, this equation makes it possible to state immediately the displacement of a point executing such a motion at any desired instant. Consider, for example, an harmonic motion of amplitude 10 centimeters and period 12 seconds. With the aid of a table of cosines, the following values of the displacement at successive seconds may be verified:

t	0	1	2	3	4	5	6	7	8	9	10	11	12	etc.
x	10	8.66	5	0	−5	−8.66	−10	−8.66	−5	0	5	8.66	10	etc.

In this way the details of any harmonic motion whose principal characteristics (that is, amplitude and period) are given may be tabulated.

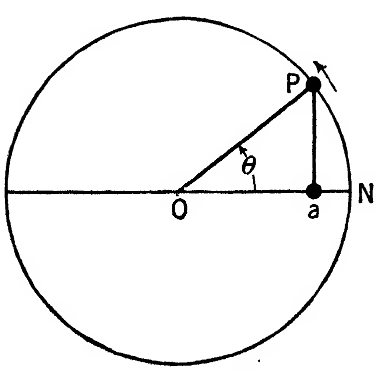


FIG. 119. PHASE ANGLE IN HARMONIC MOTION

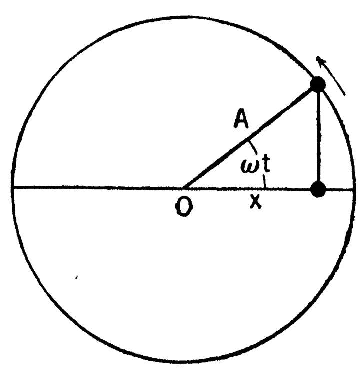


FIG. 120. DISPLACEMENT IN RELATION TO PHASE ANGLE

The nature of harmonic motion is often represented diagrammatically as in Figure 121. If a vertically oscillating body, represented at the left, were at the same time moved steadily toward the right, leaving its trace behind it, the resulting curve would constitute a graphical representation of the data of the preceding table. It is also a graph of equation (4), where A has the value 10 and T the value 12. These three devices — equation, table, and graph — are but three ways of describing harmonic motion. Each form of description has its individual merits. All will be used at various times in the following development of the subject.

Speed in Harmonic Motion

Since the projection of uniform motion in the so-called circle of reference upon a diameter of that circle is harmonic motion, the speed of a point moving harmonically may be found by the projection upon the same diameter of the speed of the point moving in the circle of reference. Representing the speed of this reference point by a tangential vector V, as in Figure 122, the projection of this will be v. It is easy to justify the alleged equality of the three angles marked ωt, and, hence, to show that

$$v = V \sin \omega t.$$

It should be noted, however, that when ωt is less than π, that is, when its sine is positive, the velocity is to the left (the case illustrated in Figure 122) and hence is negative; and that when ωt has values between π and 2π, that is, when its sine is negative, the velocity is to the right and, hence, is positive. In all cases, then, the sign of the velocity is opposite to that of the sine to which it is proportional, allowance for which may be made by introducing a negative sign into the equation, which thereupon becomes:

$$v = - V \sin \omega t.$$

But V, the speed in the circle, is $\dfrac{\text{circumference}}{\text{time of one rev.}}$,

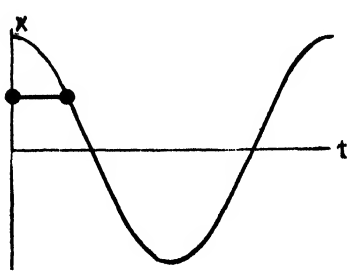


FIG. 121. HARMONIC DISPLACEMENT REPRESENTED IN A WAVE DIAGRAM

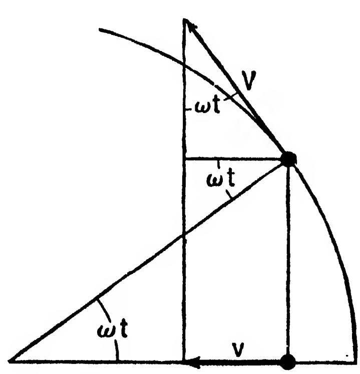


FIG. 122. SPEED IN RELATION TO PHASE ANGLE

or $$V = \frac{2\pi A}{T} \text{ and from equation (1), } \omega = \frac{2\pi}{T}.$$

Hence, finally, $$v = -\frac{2\pi A}{T}\sin\frac{2\pi t}{T}. \tag{5}$$

This, as in the case of equation (4), will repay some attention. It was observed at the outset that the speed in harmonic motion had its maximum value at the center and fell steadily to a zero value at each extreme. Inspection of equation (5) and comparison with equation (4) will verify this surmise, and in addition will yield information on the actual values of the speed at any desired instants. Thus, for the same harmonic motion as before:

t	0	1	2	3	4	5	6	7	8	9	10	11	12	etc.
v	0	−2.6	−4.5	−5.2	−4.5	−2.6	0	2.6	4.5	5.2	4.5	2.6	0	
x	10	8.66	5	0	−5	−8.66	−10	−8.66	−5	0	5	8.66	10	

The values of x are simply copied from the tabular values for equation (4). Comparison of the rows entitled v and x will show that the speed is at its maximum when $x = 0$, and is zero when x is at its maximum absolute value. In the same way, the various speeds at desired instants may be computed for any given harmonic motion.

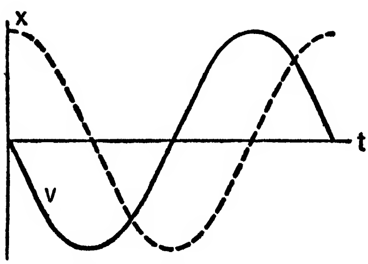


FIG. 123. HARMONIC SPEED REPRESENTED IN A WAVE DIAGRAM

Equation (5) and the data of the above table may be represented also, as before, in graphical form as in the full line of Figure 123, which therefore represents the values of the velocity in a representative harmonic motion. For comparison, the corresponding values of the displacement, copied from Figure 121, are represented by the dotted line.

Acceleration in Harmonic Motion

Equations (4) and (5) formulate the way that displacement and velocity respectively vary with the progress of time in harmonic motion. It still remains to formulate in the same way the variations of acceleration. In some respects this is the most important of the three. In previous chapters the only motions that have been studied have been those of constant acceleration, either having the value zero, as in the cases of statics and of uniform velocity, or having some constant value other than zero, as in the cases of falling bodies and of uniform circular motion. In this chapter a new circumstance appears: acceleration which is not constant, but varies periodically. To discover the law of this variation, an appeal will be made once more to the circle of reference.

The acceleration in the case of uniform circular motion was found by

Huygens and independently by Newton (page 113) to be directed toward the center of the circle and to have the value $\omega^2 A$ where A was the radius of the circle and ω the angular speed of the revolving object, measured in radians per second. So the constant acceleration of the point revolving in the circle of reference may be represented, as it is in Figure 124, by a vector directed toward the center of the circle and having the length $\omega^2 A$. The acceleration of the point executing harmonic motion will then be the component of $\omega^2 A$ parallel to the diameter along which the projections are being made, the value of which is given by multiplying $\omega^2 A$ by the cosine of ωt, or

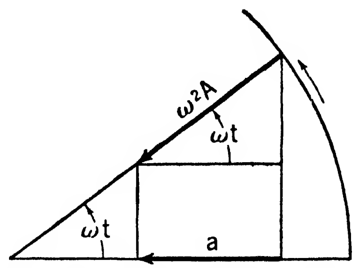


FIG. 124. ACCELERATION IN RELATION TO PHASE ANGLE

$$a = \omega^2 A \cos \omega t.$$

But, as in the case of the velocity, the sign must be attended to. In this instance, when ωt has values between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$, that is, when its cosine is positive, the velocity to the right is decreasing or to the left is increasing, which means that the acceleration is directed to the left (the case illustrated in Figure 124) and, hence, is negative; and when ωt has values between $\frac{\pi}{2}$ and $\frac{3\pi}{2}$, that is, when its cosine is negative, the velocity to the left is decreasing or to the right is increasing, which means that the acceleration is directed to the right and, hence, is positive. In *all* cases, then, the sign of the acceleration is opposite that of the cosine to which it is proportional, allowance for which may be made by introducing a negative sign into the equation, which thereupon becomes:

$$a = -\omega^2 A \cos \omega t,$$

or, substituting for ω from equation (1) as before,

$$a = -\frac{4\pi^2}{T^2} A \cos 2\pi \frac{t}{T}. \qquad (6)$$

A table of values of the acceleration may be constructed as before for an harmonic motion of amplitude 10 centimeters and period 12 seconds, in which the previously calculated values of the displacement and velocity may be included.

t	0	1	2	3	4	5	6	7	8	9	10	11	12
a	−2.78	−2.40	−1.39	0	1.39	2.40	2.78	2.40	1.39	0	−1.39	−2.40	−2.78
x	10	8.66	5	0	−5	−8.66	−10	−8.66	−5	0	5	8.66	10
v	0	−2.6	−4.5	−5.2	−4.5	−2.6	0	2.6	4.5	5.2	4.5	2.6	0

Representing graphically the acceleration described in equation (6) and in the above table, the full line illustrated in Figure 125 results. Corresponding displacements copied from Figure 121 are shown dotted as before.

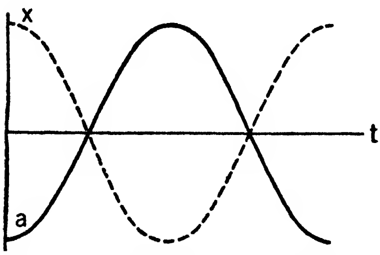


FIG. 125. HARMONIC ACCELERATION REPRESENTED IN A WAVE DIAGRAM

Equation (6), the corresponding table, and the graphical portrayal in Figure 125 — all show that the acceleration fluctuates periodically as both the displacement and the velocity have been seen to do. By comparison of a with the corresponding values of x, the acceleration is seen to pass through its numerical maxima and its zero values at the same time as the displacement, and, moreover, that it is at all times proportional to the displacement, except that it carries the opposite sign. The oppositeness of sign can mean only that when the displacement is to the right, the acceleration is directed toward the left, and vice versa.

Relation of Acceleration to Displacement

This proportionality between acceleration and displacement is perhaps the condition which is most characteristic of harmonic motion. Indeed, it is more commonly used as the defining property of harmonic motion than is the fact that it is also the projection of uniform circular motion on a diameter, the definition used above. Of course, both definitions must come to the same thing in the end. It is possible to prove that there is no other type of motion than the harmonic of which this negative proportionality is true. That is to say, whenever the ratio $\frac{\text{acceleration}}{\text{displacement}}$ is a negative constant, then an harmonic motion will always be found to be in process.

The proportionality between acceleration and displacement may be more neatly deduced from the equations themselves than from any comparison of numerical values computed from them. By comparing equation (4)

$$x = A \cos 2\pi \frac{t}{T} \tag{4}$$

with equation (6)

$$a = -\frac{4\pi^2}{T^2} A \cos 2\pi \frac{t}{T}, \tag{6}$$

the proportionality is at once evident. By substituting in equation (6) for the expression $A \cos 2\pi \frac{t}{T}$ its equivalent, x, from equation (4), there results:

$$a = -\frac{4\pi^2}{T^2} x, \tag{7}$$

an explicit statement of the negative proportionality between acceleration and displacement in harmonic motion. Frequent reference will be made to this equation in future connections. It will be found useful in all cases in which the principles of harmonic motion find application. The equation is frequently stated in a somewhat different form, thus:

$$T = 2\pi\sqrt{\frac{x}{-a}}, \tag{8}$$

which results upon solving equation (7) for T.

Since both x and a are variables, equation (8) would not be of much use were it not for the fact of their negative proportionality to each other. This proportionality, however, results in the ratio of the two being a positive constant, and thus the value of the period of a given harmonic motion is determinable.

Harmonic Motion under Elastic Displacement

One of the common applications of equation (8) is to the harmonic motion that occurs under the action of forces due to elasticity, such as the extension or compression of a spring. In a spring not stretched beyond its elastic limit, Hooke's law applies; that is, the strain is proportional to the stress (page 123). The proportionality between force and displacement involved in Hooke's law implies harmonic motion, for by Newton's second law force is proportional to acceleration, and it has just been noted that the principal characteristic of harmonic motion is the proportionality between acceleration and displacement. Hence, equation (8) may be made to yield information on the period of a body oscillating under an elastic force. Suppose a mass m to be oscillating horizontally under the action of springs as in Figure 126. Let it have been found previously that the "stiffness" of these springs is such that the force necessary to hold the body at one side of its central position was at the rate of k newtons per meter of displacement. k is termed the elastic constant of the system. The Hooke's-law proportionality may then be stated

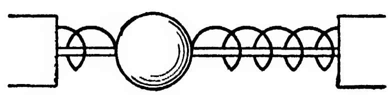


FIG. 126. INERTIA + ELASTICITY = HARMONIC MOTION

$$f = -kx, \tag{9}$$

the negative sign indicating that motion to the right calls into action a restoring force to the left and vice versa. But since Newton's second law of motion also applies, the equation

$$f = ma \tag{10}$$

may be used, as always. From equation (9), $x = -\frac{f}{k}$, and from equation (10), $a = \frac{f}{m}$. When these values are substituted in equation (8), f cancels, whence

$$T = 2\pi\sqrt{\frac{m}{k}}. \tag{11}$$

Thus, the period of any body oscillating under an elastic force is seen to depend solely on the mass m and the elastic constant k.

It is possible to stretch or compress any substance beyond its *elastic limit*, whereupon k has no longer a constant value, and motion under such conditions would therefore no longer be harmonic. But for devices ordinarily classed as elastic, such as steel springs, displacements which are a relatively small fraction of the undistorted lengths of the springs are so nearly proportional to the forces producing them that Hooke's law may be usefully invoked.

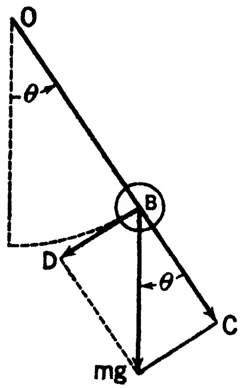


FIG. 127. ANALYSIS OF PENDULUM

Analysis of the Pendulum

Consider a pendulum, consisting of a small heavy bob[1] B of mass m, suspended by a string OB, of length l and of negligible mass (Fig. 127). Let it be drawn to one side until it makes an angle θ with the vertical and then released. It is proposed to show, by an analysis of the conditions which obtain at that instant, that the subsequent oscillatory motion of the bob is a very close approximation to harmonic motion. Let mg represent the weight of the bob, a downward force. This may be replaced by two components. One, BC, is so chosen as to act along the direction of the string. It is balanced by an equal and opposite tension in the string, and the bob is unaffected by this balanced pair of forces.[2] The other component, BD, acts at right angles to the string, and is a force of magnitude $mg \sin \theta$. From Newton's second law,

$$mg \sin \theta = ma, \quad \text{or} \quad a = g \sin \theta.$$

Introduce a negative sign to show that the acceleration is opposite in sign to the displacement, whence

$$a = -g \sin \theta.$$

The displacement is the distance of the bob from its position when the string is vertical, measured along the arc of the circle which constitutes its motion. The value of the displacement is $l\theta$. Hence, the values of both the acceleration and the displacement at the instant of release of the bob are known. They are:

$$\left.\begin{aligned} a &= -g \sin \theta \\ x &= l\theta. \end{aligned}\right\} \qquad (12)$$

[1] Strictly speaking, such a bob, while possessing mass, should be of zero dimensions; that is, it should be a "heavy point." This is, of course, an ideal impossible to realize. By using dense metals, the distribution of mass may be made small enough, so that for many purposes, the error involved in measuring the length of the pendulum to the center of the bob may be neglected.

[2] If the pendulum is swinging, the tension in the string must be larger than BC to provide the requisite centripetal force. But the tangential motion of the bob will also be uninfluenced by these radial forces.

Now the condition for harmonic motion has already been seen to be that the acceleration shall be negatively proportional to the displacement. This condition is evidently not satisfied in equations (12), for an angle is not proportional to its sine, as may be realized from the fact that the sine of $\pi/6$ radians (30°), for example, is .5, whereas the sine of $\pi/2$ radians (90°) is 1. The sines are in the ratio of 2:1, but the angles are in the ratio 3:1. These are very large angles, however. In ordinary cases, as in clocks, pendulums never swing through as great an angle as 90° each way from the vertical, and seldom through as much as 30°. If attention be restricted to small angles, the case will be found quite different, the sines becoming more and more nearly proportional to the angles as smaller and smaller values are taken. The following table shows this:

angle (degs)	30°	20°	10°	5°	2°
angle (rads)	.52360	.34907	.17453	.08727	.03491
sine	.50000	.34202	.17365	.08716	.03490

For 30° the value of the sine differs from that of the angle by about 5 per cent, for 20° the difference is about 2 per cent, for 10° about $\frac{1}{2}$ of 1 per cent, for 5° $\frac{1}{8}$ of 1 per cent, and for 2° less than $\frac{1}{30}$ of 1 per cent.

It accordingly becomes evident that upon restriction to small angles it is possible to say that the angle and the sine are proportional to each other; indeed, that if the angles be measured in radians, the angles are nearly equal to the corresponding sines, an approximation which comes indefinitely closer to actual equality as the angles are taken smaller and smaller. It is customary to say, therefore, that for small angles, the expression $\sin\theta$ may be replaced by θ to a sufficient degree of approximation for certain purposes. The degree of approximation thus introduced obviously depends on the angle involved in a particular case. If this condition is introduced and it is realized that certain limits are thereby placed on the angle of swing of the pendulum, equations (12) will be replaced by

$$\left.\begin{aligned} a &= -g\theta \\ x &= l\theta. \end{aligned}\right\} \qquad (13)$$

Acceleration and displacement are now seen to be mutually proportional and by elimination of θ between the two equations, there results:

$$\frac{x}{-a} = \frac{l}{g}. \qquad (14)$$

This value for $\frac{x}{-a}$ may be substituted into equation (8), giving for the period of a simple pendulum swinging through a small angle the expression

$$T = 2\pi\sqrt{\frac{l}{g}}. \qquad (15)$$

Isochronism of the Pendulum

There are several points of interest in equation (15). One of them is the fact that since the mass m of the pendulum canceled out during the derivation, the period is independent of the mass, a fact which Galileo experimentally observed. Also, again in verification of Galileo's observation, the period T is seen to be proportional to the square root of the length l of the pendulum.

It will also be evident that equation (15) contains no term involving the amplitude or angle of swing. This means that the period is independent of the amplitude as well as of the mass. It would be hard to exaggerate the importance of this property of the pendulum. The constancy of the rate of pendulum clocks depends upon it. It is termed the *isochronism.*

Isochronism of the pendulum was first observed by Galileo. In his *Two New Sciences* (46:97) he causes one of his characters to say:

> Thousands of times I have observed vibrations, especially in churches where lamps, suspended by long cords, had been inadvertently set into motion.... But I never dreamed of learning that one and the same body, when suspended from a string a hundred cubits long and pulled aside through an arc of 90°, or even 1° or ½°, would employ the same time in passing through the least as through the largest of these arcs.

One of the more famous statues of Galileo represents him as engaged in an experiment on the pendulum (Fig. 128). The possibility of utilizing the isochronism of the pendulum in the design of clocks apparently did not occur to Galileo until his old age. In 1641, after blindness had overtaken him, he dictated to his son and to one of his pupils the specifications for a pendulum clock and caused a drawing to be made. The original model constructed from this drawing is still in existence, but the idea did not become generally known at the time, and fifteen years later Christiaan Huygens independently invented a pendulum clock which rapidly met with general appreciation.

It is important to note, however, that the pendulum is isochronous only for small angles of swing. Galileo failed to comprehend this, but his oversight was understandable, since there is only one per cent difference between the period of a pendulum swinging through 1° and that of one swinging through 30°. That the dependence of period upon amplitude does not appear in equation (15) is a consequence of the approximation made in the transition from equation (12) to equation (13). Without that approximation the amplitude (indicated by θ) would not have been canceled out.

Another fact which has had, if possible, even greater significance than the foregoing, appears from a further inspection of equation (15). It is evident that any change in the value of g, the acceleration of gravity, will produce a fluctuation in the period of the pendulum. The origin of the circumstance that gave so much perplexity to Jean Richer in 1671 (page

FIG. 128. GALILEO ENGAGED IN AN EXPERIMENT WITH A PENDULUM

88) is thus identified. An important application of equation (15) is to the experimental determination of g, the acceleration of gravity. If the period T of a pendulum of known length l is observed and if these values are substituted in equation (15), the value of g may immediately be computed. This was first done by Huygens, who communicated to the Royal Society in 1664 his value of g, 9.81 m/sec thereby determined (60:*5*:84; 6:246), though he appears to have performed the actual experiment several years earlier. It is the method used today whenever really precise determinations of g are required, though certain modifications of Huygens' procedure are involved. The merit of the method is that by swinging the pendulum a long time, a very accurate value of the period T may be found. Unfortunately, an equally accurate measurement of the length l is hard, in fact impossible, to get. This is partly due to the difficulty of measuring to the center of the bob, and partly due to the fact, as will be seen later (page 162), that it is not really to the center of the bob that such measurement should be made anyway. The way in which the difficulty is surmounted will shortly be described (page 128).

The Compound Pendulum

In his famous work *Horologium Oscillatorium*, which appeared in 1673, after the time of Galileo but before the time of Newton's *Principia*, Huygens published his description of a pendulum clock, including with it the mathematical foundations for the dynamics of oscillating bodies. His "cycloidal pendulum" (Fig. 129) was a device for realizing isochronism even in pendulums swinging through wide arcs. He also developed the consequences of the fact that no suspension is really without mass, and that no pendulum bob is ever really a point. The effect of the distribution of mass due to both these factors leads to the concept of the *compound pendulum*. Huygens showed how to deal with this case and how, from measurements on a given compound pendulum — for example, a pendulum consisting of a straight rod — the length of the theoretical simple pendulum which would vibrate with the same period could be computed.

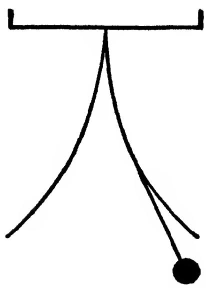


FIG. 129. CYCLOIDAL PENDULUM

If such a rod is suspended from one end, the point on it at a distance equal to the length of a simple pendulum of the same period is termed the *center of oscillation* (point O of Figure 130). Thus, from the standpoint of its behavior as a pendulum, the mass of the rod may be considered concentrated at point O. This point is considerably below the center of gravity of the rod. Thus, in a compound pendulum the center of oscillation is not to be identified with the center of gravity. This is the basis for the statement in the footnote on page 124 that the length of the ideal simple pendulum is not equal to the distance from the point of suspension of a real (and therefore necessarily compound) pendulum to the center of its bob. This point is dealt with at greater length on page 161.

FIG. 130. SHOWING THE DISPLACEMENT BETWEEN THE CENTER OF GRAVITY G AND THE CENTER OF OSCILLATION O OF A ROD SUSPENDED AT THE END S

One of the outgrowths of this was the use of the reversible pendulum in 1817 by an English sea captain, Henry Kater (1777–1835),[1] to facilitate the precise determination of g. He used it in London in a determination which is famous as the first determination which possessed the degree of accuracy characteristic of modern measurements. Kater's pendulum (Fig. 131) possesses two knife edges, the position of one of them being adjustable. When the pendulum is so ad-

[1] Kater, *Philosophical Transactions*, p. 33 (1818).

justed that its period is the same when swinging from one knife edge as when swinging from the other, the length of the equivalent simple pendulum may be shown to be simply the distance between these two knife edges. It will be clear that much greater precision could be secured in making a measurement of the distance between two knife edges than between two points, one of which is the estimated center of a pendulum bob.

FIG. 131. KATER'S REVERSIBLE PENDULUM (Courtesy of Central Scientific Company.)

Vibrations of Stretched Strings

Elastic strings as well as pendulums constitute examples of oscillation. In this field also, Galileo was a pioneer. He had observed (46:100) that the frequency of vibration n of a stretched string was inversely proportional to the length l, directly proportional to the square root of the tension T, and inversely proportional to the "size." Elsewhere he makes it clear that by size he means mass m per unit of length. Stated algebraically his observations were

$$n \propto \frac{1}{l}\sqrt{\frac{T}{m}}.$$

The factor of proportionality may be shown to be $\frac{1}{2}$, whence the relation becomes

$$n = \frac{1}{2l}\sqrt{\frac{T}{m}}. \qquad (16)$$

This resembles equation (15) for the period of a pendulum. The resemblance would be still closer if (15) were stated in reciprocal, thus giving the value of the frequency in both cases. Equation (16) is especially useful in the study of stringed instruments of music. It will be referred to in the section on sound.

Retrospect

The successful treatment of the pendulum by Galileo and later by Huygens, as well as the treatment of vibrating strings by Galileo, were among the major accomplishments of early science. They are especially striking in view of the limited mathematical equipment available at those times. The identification of the center of oscillation was definitely Huygens' greatest work in mechanics, which is saying a great deal. He is generally accorded an equal rank with Galileo and Newton, the three constituting the triumvirate which laid the foundations of physical science. It has been said with much reason that the work of the physical sciences for two hundred years after the time of these men was confined to developing the implications in the principles which they discovered and formulated. Even if these men had done no more than what has been described thus far in this account, the foregoing statement would seem pretty well justified. But, as a matter of fact, all three men made major discoveries in other fields of physics than mechanics, which are yet to be related. Truly, the seventeenth century was the mother of giants.

Questions for Self-Examination

1. Define harmonic motion and tell how s, v, and a vary.
2. Interpret the expressions
$$T = 2\pi\sqrt{\frac{x}{-a}}; \qquad T = 2\pi\sqrt{\frac{m}{k}}.$$
3. What is meant by the term *isochronism* as applied to pendulums? State the limitation of the principle.
4. How may a simple pendulum be used to determine g? What are its shortcomings?
5. Distinguish between simple and compound pendulums.
6. Describe Kater's pendulum as an example of the compound pendulum. How is it used?
7. How does the frequency of a vibrating string depend on the length, tension, and linear density of the string?

Problems on Chapter 12

1. Each prong of a tuning fork which makes 100 double vibrations a second vibrates through a distance of 2 millimeters. Find the speed of the prong at the middle of the swing. 63 cms/sec.
2. How long must a pendulum be to beat seconds (per single swing) at a place where $g = 9.8$? If made 1 millimeter too long how much will its clock lose per day? 99 m, 43 secs.
3. One of the first pendulum clocks to be used in scientific work (Jean Richer, 1671) was found to run slower by $2\frac{1}{2}$ minutes a day when taken from Paris to the tropics. Not until later was it realized that this was caused by change in the value of g. If g was 9.80 at Paris, what was it at the new location? 9.79.
4. What is the frequency of vibration of a wire 50 centimeters long when stretched by a weight of 25 kilograms, if 2 meters of the wire are found to weigh 4.79 grams? 320 vib/sec.
5. A string has its length and tension divided by 4 and its mass per unit length multiplied by 4. What is the effect on its frequency?
6. A wire is L meters long and weighs M grams. What tension F in kilograms must be applied if the wire is to produce a tone whose frequency is n vibrations per second?

	L	M	n	F
6.	1.00	2	200	33
	1.00	4	150	38
	1.00	6	100	24
	1.00	8	75	18

	L	r	l
7.	120	$\frac{16}{15}$	7.5
	120	$\frac{10}{9}$	12.
	120	$\frac{9}{8}$	13.
	120	$\frac{6}{5}$	20.
	120	$\frac{5}{4}$	24.
	120	$\frac{4}{3}$	30.
	120	$\frac{3}{2}$	40.

7. The strings of a 'cello have a length L centimeters. By how many centimeters l must they be shortened, by fingering, to change the pitch by a frequency ratio r?

8. A shelf moves up and down with an harmonic motion of amplitude A meters. What is its smallest period T in seconds for which articles placed upon it will remain steadily in contact with it?

A	T
.01	.20
.1	.63
1.	2.
10.	6.3

9. The piston of an automobile engine weighs m kilograms. Its stroke is l centimeters. It executes harmonic motion approximately. What maximum force f in kilograms is involved in reversing the motion of the piston (at the end of its travel) if the engine is making n revolutions per minute?

m	n	l	f
2	600	10	40
2	1200	10	160
2	2400	10	640
2	3000	10	1000

10. If a mass of m kilograms is loaded onto a spring and thereby extends it s centimeters, with what period T in seconds will it vibrate if set into motion? Neglect the weight of the spring.

m	s	T
.1	2	.28
.2	5	.45
.3	9	.6
.4	14	.75

11. If it were possible to bore a hole through the earth, anything falling through it from one side of the earth to the other would be subject to a force proportional to its distance from the center of the earth. Its motion would therefore be harmonic. How many minutes would an object require to fall from one side of the earth to the other? What would be the speed in miles per hour when passing the center of the earth? Disregard air resistance.
43 min.
18,000 mi/hr.

12. How many miles s has the car of the preceding problem fallen, and what speed v in miles per hour has it attained after having traveled t minutes?

t	s	v
1	12	1,400
3	100	3,700
10	1,100	12,000
20	3,600	18,000
30	7,000	12,000
40	7,900	3,000

CHAPTER 13

Impact

The Principles of Impact

Like many other scientific problems, the problem of impact was in the air during the time of Newton. The newly founded Royal Society of London issued a request for contributions on this subject, and this request evoked in 1668 the first systematic treatment. Three eminent physicists, John Wallis (1616–1703), Christopher Wren (1632–1723), and Christiaan Huygens, complied with the invitation of the society. Wallis was a man of considerable mathematical attainment. Wren had given his principal energies to scientific pursuits up to this time, but was shortly to turn exclusively to architecture, in which field he became one of the best-known men of all time. Huygens is already a familiar figure in these pages and is destined to become still more so. The work of these men was supplemented in 1673 by that of Edme Mariotte (1630–84) and was rounded out to completion by Newton (91:22 ff.), who utilized and acknowledged the contributions of these four men.

Collisions between two moving bodies may be conveniently divided into three classes:

1. Inelastic collisions, such as a collision between two balls of clay, characterized by the fact that there is no rebound, the colliding bodies remaining together after the impact.
2. Perfectly elastic collisions, very nearly represented by a collision between two identical ivory or steel balls, characterized by the fact that the velocity with which the colliding bodies rebound from each other is the same as the velocity with which they approached.
3. Imperfectly elastic collisions, comprising most actual cases, characterized by the fact that the mutual velocity of recession is only a certain fraction of the velocity of approach. It will be evident that this last is the most general of the three cases, for if the fraction mentioned has the value zero, inelastic impact is thereby described. If it has the value unity, perfectly elastic impact is involved. All possible cases must lie somewhere between these two extremes.

Wallis' paper was the first to be submitted to the Royal Society. He dealt exclusively with the case of inelastic impact. His theory was later amplified and published in 1671 (129:1002 ff.). Wallis pointed out that the decisive factor in impact is *momentum*, definable as the product of mass by velocity. If two inelastic bodies which have equal momenta approach from opposite directions and strike each other, rest will ensue after the impact. If their initial momenta are unequal, the momentum remaining after the impact will simply be the arithmetical difference. Similarly, if they are initially traveling in the same direction but with different velocities, the momentum after impact will be the arithmetical sum.

Conservation of Momentum

Momentum is a vector quantity. It possesses not only magnitude but direction, the latter being determined by the direction of the velocity involved. To oppositely directed momenta opposite signs must be attributed. Accordingly the two preceding cases, of unequal motion either in opposite directions or in the same direction, may be subsumed under one law, known as that of the *conservation of momentum*.[1] Limiting it to the case of direct impact (that is, not including "glancing" impact), the law of conservation of momentum may be stated thus:

For any collision, the algebraic sum of the momenta of the colliding bodies is the same after the impact as before. This statement is sufficiently general to cover all the fifteen propositions contained in Wallis' *De Percussione*. Indeed it covers much more, for as will be seen by inspection, it is not limited to the only case which Wallis treated, that of inelastic impact. It has been found to apply to all types of impact, whether inelastic, perfectly elastic, or intermediate.

Inelastic Impact

The principle may be formulated algebraically in various ways and thereby may be made more directly applicable to the respective particular cases mentioned above. Thus, for inelastic impact, let M be the mass of one of two colliding bodies, m that of the other, U and u their respective velocities before impact, and V their common velocity after impact. Then the principle of conservation of momentum may be stated for the inelastic case:

$$MU + mu = (M + m)\,V. \qquad (1)$$

By the use of this equation the velocity after an inelastic impact may be computed, given the masses of the colliding bodies and their velocities before the impact. The usual experimental case is that in which one of the bodies is initially stationary. In this case, since one of the initial velocities, say U, has a zero value, equation (1) becomes

$$mu = (M + m)\,V. \qquad (2)$$

[1] This should not be confused with the law of *conservation of energy*, a completely different doctrine which will receive attention later.

FIG. 132. CHRISTIAAN HUYGENS (1629–1695)
(Courtesy of *Scripta Mathematica.*)

Measurement of Velocity in Impact

The principal difficulty involved in an experimental study of impact lies in the measurement of the velocities. This is the same difficulty that Galileo encountered in his experimental study of accelerated motion (page 18). Galileo solved his problem by so modifying his experiments as to permit the measurement of, not velocities, but distances. Some tactics which are quite similar are usually resorted to in the study of impact. Mariotte was the first to point out (80:*1*:4) that measurement of the speeds of colliding bodies would be facilitated by suspending them as pendulums. The speed of a pendulum at its lowest point is a function solely of the height through which it has descended in describing its arc; and, conversely, the height through which it rises along its arc depends solely upon the speed imparted to it at the lowest point. Thus the speed of either the striking or the struck body becomes measurable in terms of distance. Note that this distance is the *vertical* height through which the pendulum passes in the course of its motion.

Hence, it becomes necessary to formulate the relation between the maximum height traversed by a pendulum and the speed at its lowest point. It will be shown in the following chapter that this relation between speed and height in the case of a pendulum is none other than the familiar relation for freely falling bodies (page 20, equation 3),

$$v = \sqrt{2gh}. \tag{3}$$

A measurement of h and the knowledge of the value of g yield immediately the value of the corresponding speed v. This solves in principle the problem of determination of the velocity of a suspended body at the instant before or after impact at the lowest point of its arc. A certain further amount of indirection is usually injected into the situation in practice through the fact that it is ordinarily more convenient or more accurate to measure h not directly, but indirectly through calculation from the arc which the pendulum describes or the horizontal distance which it moves. The derivation of the relation between h and either of these quantities is a mere matter of geometry, which is adequately covered in laboratory manuals and need not be considered here.

It will be evident that the foregoing method readily lends itself to the experimental determination of the speeds of bullets. A bullet imparting its momentum to the wooden or lead bob of a pendulum in which it embeds itself constitutes a perfect example of inelastic impact. Application of equations (2) and (3) or their equivalent yields the required speed of the bullet, a quantity which is harder to determine by other means. A pendulum so used is termed a ballistic pendulum. This method was first described in 1742 by Benjamin Robins (101:83).

Perfectly Elastic Impact

Turning now from the case of inelastic impact to that of perfectly elastic impact, let us recall first that the principle of conservation of momentum is as applicable to elastic as to inelastic impact. Just as Wallis had confined his attention to the case of inelastic impact, so both Wren and Huygens,[1] in their respective responses to the invitation of the Royal Society, confined attention to the case of perfectly elastic impact. Their findings agreed at all essential points. Wren's paper was somewhat more complete than that of Huygens, which came a month later, in that he accompanied his presentation with experimental demonstrations. It is mainly from Huygens' work that modern theory of perfectly elastic impact rises, however, since he continued study in this field and perfected the theory.

Huygens' complete treatment of elastic impact is to be found in his posthumous treatise *De Motu Corporum ex Percussione*, published in 1703. In effect, Huygens defines perfectly elastic impact as above (page 132), namely, impact in which the relative speed of recession after collision is equal to the relative speed of approach before collision. His point of departure for nearly all his treatise is the assumption that elastic bodies of equal mass, colliding with equal and opposite velocities, suffer merely a reversal of their speeds. He extends this to other cases by straightforward logical processes which, as for elastic impact, can best be replaced by a simple algebraic formulation.

The same notation will be used as for elastic impact. That is, let M be the mass of one of the two colliding bodies, m that of the other, U and u

[1] *Philosophical Transactions of the Royal Society of London*, *1*, 547, 548 (1668).

their respective velocities before impact, and V and v their respective velocities after impact. Then the principle of conservation of momentum may be thus stated for the elastic case:

$$MU + mu = MV + mv, \tag{4}$$

or for the special case of one of the bodies being initially stationary

$$mu = MV + mv. \tag{5}$$

These correspond to equations (1) and (2) for the inelastic case. It will be evident that equation (4) is not sufficient to give the values V and v of the final velocities for elastic impact in the way that equation (1) was sufficient for inelastic impact. There are now two velocities to be computed instead of one; hence, two independent relations are required. The way in which the additional relation may be secured will be presented on page 139.

Imperfectly Elastic Impact

The third case, that of imperfectly elastic impact, is characterized by the fact that the relative velocity of recession bears a ratio to that of approach which is neither zero, as in the inelastic case, nor unity, as in the perfectly elastic case, but possesses some value between zero and unity which must be determined experimentally for each pair of colliding bodies. This was the case which Newton discussed (91:25) and will be seen to be completely general, including both inelastic and perfectly elastic impact as special cases. For this type of impact the ratio of velocity of recession to velocity of approach is a constant characteristic of each pair of colliding bodies. This ratio may be stated algebraically as follows:

$$e = \frac{v - V}{U - u}. \tag{6}$$

The constant relation e between the relative velocities is usually termed the ***coefficient of restitution.*** For inelastic impact, by the very definition of the term, $v = V$ and hence $e = 0$. Similarly, for perfectly elastic impact, $v - V = u - U$, and, hence, $e = 1$. For all other cases e possesses some value between these two extremes. If it were possible to determine the value of e in some other way — than by using values of velocities involved in impact — equation (6) would constitute the required additional relation between v and V referred to above. But this is so seldom true that further search must be made.

Equations (4) and (5) though stated as a formulation of the case of perfectly elastic impact, apply also without any change to this third case. Hence, they are perfectly general, applying to all three cases, for they assume the form of equations (1) and (2) if v is set equal to V.

Impact and Newton's Laws of Motion

Thus far the equations expressing conservation of momentum — (1), (2), (4), and (5) — have been presented merely as statements of experimental fact as, of course, they are. But it is worth while to note that they are implicit in Newton's laws of motion and are derivable from them, though, as a matter of history, they were not so derived in their first inception. They could not be, because they antedated Newton's laws of motion by nearly twenty years. It is possible that the course of development proceeded in the reverse direction and that Newton's laws of motion may owe their form in some measure to his knowledge of the principle of conservation of momentum, though there is no clear evidence on this point. Today, at any rate, Newton's laws are invariably considered the more fundamental of the two.

Consider, then, two masses M and m (Fig. 133), each moving toward the right, but M initially moving the faster.[1] Let U be the velocity of M and u that of m before the collision. Since this is a "rear-end" collision of a larger with a smaller body, m will be speeded up and M slowed down, but the direction will not be changed in either case. Let V and v be the velocity of M and m respectively after the collision. Then the change of velocity of M is $(U - V)$ and that of m is $(v - u)$.

At some instant during the impact suppose the small body to experience a force f and in consequence to be under an acceleration a. The force and the acceleration are related to the mass by Newton's second law of motion

$$f = ma. \tag{7}$$

At the same instant let the larger body experience the reaction F of the smaller body on it, and be accelerated (negatively) at the rate A where

$$F = MA. \tag{8}$$

[1] This choice of directions is made to avoid the minor difficulty that might be involved with sign in case reversal of the direction of motion of one of the bodies were involved.

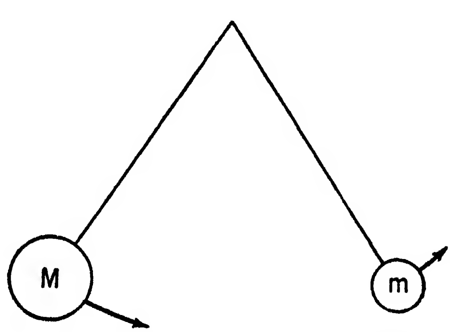


FIG. 133. THE COLLISION OF ELASTIC PENDULUMS

But by Newton's third law of motion F and f, constituting action and reaction, are numerically equal, or

$$F = f, \text{ and, hence, } MA = ma. \tag{9}$$

This relation, assumed for one instant during the impact, applies to all instants. To be sure, the forces will not remain unchanged during the interval covered by the impact, but whatever changes occur in MA will be duplicated in ma, because Newton's third law of motion requires the continued equality of these two. Because of this duplication at every instant during the interval, A and a retain a strict proportionality to each other through all their changes. Hence, the respective changes in velocity $(U - V)$ and $(v - u)$ bear the same proportionality. It follows that

$$M(U - V) = m(v - u). \tag{10}$$

From this by transposition of terms

$$MU + mu = MV + mv.$$

This is equation (4). Equations (1) and (5) have already been shown to be special cases of (4), and (2) is a special case of (1). Hence, the whole doctrine of conservation of momentum, for elastic and inelastic impacts alike, is seen to be derivable from Newton's laws of motion.

Energy Relations in Impact

Over thirty years after his first solution of the problem of elastic impact Huygens made another contribution to its theory, which in some respects considerably overshadowed the previous one. In 1699 [1] he wrote to two scientific magazines as follows:

> The sum of the products of the masses of every hard [that is, elastic] body multiplied by the *square* of its velocity is always the same before and after the encounter.

Hence, in perfectly elastic impact, in addition to conservation of momentum, there appears to be conservation of another entity involving the *squares* of the velocities. The potentialities inherent in this discovery, as subsequent chapters will show, were almost unparalleled in the history of science, though they remained undeveloped for a century and a half. The discovery itself will accordingly repay close inspection.

Let no one say, "Why, of course! If, for example in equation (4),

$$MU + mu = MV + mv,$$

then it naturally follows that also

$$MU^2 + mu^2 = MV^2 + mv^2."$$

[1] *Journal des Savants*, *2*, 534 (1699); *Philosophical Transactions of the Royal Society of London*, *4*, 928 (1699).

Not at all! In fact, the natural conclusion would be exactly the reverse; namely, that if the first equation were true, then the second would almost certainly not be true. To take an analogy from arithmetic, if $5 = 2 + 3$, it not only does not follow that $5^2 = 2^2 + 3^2$, but the statement is clearly not true. The presumption against the possibility of the phenomenon which Huygens described in 1699 was so strong that there is no occasion for surprise that it did not suggest itself to him until thirty years after his first work in this field. The mere fact of the second type of conservation, the type involving the squares of the velocities, would lead one to suspect the existence of some underlying principle, not theretofore discovered. The suspicion would be amply justified, as will shortly be seen. But for the present the new entity which Huygens found to be conserved in elastic impact will be given the name by which it ultimately came to be known, and the more complete development of the principle in which it is involved will be treated later.

The new concept had to wait a long time before receiving the name *kinetic energy*, by which it is known today. It was Thomas Young who, in 1807, proposed the name *energy* and Lord Kelvin who, in 1856, added the objective *kinetic* to differentiate energy due to velocity from that embodied in other phenomena.

In the meantime — 1835 — the French physicist, G. G. Coriolis (1792–1893), had pointed out that this concept, by whatever name it be called, could more conveniently be applied to one half the product of mass by square of velocity. Hence, today kinetic energy is defined by the equation

$$K.E. = \tfrac{1}{2}\, mv^2.$$

Thus for elastic impact, besides equation (4) expressing conservation of momentum, another, expressing conservation of kinetic energy, may be stated, namely,

$$\tfrac{1}{2}\, MU^2 + \tfrac{1}{2}\, mu^2 = \tfrac{1}{2}\, MV^2 + \tfrac{1}{2}\, mv^2;$$

or, multiplying all terms by 2,

$$MU^2 + mu^2 = MV^2 + mv^2, \qquad (11)$$

which is an algebraic statement of Huygens' discovery of 1699.

If, now, equations (4) and (11) are solved simultaneously for V and v, the result is

$$V = \frac{MU + m(2\,u - U)}{M + m} \qquad (12)$$

$$v = \frac{mu + M(2\,U - u)}{M + m}. \qquad (13)$$

Equations (12) and (13) are the two independent relations referred to on page 136.

Equation (11) thus expresses for perfectly elastic impact the conservation of kinetic energy corresponding to the conservation of momentum expressed by equation (4). It should be pointed out that the conservation principle does not apply in just this form either to inelastic or imperfectly elastic impact. For these two cases there are no equations involving squares of velocities that are mates to equation (1) in the sense that equation (11) is a mate to equation (4). This does not mean, however, that the general principle of conservation of energy does not apply to inelastic impact. That principle, in its full generality, is yet to be stated (Chapter 19) and the solution of the apparent anomaly between the cases of elastic and inelastic impact must await that statement.

Boyle's Law as an Outgrowth of Molecular Impacts

In Chapter 7 (page 70) Boyle's law was mentioned as an example of equilibrium conditions in a gas. The fact was pointed out, however, that this was not a case of static equilibrium, like that exemplified in the chapters on forces and on torques, but was a case of dynamic equilibrium, in which the balance existed between the force exerted by the walls of the container and the changing momenta of the flying molecules which rebounded from them. This is, of course, an impact phenomenon on a large scale, with the emphasis on the statistical resultant of myriads of collisions rather than on a single collision.

This is a relatively recent point of view. An earlier view attributed gas pressure to a repulsion between molecules. This was perhaps a natural way to try to account for the expansive behavior of gases, though it had nothing more than a remote plausibility to support it. The theory was entirely *ad hoc*, as there was no other evidence that repulsive forces existed between molecules. The inadequacy of this repulsion theory became still more evident when it was appealed to as a means of accounting for Boyle's law. In several respects the repulsion theory failed to meet this quantitative, mathematical test, a test which has found the weak points in many hypotheses which have been otherwise prepossessing.

Daniel Bernoulli (1700–82) was the first to attribute the steady pressure exerted by a gas to the aggregate effect of molecular impacts — in anything resembling the modern scientific sense (18:Chap. X). He made a brave attempt at establishing a mathematical relation between molecular impacts and gas pressures and, with the benefit of a few doubts, may be said to have been successful.

This preliminary success paved the way for the gradual establishment of what has come to be called the *kinetic theory of gases*.[1] This theory constitutes one of the most important stages in the development of physics. One of the earliest and most successful chapters in it was written by Joule when he actually deduced Boyle's law by a statistical study of the connec-

[1] Waterston, 33:248; Kronig, *Poggendorff's Annalen, 99*, 315 (1856); Joule, *Philosophical Magazine*, Series 4, *14*, 211 (1857); Clausius, *Poggendorff's Annalen, 100*, 315 (1856).

tion between gas pressure and the aggregate effects of innumerable molecular collisions with the walls of the container. The line of reasoning will not be traced here, but the paper is easily available (77:256).

Joule's theory suggested that molecules and atoms were traveling at a surprisingly high speed. He deduced the value 1700 meters per second, for example, for the speed of hydrogen molecules. He also gave the first quantitative development to the relation between molecular speed and temperature and remarked that

> 273° below the freezing point of water must work an absolute zero of temperature corresponding to a zero value of kinetic energy of the particles.

Joule's conclusion that molecules of hydrogen were traveling at speeds of the order of a mile a second must have occasioned much shaking of heads at the time that it was proposed. The identification of temperature with molecular kinetic energies, while not a new idea, received at the hands of Joule its first quantitative development. The same remark applies to the molecular interpretation which Joule gives to the conditions which exist at the so-called absolute zero of temperature. At the root of all of these outgrowths was a basic principle, which has become universally familiar since that time under the name *Conservation of Energy*, but about which little was understood when Joule and others made their contributions. Indeed, Joule is better known for his experimental confirmations of the principle of conservation of energy than he is for the foregoing development of the kinetic theory of gases.

This raises the large and interesting subject of energy and its conservation. We consequently turn from a field in which the subject of energy has been introduced as an incidental matter to a direct study of that basic concept of physics.

Questions for Self-Examination

1. What contribution did each of the following men make to the development of the theory of impact: Galileo, Wallis, Wren, Huygens, Mariotte, and Newton?
2. Distinguish between elastic and inelastic impact.
3. Describe the ballistic pendulum.
4. Define coefficient of restitution. Within what range do its numerical values lie?
5. Demonstrate that for elastic impact the doctrine of conservation of momentum is simply a special formulation of Newton's second law of motion, i.e., from $f = ma$ derive $MU + mu = MV + mv$.
6. What is the kinetic theory of gases?

Problems on Chapter 13

1. The earliest known use of heavy artillery was at the siege of Constantinople in 1453. Brass cannon weighing 19 tons each threw stone spheres weighing 600 pounds. If the projectiles left the cannon at 500 feet per second, what was the speed of recoil? 7.9 ft/sec.

2. Which is the least hazardous, to collide with a wall when going 40 miles an hour or while traveling 20 miles an hour, to collide head on with an approaching car having the same speed? Why?
3. A bullet having a mass of 30 grams is fired into a wooden block of mass 2 kilograms and administers to the block a speed of 6 m/sec. What was the speed of the bullet? 400 m/sec.
4. What was the kinetic energy of the bullet of the preceding problem before the impact? What was the kinetic energy of the block after the impact? What became of the difference? 2400. 36.
5. A monkey clings to one end of a rope passing over a frictionless pulley and is balanced by an equal weight on the other end of the rope. What will happen if the monkey starts climbing up the rope? The masses of the rope and pulley are to be neglected.
6. A bullet of mass m grams strikes a suspended wooden block of mass M kilograms, causing it to swing through such an arc that it rises through a height h centimeters. What was the velocity V of the bullet in meters per second?

m	M	h	V
1.9	1.926	.398	280
4.61	5.706	.351	320
9.68	5.706	.692	220
14.91	6.11	1.972	250

7. A steel marble falls from a height H meters upon a concrete surface and rebounds to a height h meters. What is the coefficient of the restitution e?

H	h	e
1	.9	.95
1	.7	.84
1	.5	.7
1	.3	.55

8. A stream of n bullets per second from a machine gun, each of mass m grams and having a velocity of v meters per second, strikes and is embedded in a target. With what force f in kilograms is the target pushed back?

n	m	v	f
6	10	500	3.1
8	10	600	4.9
10	10	700	7.1
12	10	800	9.8

9. Two steel balls of mass m and M kilograms respectively are launched on level, frictionless ice with velocities u and U meters per second. They collide head on. What is the velocity v and V of each after the impact? The coefficient of restitution is e. Assume all velocities in one direction positive and those in the opposite direction negative.

m	M	u	U	e	v	V
.4	.6	− 2	2	.9	2.6	1.
.4	.6	− 1	2	.9	2.4	.28
.4	.6	0	2	.9	2.3	.48
.4	.6	1	2	.9	2.1	1.2

CHAPTER 14

The Conservation of Mechanical Energy

A Famous Controversy

Late in the seventeenth century a curious dispute arose in scientific circles. Newton had taken "quantity of motion" to be measured by the product of mass by velocity. Descartes had still earlier adopted this idea and written extensively in that vein, though his writings were somewhat confused and lacked the precision of Newton's treatment of mechanics. Leibnitz, who has already been introduced as having been the first to formulate the concept which ultimately developed into that of kinetic energy, took issue with Descartes. He insisted that a more appropriate measure of the quantity of motion would be the product of the mass by the square of the velocity.

It is important to envisage exactly the point at issue in this famous controversy. In modern terminology Descartes' argument was as follows (30:135):[1] If you wish to compare two forces, f and F, allow them to act for any given time, t, upon two masses, m and M respectively; then the ratio of these forces will be

$$\frac{f}{F}=\frac{ma}{MA}=\frac{mat}{MAt}=\frac{mv}{MV}, \tag{1}$$

that is, as the ratio of the resulting momenta.

Leibnitz, on the other hand, based his opinion upon the common experience that it requires the same "force" (as he called it) to raise a body weighing m pounds through a height of $4\,h$ feet as it does to raise a body of $4\,m$ pounds through a height of h feet. Now it is well known, argued Leibnitz, that a body in falling through $4\,h$ feet acquires a velocity just twice as great as when it falls through h feet. Hence, if the "force" required is the same in each of the two cases just mentioned, this "force" must be measured by the product of the "body" (mass) by the square of the speed. In symbols, letting s be the given distance through which the body is moved,

$$\frac{f}{F}=\frac{ma}{MA}=\frac{mas}{MAs}. \tag{2}$$

[1] With slight alterations.

But for uniform acceleration, $v = \sqrt{2\,as}$ (see page 20).
Hence, the ratio of the forces becomes

$$\frac{f}{F} = \frac{\frac{1}{2}\,mv^2}{\frac{1}{2}\,MV^2} = \frac{mv^2}{MV^2}. \tag{3}$$

Thus the issue was squarely joined between the two men. Leibnitz fired the opening broadside in 1686 by the publication of a two-page treatise bearing the long title (67:*3*:180):

> A short Demonstration of a Remarkable Error of Descartes & Others, Concerning the Natural Law by which they think the Creator always preserves the same Quantity of Motion; by which, however, the Science of Mechanics is totally perverted.

The subsequent dispute raged between these men and their partisans for more than half a century. It was finally settled by d'Alembert in 1743 (1:xvii) when he showed that the contest had been a mere battle of words. Thinking that they were talking about the same thing, the disputants had actually been thinking about different things. Both were right, each in a separate field. Descartes' adoption of momentum as the measure of quantity of motion was correct for a force acting for a certain *time*. Leibnitz's adoption of energy as the measure of a quantity of motion was correct for a force acting through a certain *distance*. Each had its use, and once it was clear that momentum and energy were two entirely separate entities, no further occasion for disagreement existed.

The fact is — though nobody up to the time of d'Alembert had realized it — that mechanics had been from the beginning confronted with the necessity, not of choosing between these alternative points of view, but of incorporating both of them into its structure. Until it did so, thoroughly and completely, it was but half a science, limping along on one leg while its normal logical progress required two.

Unknown, even to the pioneers of science themselves, the energy principle had been lurking in the hiding-places of mechanics for a long time. It had been employed by Huygens in his *Horologium Oscillatorium*, though with what degree of appreciation of its generality is hard to say. It had also been used unconsciously by Galileo, and it was even implied in the work of Archimedes on the balance. Great as was the work of these men, we now realize that they were in large measure blind to what is possibly the most fundamental single principle in mechanics.

The Definition of Work

Leibnitz had commented on the equivalence between raising a mass m through a height of $4\,h$ and raising a mass of $4\,m$ through a height of h. The significant point in his mind was thus *the product of the force exerted and the distance traversed in the direction that the force acted.* This product of force by distance is now known as *work,* in the technical or scientific sense; Leibnitz called it "force."

In the literature of physics, work means just one thing, *force multiplied by the distance traversed in the direction of the force.* Algebraically,

$$W = fs. \tag{4}$$

Work, being defined as the product of a force by a distance, is expressible in corresponding units. A common English unit of work is the *foot-pound,* which is the work performed in lifting one pound through a distance of one foot against gravity, or one half-pound through two feet, or any similar product of distance in feet by force in pounds which gives the value unity. This is, however, a somewhat unsatisfactory unit, since the gravitational attraction upon a given mass is measurably different for different localities. The difficulty is avoided by taking advantage of the so-called absolute system of units in common scientific use — involving the metric system, with the meter as the unit of length and the newton as the unit of force (see page 96). This metric unit of work, which would otherwise have to go by the awkward name of meter-newton, has been given the name *joule,* in honor of James Prescott Joule (1818–69) of Manchester, England, whose contributions will be encountered in the study of heat.

The Idea of Potential Energy

If we state that a falling body is converting into kinetic energy the work which was previously done upon it in the process of raising it, we imply that the body thus raised must somehow possess, before it starts to fall, the energy which is later to appear in kinetic form. Once the concept of energy is formed, the possession of energy by a *moving* body is pretty obvious. Kinetic energy thus came early into the scientific scheme. But the possession of additional energy by an object which has been raised from a lower to a higher level is considerably less obvious. It is, in fact, a rather subtle concept. It was a concept which eluded both Leibnitz and Descartes. The concept of potential energy struggled for a place in the scientific sun through the first half of the nineteenth century under various names. Finally an engineer[1] coined the term *potential energy.* It immediately found favor and has been used ever since. Hence, from the welter of scientific confusion about the nature of energy two fundamental concepts emerge: kinetic energy and potential energy. The former is energy possessed by reason of motion; the latter is energy possessed by reason of position or condition, as in a raised weight or a stretched spring. It will be clear that either kind may come into being in consequence of the performance of work; in the first case, work devoted to producing speed; in the second case, work involving some reversible process which subsequently can be made to yield up the energy thus stored.

Some Applications of the Doctrine of Conservation of Energy

If Galileo had possessed the concept of energy, he would not have been satisfied with merely observing that a pendulum bob rose to the same

[1] W. J. M. Rankine (1820–72) (107:203).

height on each side of its swing. He would have pointed out that at the extremes of its travel it possessed no kinetic energy, and that for those positions the potential energy was at its maximum value; also that in the course of reaching its lowest level, the mid-point of its swing, the potential energy had diminished and the kinetic energy had increased to its maximum value, having grown, of course, at the expense of the potential energy previously possessed by the bob; also that as the pendulum bob traveled past the center further and further up the arc of its swing, its kinetic energy once more diminished, being progressively converted again into potential energy. And finally he would have shown that, barring frictional losses, the entire action of the pendulum consisted in these successive alternations between potential and kinetic energy and that at all times the sum of the kinetic and potential energy remained unchanged.

One of the laws of falling bodies developed in Chapter 3 was that the speed attained by falling through a height h was $\sqrt{2\,gh}$ (page 20). Even limited to free fall this was a useful relation but with the aid of the conservation principle it may now be extended. The important point is not the mode of fall but merely the difference in height between the initial and the final positions, regardless even of any simultaneous horizontal travel. Recall that in raising a mass m through a height h the amount of work done upon it is mgh, and that in the process a corresponding increase of potential energy is imparted to the mass. If now the object descends without friction through the same height h, thereby acquiring a velocity v, its kinetic energy becomes $\frac{1}{2}\,mv^2$, acquired at the expense of the potential energy previously possessed. Equating the potential and kinetic energies,

$$mgh = \tfrac{1}{2}\,mv^2 \qquad (5)$$

whence

$$v = \sqrt{2\,gh}. \qquad (6)$$

Though this relation has been encountered before, it appears now in a new light. It is now entirely free of any limitations such as free fall. No particular path whatever has been assumed; hence, it now appears that, *whatever the path*, any body which moves without friction under gravity with zero initial vertical velocity from one level to another, lower by a distance h, will acquire the speed $\sqrt{2\,gh}$. It need not descend vertically; it may travel an indefinite distance horizontally; it may descend far below the final level and rise to it again; it may do anything, provided only that it completes its course, without friction or other losses, at the required distance below its starting point.

The possibility which this opens up of treating mechanical problems by a general consideration of energy relations, without being concerned with the purely incidental minutiae, is sometimes very useful. A good example is the study of water jets.

Speed of Efflux

The energy principle is naturally not limited to solids. It applies to fluids as well. If it is desired to know the speed of efflux of a liquid from an opening a distance h below the level in a container, the problem might be rather complicated if it were necessary to consider all the forces and accelerations involved within the container. The energy principle makes all that unnecessary and tells immediately that since the kinetic energy of an emerging drop must equal the potential energy which it possessed before it started to move, its speed as it issues must be $\sqrt{2gh}$. This, as has been observed, is the speed acquired by any body which falls freely through a height h.

The first man to associate the speed of efflux of a liquid with the speeds of falling bodies was Evangelista Torricelli, the same man who constructed and experimented with the first barometer (page 83). In 1641, Torricelli investigated the motions of projectiles and the flow of liquids (124:*88*:185). In it he said:

> Liquids which issue with violence (from an opening in a vessel) have at the point of issue the same velocity which any heavy body would have, or any drop of the same liquid, if it were to fall from the upper surface of the liquid to the orifice from which it issues....

There is, of course, no hint here of the energy concept. But it has just been pointed out that the relation of speed to distance which Torricelli invokes here is most simply derived in its most general form by using the energy principle.

There is another point of interest about the efflux problem. It illustrates a weakness as well as the strength of the energy method. It yields the *value* of the speed but gives no information about *direction*. By using different shapes of spout, the direction can, in fact, be controlled quite without limitation as indicated in Figure 134. The energy principle applies to all cases and is entirely noncommittal on a question which would be answered immediately if forces and accelerations were considered instead of merely energies.

The speed of efflux is, indeed, not limited as to direction. In the absence of a spout to give guidance to the liquid, the various parts of the emerging

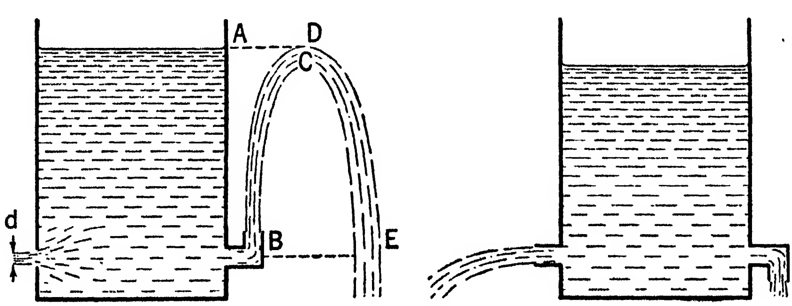


FIG. 134. THE ENERGY PRINCIPLE APPLIED TO FLUIDS

jet are not even parallel but converge, producing a "waist" or constriction in the jet, known as the *vena contracta,* a short distance from the point of emergence (point *d* of Figure 134). This information, also, comes from other sources than the energy equation. Thus the energy principle, useful as it will be found, is far from all-inclusive in its generality. While all physical phenomena proceed in accordance with it, application of the principle to these phenomena will yield information only about certain aspects of them, which may not be the aspects which are of greatest interest.

Stream-Line Flow

Another class of phenomena to which the energy principle may usefully be applied is the stream-line flow of fluids. The term *stream-line* is used so widely and so indiscriminately that it will be well to clarify it. When the velocity at each point of a given body of moving fluid remains unchanged, both in magnitude and in direction, the flow is clearly occurring along certain unchanging lines.[1] These lines of steady flow are called streamlines. Any narrow, cylindrical or conical portion of such a fluid is called a tube of flow. Obviously, by definition of tube of flow, the fluid constituting such a tube remains in it during the entire course of the flow. Figure 135 represents such a tube. It might be a portion of a larger body of fluid, or it might consist of as common a thing as the tapered interior of the nozzle of a hose. In either case, all the fluid entering at the left end leaves by way of the right end, none of it crossing the boundaries in the interim.

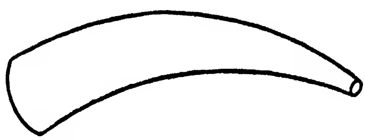


FIG. 135. A TUBE OF FLOW IN STREAM-LINE MOTION

The Bernoulli Effect

If the tube is tapered, as in the illustration, the speed of the fluid will necessarily be greater at the smaller end than at the larger. Experiment has shown, contrary to what might be supposed, that the pressure is less at the small portion of a constricted tube than at the larger portion. This is indicated in Figure 136 by small manometers placed at the points in question on a constricted tube. This association of regions of diminished pressure with increased fluid velocity and vice versa is termed *Bernoulli's Principle.* It was first described by the same Daniel Bernoulli who first

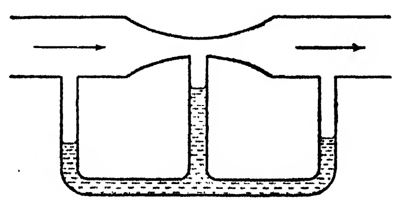


FIG. 136. PRESSURES IN A CONSTRICTED TUBE

[1] When the relative speeds of adjacent fluid surfaces exceed certain values, or when the boundaries are rough or otherwise discontinuous, no steady lines of flow can exist. The irregular variations thus introduced constitute *turbulent* flow, which is to be set in contrast with steady flow, to which the present discussion is confined.

connected Boyle's law with molecular impacts. It may be accounted for by an application of the principle of conservation of energy to stream-line flow, as follows.

An increase in speed can occur only if work is done on the fluid as it moves from one end of the tube to the other, this work being converted into kinetic energy. (Losses by friction, etc., will be disregarded. The flow will be assumed horizontal so that disturbing effects due to change in height can also be disregarded.) Work done *on* a volume V of fluid as it enters the tube will be given by

$$\begin{aligned} \text{force} \times \text{distance} &= \text{pressure} \times \text{area} \times \text{length} \\ &= \text{pressure} \times \text{volume} = P_1 V \end{aligned}$$

where P_1 is the pressure at the entering end of a tapered tube. The work done *by* the same volume of fluid as it emerges will similarly be given by the expression $P_2 V$, where P_2 is the pressure at the emerging end. The difference is the net work done on the given volume, or $(P_1 - P_2)V$.

The work thus done on the moving volume of fluid will change the speed from v_1 to v_2, and the corresponding kinetic energy from $\frac{1}{2}\,mv_1^2$ to $\frac{1}{2}\,mv_2^2$. If the density of the fluid is d, then m may be given the value dV. Whence

$$(P_1 - P_2)V = \tfrac{1}{2}\,dV(v_2^2 - v_1^2), \qquad (7)$$

or

$$P_1 + \tfrac{1}{2}\,dv_1^2 = P_2 + \tfrac{1}{2}\,dv_2^2. \qquad (8)$$

This equation shows that regions of high pressure are characterized by low velocity and vice versa, as observation has shown. If the difference in pressures is known, the difference in the squares of the velocities can be deduced, or vice versa. This relation is utilized in the so-called *Venturi meter*, a device for indicating through pressure differences in a constriction, the rate of flow of a fluid through the interposed device. Bernoulli's principle is invoked to account for many other phenomena, such as the action of a gas jet in sucking in air, of an atomizer, and of the carburetor-jet of a motor car in sucking liquid fuel into a blast of air; the curving of baseballs and golf balls; and so forth. Strictly speaking, Bernoulli's principle applies only to incompressible fluids, but with certain reservations it may be used qualitatively to explain these other effects.

The Energy Principle Applied to Machines

Another area in which the idea of conservation of mechanical energy may be usefully applied is the design of machines. Newton had a somewhat vague prevision of this in the Scholium to his third law of motion (91:28). A modern version of it would be a statement of the equality of work done to energy absorbed in a given process. This can be applied to the lever, which has already been treated by the force method. By that method it was shown that if, for example (Fig. 138), the long arm of a lever is twice as long as the short arm, a force on the short arm will be balanced by one of half the magnitude on the long arm. Now when the lever is in action, motion is produced at the point of application of each

FIG. 137. VENTURI METER TUBE
(Courtesy of Builders — Providence.)

force, and by the simplest geometry it will be seen that the point of application of the small force will move twice as far as that of the large force. Hence, if we take products of forces and distances moved, the work done at one end of the lever is found equal to that done at the other.

The so-called compound pulley is sometimes treated as a modified lever, with the fulcrum at one end, the weight in the middle, and the force applied at the other end. This is entirely justifiable, as an examination of the shaded portion of the pulley of Figure 139 will show, but somehow the analogy appears a bit strained when it is realized that the shaded portion represents successively different portions of the pulley. But if Newton's work principle is applied the mechanical minutiae can be avoided entirely. From the mere observation that the weight W is supported by two ropes symmetrically placed, the force exerted through each is seen to be necessarily $W/2$. The application of the work principle will then require the conclusion that if the weight is to be raised a certain distance, the free end of the rope must be raised twice that distance. Similarly for two pulleys (four ropes over the lower pulleys) (Fig. 140), the weight is distributed equally among four ropes, the force exerted will be in the ratio one to four, and the distance consequently four to one. The ratios 1 : 6 and 6 : 1 will apply to three pulleys, and so forth. As a matter of convenience to the user, in all three of these cases, the free end of the rope would in practice be carried over an upper fixed pulley. A downward pull on a suspended rope enables the user to use his muscles to better advantage, of course.

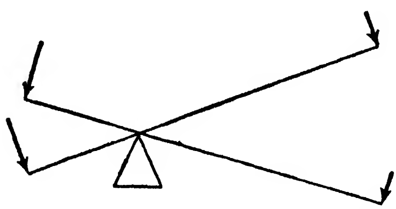


FIG. 138. THE WORK PRINCIPLE APPLIED TO THE LEVER

The term *mechanical advantage*, commonly applied to such devices, indicates the ratio between weight and force. For example, in the lever of Figure 138 and in the pulley of Figure 139 the mechanical advantage is *2*. In Figure 140 it has the value *4*, and so forth. The use of the extra fixed pulley referred to in the preceding paragraph does not affect the mechani-

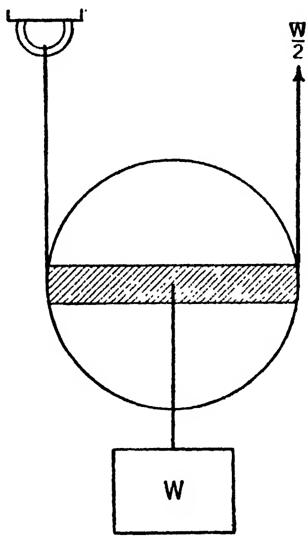


FIG. 139. THE PULLEY AS A LEVER

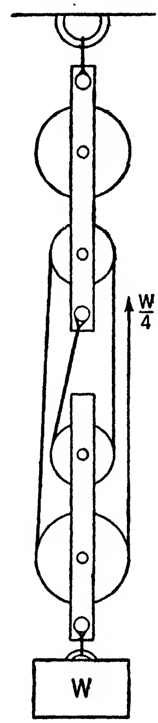


FIG. 140. COMPOUND PULLEY

cal advantage, which for a pulley system is affected only by the number of ropes attached to the moving block.

Efficiency

It is usually not permissible, in actual practice, to disregard the energy wasted by a machine, as has been done in the preceding examples. The energy expended upon a machine always exceeds that yielded by it, the excess being the energy dissipated, principally through friction. In many mechanical operations the ratio of energy yielded to energy absorbed is sufficiently constant to be taken as descriptive of the performance. This ratio is termed *efficiency*, and the efficiency of an operation is thus describable in numerical terms by some number less than unity. Examples are not hard to find. An eight-hundred-pound elevator raises a two-hundred-pound man. The useful work done, the raising of the man, is proportional to his weight, but the total work done is proportional to the total weight. The efficiency of the operation is then the ratio of two hundred to one thousand or 0.2. It is true that at the conclusion of the upward trip the elevator possesses potential energy, which might in theory be utilized and thus raise the efficiency of its operation. In practice the remaining

0.8 is nearly always a total loss, since the subsequent descent without the man is usually kept under control by friction, thus dissipating the potential energy in non-useful heat.

Power

There is a final application of the concept of work to be made before the next subject is taken up. While the same amount of work would be required, say, to ascend a flight of stairs in five minutes as in five seconds, the rate of doing work would be quite different in the two cases. The rate of doing work is of sufficient importance to have received a name and a unit.

Rate of doing work is termed *power.* In this case as in the case of other technical terms careful attention must be paid to the exactness of this definition, and how it contrasts with the wide range of vague meanings attached to it in non-technical use. *Power is work per unit time.* In the English system the foot-pound has been mentioned (page 145) as a unit of work. Thomas Savery, whose pumping engine, patented in 1698, was the first device actually to use steam power in industry, suggested as a standard of power the rate at which a horse could do work. The figure 550 foot-pounds per second was somewhat casually set by James Watt in 1782 for this rate. Watt was making major improvements in the design of steam engines at the time (cf. Chapter 20) and either by guess or by rather crude observation set at 180 pounds the force that an average draft horse would be able to exert while walking at the rate of three feet per second. The resulting value for the power of a horse, 540 foot-pounds per second, was apparently rounded out later to 550 [1] and the unit thus determined was named the *horsepower.* This is an awkward unit in more ways than one, but it was a natural one to introduce at a time when every prospective customer of the engine-builder was asking the question, "If I buy one of your engines, how many horses will it replace?"

In the metric system the unit of power is, naturally, one joule per second, and it is named the *watt,* in honor of James Watt, whose inventions improved the steam engine nearly to the plane of its present performance. The watt, as a unit of power, is used so commonly in connection with electrical devices that the association produces the impression that it is an electrical unit. Such is, however, not the case. It is fundamentally a unit of mechanical power, similar to the horsepower, though much smaller. It requires 746 watts to be the equivalent of one horsepower. The metric unit of power which is analogous to the horsepower is the kilowatt (1000 watts). The horsepower is thus 0.746 of a kilowatt.

Questions for Self-Examination

1. Carry through the two courses of reasoning that led to the collision between Leibnitz and Descartes on the proper measure of "quantity of motion."

[1] *The American Physics Teacher,* 4, 120 (1936).

2. Evaluate Galileo's early contribution leading up to the concept of energy as formulated later by others.
3. Describe Galileo's treatment of the pendulum.
4. Show that the velocity of a pendulum at its lowest point, the descent having been through a height of h, is $\sqrt{2gh}$ and comment on the significance of the method of derivation.
5. Distinguish between kinetic energy and potential energy. Mention instances of conversion from one to the other.
6. Discuss the velocity of efflux of a liquid jet.
7. State Bernoulli's principle and connect it with the conservation principle.

Problems on Chapter 14

1. A gun shoots a projectile weighing a ton with a speed of 2250 ft/sec. Calculate the energy in foot-pounds. 1.6×10^8.
2. A belt from a pulley 4 feet in diameter making 300 revolutions per minute transmits 114 horsepower. How strong must it be?
   It must withstand at least 1000 lbs.
3. A 150-pound man runs upstairs, rising 15 feet in 3 seconds. What horsepower does he develop? 1.4.
4. An automobile requires 50 horsepower to drive it 60 miles per hour. What is the force being exerted on it by the engine? 310 lbs.
5. A 30-gram bullet initially traveling 400 m/sec penetrates 10 centimeters into a wooden block. What average force does it exert? 2400 kg.
6. The starting point of a "loop-the-loop" must be at least how much higher than the highest point of the loop? The center of gravity of the car is above the track at the start and below it while "topping" the loop. What bearing does this have on the answer to the foregoing question? $\frac{1}{2}$ of the radius.
7. A Prony brake of length l feet registers a force of f pounds exerted by an engine of speed n revolutions per minute. What is the "brake horsepower" P?

	l	f	n	P
7.	4	100	100	7.6
	4	150	80	9.1
	4	200	60	9.1
	4	250	40	7.6

	m	l	α	W
8.	100	10	30°	500
	100	10	50°	770
	100	10	70°	940
	100	10	90°	1000

8. How much work W in foot-pounds is done when an m-pound box is pulled without friction up an incline l feet long which makes an angle α with the horizontal?
9. In the preceding problem, what horsepower P is necessary to accomplish the task in t seconds?

	W	t	P
9.	500.	1	.91
	766.	2	.7
	939.7	3	.57
	1000.	4	.45

	m	l	α	K
10.	1	1	20°	3.4
	1	1	30°	4.9
	1	1	50°	7.5
	1	1	60°	8.5

10. A mass of m kilograms slides down a frictionless inclined plane of length l meters, which make an angle α with the horizontal. Find the kinetic energy K at the foot of the incline in joules.

11. Solve the preceding problem assuming a coefficient of friction k.

	m	l	α	k	K
11.	1	1	20°	.3	.59
	1	1	30°	.3	2.4
	1	1	50°	.3	5.6
	1	1	60°	.3	7.

	m	P	t	v
12.	4000	50	1	14
	4000	50	5	32
	4000	50	10	45
	4000	50	20	64

	v	a
13.	5	21.
	20	5.1
	40	2.6
	60	1.7

12. An m-pound automobile possesses P horsepower in excess of that required to overcome friction and air resistance. What speed v in miles per hour can it acquire in t seconds on a level road? (Note that this involves motion under constant *power*, which is of quite a different type than motion under constant *force* [see problems 1 and 2 of Chapter 10].) In this case the conservation principle is useful. The surplus work, in absolute units, done by the engine is equal to the final kinetic energy of the car.

13. What is the acceleration a, in miles per hour per second, of the automobile of the preceding problem when the speed has reached v miles per hour?

14. A steamship is driven v miles per hour by engines of P horsepower working at an efficiency e per cent. What is the thrust F of the screw in tons?

	v	P	e	F
14.	5	1000	30	11.
	10	4000	30	22.
	15	9000	30	34.
	20	16000	30	45.

	p	v	n	V	P
15.	100	10	1	20	400
	400	20	2	30	1300
	1350	30	3	40	4300
	4267	40	4	50	13000

15. It requires p horsepower to drive a ship through the water at a speed of v miles per hour. If the resistance of the water is proportional to the nth power of the speed, what horsepower P is required to drive the ship at a speed of V miles per hour?

16. An inverted simple pendulum of weight m kilograms and length l meters, having a rigid suspension, is released. With what speed v in meters per second does it pass its lowest point and what is the tension T in kilograms in the suspension at the instant?

	m	l	v	T
16.	4	.64	5.0	20
	4	1.	6.3	20
	4	1.44	7.5	20
	4	1.96	8.8	20

	m	l	v	T
17.	4	.64	5.6	24
	4	1.00	7.0	24
	4	1.44	8.4	24
	4	1.96	9.8	24

17. The pendulum of problem 16 is given an initial velocity such that as it passes the top it exerts no force on the support (cf. problem 14 of Chapter 11). With what velocity v in meters per second does it pass its lowest point and what is the tension T in kilograms in the suspension at the instant?

CHAPTER 15

Rotation

Human Intuitions on Rotation Are Unreliable

Nearly all the discussion of mechanics, in the foregoing chapters, has been confined to translatory motion. The one exception was in Chapter 7, which presented some aspects of rotational statics. This bore somewhat the same relation to the dynamics of rotation that the subject of resolution and composition of vectors as presented in Chapters 4 and 5 did to Newton's laws of motion.

In developing the statics and dynamics of translatory motion, it was possible to make rather free use of the accumulation of intuitions and ideas which develop in the ordinary daily experience of the average individual. While they needed to be clarified at many points, the development of many of the fundamental concepts of translatory mechanics consisted primarily in extending and generalizing these common intuitions.

It is quite otherwise with rotation. The muscular reflexes which humanity has perfected are associated almost exclusively with translatory motions and are of very little help when an occasion arises for the acquirement of insight into rotatory phenomena. The behavior of a spinning top is perhaps the most familiar example of rotation, but the deficiency of popular comprehension of its principles is indicated by the mystification with which the action of a gyroscopic ship stabilizer or the stability of a gyroscopic monorail car is regarded — both of which embody in somewhat simplified form the principles of the top. Because of this lack of intuitive material with which to work, it will be necessary to make a somewhat different approach to the subject of rotation from that which has characterized the treatment of translation. Instead of building up the logical structure of rotatory dynamics from the scanty materials provided by daily experience, the general plan will be to make extended use of analogy with the dynamics of translation as developed in the preceding chapters. This will be greatly facilitated by the fact that there is virtually a complete parallelism between these two great sections of mechanics. For every phenomenon of translatory motion, there is a corresponding phenomenon of rotatory motion. By minor modifications, Newton's laws of motion, the phenomena of harmonic motion, the laws of impact, and the great generalizations on work and energy — all treated earlier exclusively from the standpoint of translation — may be extended to include rotation.

The Analogy of Rotation with Translation

In setting up an analogy between translation and rotation, it will be evident immediately that the rotational entity which corresponds to linear motion is angular motion. In place of the measurement of linear displacements in feet or meters, the measurement of rotatory displacements in revolutions, in degrees, or in radians, preferably the latter, is substituted. Similarly in place of linear speeds and linear accelerations, angular speeds and angular accelerations are substituted, defined in precisely analogous terms. This makes it possible to formulate immediately for rotation the kinematical relations deduced for translation in Chapter 2 and Chapter 3.

There will naturally be some changes in notation. It is the custom in mathematics, physics, and engineering to use Greek letters to denote angular quantities. The common notation for angular displacement (corresponding to s for linear distance) is θ (theta); for angular speed (corresponding to v for linear speed) is ω (omega); for angular acceleration (corresponding to a for linear acceleration) is α (alpha). Using this notation, the following equations may be stated at once for rotation by analogy with the corresponding equations for translation.

Chap.	From Page	Eqn.	For Translation	For Rotation	
2	14	1	$v = \frac{s}{t},\ s = vt,\ t = \frac{s}{v}$	$\omega = \frac{\theta}{t},\ \theta = \omega t,\ t = \frac{\theta}{\omega}$	(1)
10	94	1	$v = at$	$\omega = \alpha t$	(2)
	94	2	$s = \frac{1}{2} at^2$	$\theta = \frac{1}{2} \alpha t^2$	(3)
	94	3	$v = \sqrt{2\, as}$	$\omega = \sqrt{2\, \alpha\theta}$	(4)
	94	4	$v = v_0 + at$	$\omega = \omega_0 + \alpha t$	(5)
	94	5	$s = v_0 t + \frac{1}{2} at^2$	$\theta = \omega_0 t + \frac{1}{2} \alpha t^2$	(6)
	94	6	$v = \sqrt{v_0^2 + 2\, as}$	$\omega = \sqrt{\omega_0^2 + 2\, \alpha\theta}$	(7)
12	122	7	$a = -\frac{4\,\pi^2}{T^2} s$	$\alpha = -\frac{4\,\pi^2}{T^2} \theta$	(8)
	123	8	$T = 2\,\pi\sqrt{\frac{s}{-a}}$	$T = 2\,\pi\sqrt{\frac{\theta}{-\alpha}}$	(9)

It is of great importance to comprehend the situations described in the equations at the right. One way of cultivating such comprehension is to state the situation in words, thus: Equation (3) gives the total number of radians, θ, described by a rotating body possessing an angular acceleration of α radians per second per second, during the t seconds elapsed after starting from rest. A similar verbal formulation may be made for each of the other equations to great advantage.

The Analogy of Torque with Force

Besides the conversion of distances, speeds, and accelerations into corresponding rotational entities, there is a second type of conversion which is involved. The groundwork for this has already been laid in Chapter 6, where the conception of *torque* was introduced. It was noted there that torque was related to rotatory acceleration in the same way that force is to translational. Hence, in converting translatory equations to rotatory, it will be necessary to substitute *torque* (L) for force (f) wherever the latter occurs. The two principal examples encountered up to this point are as follows.

From Chap.	Page	Eqn.	Translation	Rotation	
12	123	9	$f = -ks$	$L = -k\theta$	(10)
14	145	4	$W = fs$	$W = L\theta$	(11)

One of the principal problems encountered in statics is that of the composition and resolution of forces. It is natural to inquire whether the analogy between forces and torques extends to the composition and resolution of torques. Intuitively, it is rather difficult to conceive how torques can be either combined or resolved into components in the sense in which forces have been seen to be subject to these operations.

But as has already been observed, intuition is a weak reed upon which to lean when rotation is under consideration. The fact is that torques are as amenable to composition and resolution as are forces. Since torques possess both magnitude and direction, they are vectors and as such are subject to the same laws of vector addition as are forces, displacements, velocities, and accelerations.

The Analogy of Moment of Inertia with Mass

Besides the transformation of linear displacements into angular displacements and forces into torques, there is a third and last conversion which is involved before translatory relations can be applied unreservedly to rotation: namely, that of mass into its rotatory analogue. The fundamental dynamical equation of translatory motion has been seen to be

$$f = ma,$$

the algebraic formulation of Newton's second law of motion. As has already been seen, to correspond to this, a rotational equation would transform the force f to a torque L and the linear acceleration a to an angular acceleration α. The remainder of the problem is to find or to define an attribute of a rotating body which is related to torque and angular acceleration in the same way that mass, in the above equation, is related to force and linear acceleration.

The first man apparently to solve the problem with a full recognition of its significance was Euler, in 1765, though Huygens had treated it successfully, albeit somewhat casually, in 1673. Euler (37:Cap. III) set up the definition of the required entity and named it *moment of inertia*. Pending a consideration of its properties, which will be presented shortly, it may be tentatively introduced under the notation I, which is commonly used. It becomes possible immediately to write the rotational equation which corresponds to $f = ma$, namely,

$$L = I\alpha.$$

In the same way the other translational relations involving mass, which have been developed in preceding chapters, may now be converted into their rotational equivalents, as in the following table. The preceding pair of equations is included for the sake of completeness.

From Chap.	Page	Eqn.	Translation	Rotation	
10	96	9	$f = ma$	$L = I\alpha$	(12)
12	123	11	$T = 2\pi\sqrt{\frac{m}{k}}$	$t = 2\pi\sqrt{\frac{I}{L_0}}$	*(13)
13	139		$K.E. = \frac{1}{2}mv^2$	$K.E. = \frac{1}{2}I\omega^2$	(14)

* L_0 is an elastic constant of the "torsion pendulum" to which equation (13) applies. As in equation (9) (page 123), in which k is the ratio of any force f to the corresponding displacement x, so L_0 is the ratio of any torque, L, to the corresponding angular displacement or twist, θ, for the given torsion pendulum.

What Moment of Inertia Depends Upon

The moment of inertia is thus an entity which, in rotational phenomena, plays a rôle entirely analogous to that played by mass in translational phenomena. It is not, however, quite as definite an attribute as mass, since a given object may have an indefinite number of moments of inertia, depending upon the relative position of the axis about which it is rotated, whereas there is only one value for the mass of any object. But for a given object, rotating about a given axis, the value of the moment of inertia is perfectly definite.

It is, of course, as necessary that moment of inertia shall be measurable and statable in numerical terms as it is that mass or any of the other physical entities shall possess that characteristic. One of the ways in which the measurement of the moment of inertia may be effected is by the application of equation (12). If the object whose moment of inertia about a given axis is desired is set in rotation about that axis by the application of a measured torque, and the angular acceleration thereby produced is observed, the ratio of torque, L, to angular acceleration, α, will obviously give the value of I, the moment of inertia.[1] An approximate measure of the moment of inertia of any ordinary object may be made by simply giving

[1] Due attention must, of course, be paid to units. For equation (12), as for all other equations connecting force and motion, to be applicable, the so-called absolute units must be used. For the system prescribed in this text, these would be forces in newtons, masses in kilograms, distances in meters.

the object a twist about the prescribed axis. For example, a book, grasped in the middle of one side and twisted quickly will show a certain sluggishness about taking up the rotation imparted to it, and this will give a rough muscular estimate of its moment of inertia about a central axis in its plane. This sluggishness will be found much more pronounced if the same book be grasped at a corner and rotated about one of its edges. The moment of inertia for this case is, in fact, four times that of the previous case. The dependence of the value of the moment of inertia upon the position of the axis is especially pronounced for long narrow objects. A broomstick turned about its own axis presents an almost negligible inertia; but if it is grasped by the center and rotated about an axis perpendicular to its length, a man must be unusually strong to oscillate it through ninety degrees more than about twice a second.

Thus the value of the moment of inertia of an object depends upon the relative position of the axis. Naturally, it depends also upon the value of the mass. If the broomstick had been an iron bar of the same dimensions, the moment of inertia about any axis would have been greater than that of the broomstick in the ratio of the masses. Hence, as the matter is commonly stated, the value of the moment of inertia of an object depends both upon the mass and upon its distribution around the axis of rotation. This suggests the question whether it is possible to compute moments of inertia about given axes, merely from knowledge of masses and dimensions.

Calculation of Moment of Inertia for the Simplest Cases

The answer to the foregoing question is in the affirmative. For regular solids, the computation of moment of inertia from mass and dimensions is possible. For a few cases the process is extremely simple. For most, however, mathematical processes are involved which are not found in the equipment of the average man-in-the-street. Nevertheless, the principle involved is sufficiently simple, so that it will repay some attention.

To develop this principle, consider a particle of mass m at the end of a rod of length r and of negligible weight pivoted at the end opposite the mass (Fig. 141). A force f acts upon it, always perpendicularly to the rod. The torque thereby produced will have the value fr. As regards the force itself and the acceleration, a, which it produces in the mass, m, Newton's second law of motion applies, that is,

$$f = ma. \qquad (15)$$

It is now desirable to show how equation (12) may be developed out of this, without using the somewhat arbitrary analogies of force with torque, linear with angular acceleration, and mass with moment of inertia, which have characterized the preceding treatment. One step in this direction is quite obvious. Multiplication of both sides of equation (15) by r will give on the left fr which is the torque L, that is,

$$fr = L = mar.$$

Also the angular acceleration α can be introduced by noting that in radian measure, $a = \alpha r$, whence

$$L = m \cdot \alpha r \cdot r = mr^2\alpha. \tag{16}$$

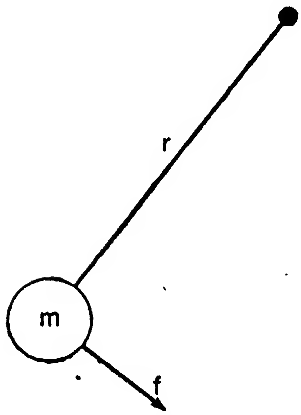


FIG. 141. CALCULATING THE MOMENT OF INERTIA OF A PARTICLE

Now comparing equation (16) with equation (12) it will be seen that the two are of the same form except that for the I which appears in (12) the corresponding term in (16) is mr^2. The natural and indeed almost the only possible inference is that the moment of inertia I of a particle of mass m concentrated at a distance r from the axis of rotation has the value

$$I = mr^2. \tag{17}$$

It will be fairly evident that equation (17) would still represent the moment of inertia of the mass m, if the mass, instead of being concentrated at a single point, were extended around into a circular filament having its center on the axis and being contained in a plane perpendicular to the axis. This extension is permissible because all parts of the mass of the ring are still at a distance r from the axis. Again, and for the same reason, equation (17) also applies to a thin hollow cylinder whose geometrical axis coincides with the axis of rotation.

Moments of Inertia of Geometrical Forms

It is not possible to deal as simply as this with any other geometrical forms. The difficulty is that in all other objects the material constituting them, instead of being concentrated at one distance from the axis of rotation, is distributed. Approximate values of their moments of inertia can be secured by dividing such objects, in imagination, into sections so small that their masses may be considered concentrated at points, multiplying each such elementary mass by the square of its distance from the axis of rotation, and finding the sum of all these products. The smaller the sections, the more accurate the result, the ideal case of perfect accuracy being the division of the object into infinitely small sections and the addition of the resulting infinitely large number of products. The logical machinery for such a process is found in a branch of mathematics known as the integral calculus. The possession of this intellectual tool cannot be assumed on the part of every reader (though many will no doubt have it). It would be possible to deduce the moment of inertia of certain geometrical forms in the way that early investigators did, namely, by indulging in laborious detours around the mathematical obstruction. It is hardly worth while, however, and the computation of moments of inertia of the various standard geometrical forms will simply not be undertaken. The following table gives the values for certain standard forms.

Object	Axis Through	Moment of Inertia
thin rod	center perpendicular to length	$\frac{1}{12}ml^2$
thin rod	end perpendicular to length	$\frac{1}{3}ml^2$
rectangle	center perpendicular to plane	$\frac{1}{12}m(a^2+b^2)$
rectangle	center parallel to side a	$\frac{1}{12}mb^2$
disk or solid cylinder	center (geometrical axis)	$\frac{1}{2}mr^2$
disk	center in plane	$\frac{1}{4}mr^2$
ring or hollow cylinder	center (geometrical axis)	mr^2
sphere	center	$\frac{2}{5}mr^2$

The Radius of Gyration

It will be evident that in the case of the solid cylinder, for example, a thin hollow cylinder of the same mass but with a radius equal to $\frac{1}{\sqrt{2}}$ of that of the solid cylinder would have the same moment of inertia, namely, $m\left(\frac{r}{\sqrt{2}}\right)^2$. All objects may be considered similarly replaceable by equivalent hollow cylinders. The radius of the equivalent hollow cylinder goes by the inappropriate name of *radius of gyration*, possibly suggested by Huygens' cumbersome term *distantia inter axem et centrum oscillationis.* The "center of gyration" of an object is remotely analogous to the center of gravity, some of the properties of which were developed in Chapter 6, in that, for the purpose of rotation about a given axis, all the mass of the body may be considered concentrated at one distance from the axis. If this distance k be known, the value of the moment of inertia I of a body of mass m may be immediately written:

$$I = mk^2. \tag{18}$$

It would be interesting and quite simple to deduce the value of the radius of gyration for all of the forms in the foregoing table, a task which is left to the reader.

The Center of Oscillation

Knowledge of the radius of gyration of a compound pendulum simplifies the problem already referred to (on page 128) of the length of the equivalent simple pendulum. In his *Horologium Oscillatorium*, Huygens formulated this problem along with several others associated with the compound pendulum (59:117 ff). The inadequacy of the mathematical methods then available necessitated a decidedly roundabout type of solution, but he finally arrived at the conclusion that the length l of a simple pendulum having the same period as a given compound pendulum was

$$l = \frac{\text{square of radius of gyration}}{\text{distance from axis to center of gravity}}$$

Huygens coined the term *center of oscillation* to describe the point on a compound pendulum which would lie abreast of the bob of an equivalent simple pendulum. From the foregoing table of moments of inertia, it is possible to show, for example, that for a composite pendulum composed of a thin rod swinging from one end, the length of the equivalent simple pendulum is two thirds the length of the rod. For the case of a spherical bob at the end of a long suspension, the center of oscillation is much closer to the center of gravity than for a pendulum composed of a rod, but the two are still not identical. This is in part the origin of the difficulty in identifying exactly the length of simple pendulum (see page 128).

An interesting application of the principles of rotational inertia may be found in the *gyroscope*, a model of which is illustrated in Figure 142. This is so mounted that in addition to its spin (about OX in Figure 143) it may also turn about OY and OZ. If a torque be applied to such a gyroscope while it is not spinning, say by hanging a weight F on the end of the axle, rotation about OY will occur. But if the gyroscope is spinning in the direction shown, the consequent rotation will actually be about OZ and is termed the *precession*. At first such behavior seems paradoxical. To understand it consider the accelerations of (and hence the forces on) various portions of the gyroscope.

Look at the gyroscope from the point X (Fig. 144). A weight on the end of the axle (now projecting toward the reader) would, in the absence of spin, cause rotation about OY, the top of the gyroscope moving toward the observer and the bottom moving inward. Consider now the effect of

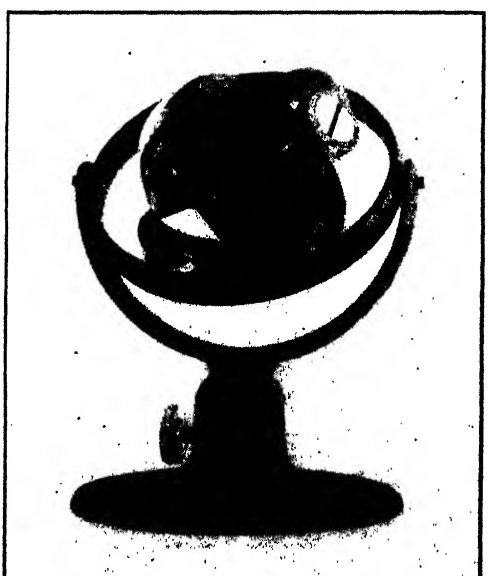


FIG. 142. A GYROSCOPE MOUNTED IN GIMBAL RINGS
(Sperry Gyroscope Company, Inc.)

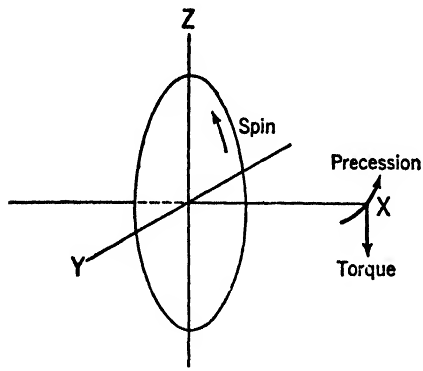


FIG. 143. ELEMENTS OF GYROSCOPIC MOTION

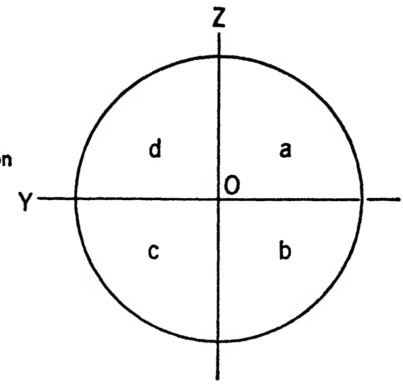


FIG. 144. GYROSCOPE VIEWED FROM THE POSITIVE END OF THE *OZ* AXIS

the inertia of a particle in each of the four quadrants in resisting this rotation about *OY*, and producing, as will appear, a rotation about *OZ* instead.

A particle in quadrant *a* is being carried by the spin constantly further away from the *OY* axis, and hence any rotation about *OY* would cause this particle to move faster (out of the page) as it receded from that axis. Its inertial reluctance to increase its speed out of the page is the equivalent of a force *away from* the observer. A particle in quadrant *b* is at the same time getting closer to *OY* on account of the spin, and hence any rotation about *OY* would cause it to move more slowly away from the observer. Its inertia, acting to maintain this motion of recession, consequently provides the equivalent of another force away from the observer, as was that on *a*.

Similar reasoning about the inertial reactions of particles in the quadrants *c* and *d* will show that they are in the opposite direction; that is, *toward* the observer. Clearly these two pairs of reactions will cause the spinning gyroscope to rotate about the axis *OZ*, if free to do so, instead of about *OY* as it would under the given torque if not spinning. That is to say, under a torque about *OY* the precession will be about *OZ*.

The precessional response being at right angles to the torque in a gyroscope is at first puzzling. At first inspection it appears to contravene Newton's second law. But the paradox is only superficial, as the foregoing analysis shows. The directional relation between spin, torque and precession is sometimes formulated as follows:

> Let the torque applied to a gyroscope be represented by a force on the axis and perpendicular to it (Fig. 143). Rotate the vector representing this force through 90° in the direction of spin of the gyroscope to find the direction of precession.

The foregoing "gyroscope rule" deals only with the direction of the precession. The magnitude of the precessional angular velocity Ω, as related to torque L and spin velocity ω for a gyroscope having a moment of inertia I, is given by the relation

$$\Omega = \frac{L}{I\omega}. \tag{19}$$

To deduce this, consider one of the attributes of uniform circular motion. In such motion tangential velocity is associated with radial acceleration and hence radial (or "centripetal") force. Thus the force and the velocity are at right angles analogously to torque and precession in the gyroscope. The equations of uniform circular motion will make this clear. From equation (4) on page 110,

$$a = \frac{v^2}{r} = v \times \frac{v}{r} = v\omega, \tag{20}$$

where a represents the radial acceleration of an object traveling in a circle; v represents the tangential velocity of the same object and ω represents the angular velocity of the same object about the axis of rotation. The acceleration and the velocity are mutually perpendicular as observed above. Multiplying both sides by m (the mass of the revolving object),

$$ma = mv\omega$$

or by Newton's second law

$$f = \text{momentum} \times \omega. \tag{21}$$

This is a translatory motion, the translation being in a circular path. Like the other translatory equations previously converted into equations describing corresponding rotatory motions, change force f to torque L, momentum mv to angular momentum $I\Omega$ and the result is

$$L = I\Omega\omega. \tag{22}$$

Solved for Ω this is equation (19). The angular velocities must both be expressed in radians per second and the data on which L and I depend must, in the metric system, be in meters, newtons and kilograms.

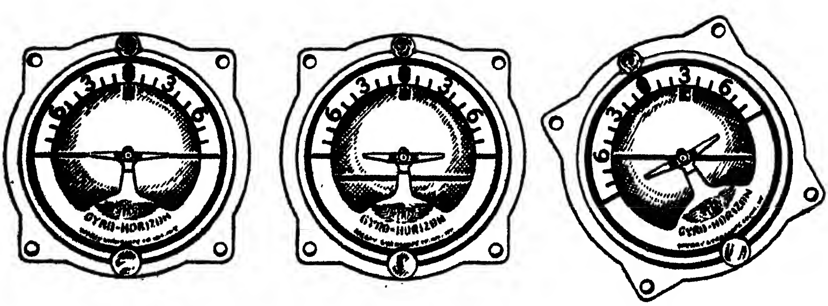


FIG. 144a. GYRO-HORIZON FOR THREE POSITIONS OF THE PLANE: LEVEL FLIGHT, CLIMB, AND BANK

Equation (19) applies to all cases of gyroscopic motion. This includes the motion of an ordinary top, the "precession of the equinoxes" of the earth, the upsetting torques on motorcycles when negotiating turns, the "somersaulting" tendency of a car when turning too sharply at high speed, and all similar phenomena. In recent years the gyroscope has been applied in many ways, especially to navigation and aviation. The Gyroscopic Compass and the Gyroscopic Stabilizer are well known. The Gyro-Horizon, the Directional Gyro, and the Automatic Pilot are common devices in airplanes. All depend upon either gyroscopic inertia or precession or upon the combined action of the two. Figure 144a shows the dial of a gyro-horizon for three positions of the plane: level flight, climb, and bank. Its utility when the actual horizon is obscured requires no emphasis.

Questions for Self-Examination

1. Describe the three changes that are made in translational equations to convert them to rotational equations and give examples.
2. For the following translatory equations, state the corresponding rotatory equations; then state the substance of three of the latter in words: $V = at$, $v = v_0 + at$, $T = 2\pi\sqrt{\frac{s}{-a}}$, $W = fs$, $f = ma$, $K.E. = \frac{1}{2}mv^2$.
3. Write a short exposition on the nature of the entity known as moment of inertia.
4. Derive the equation $I = mr^2$ for the moment of inertia of a hollow cylinder, starting with $f = ma$.
5. Define "radius of gyration" and "center of oscillation."
6. State the "gyroscope rule." Show how it applies to the flywheel of an automobile rounding a corner; to the wheels of a motorcycle making a turn.
7. Tell the meaning of each term of the precession equation (equation 19) and derive the equation.

Problems on Chapter 15

1. The mass of the earth is about 6.6×10^{21} tons and its radius about 4000 miles. Calculate the moment of inertia of the earth about its axis of rotation. 23×10^{38} lb-ft^2.
2. If the friction of the tides slows down the rotation of the earth to such an extent that a hundred years from now the day will be about one second longer than it is now, how great is the drag due to tidal friction? $4.7 \cdot 10^8$ tons.
3. A sphere rolls along a level surface. Find the ratio of its rotational kinetic energy to its translational kinetic energy. 2/5.
4. A ball is launched down a bowling alley, without initial rotation, at a speed of 6 m/sec. The coefficient of friction is .2. How much time elapses before sliding ceases, the subsequent motion being one of pure roll? (*Hint:* Pure roll sets in

when the increasing speed of the surface due to rotation becomes equal to the diminishing speed of the center of the ball.) .88 second.

5. If a rapid "backward" spin is imparted to the ball of the preceding problem at the instant that it is launched, consider the nature of the subsequent motion.

6. Find the moment of inertia I of a hollow cylinder of mass m kilograms and mean radius r centimeters in kilogram-meters squared.

	m	r	I
6.	1	10	.01
	1	15	.02
	1	20	.04
	1	25	.06

	I	n	N	L
7.	.01	1	5	.0063
	.01	2	5	.025
	.01	3	5	.057
	.01	4	5	.1

7. A wheel whose moment of inertia is I is spinning with an angular speed of n revolutions per second. What is the frictional torque L in newton-meters that will bring it to rest in N revolutions?

8. A force of f kilograms acts on the axle of a wheel of radius R centimeters and mass M kilograms while the wheel turns through N revolutions. The radius of the axle is r centimeters. Find the resulting angular speed n in revolutions per second. Assume the mass of the wheel to be concentrated at the rim. For such a case $I = MR^2$.

f	r	R	M	N	n
.5	1	10	1	20	5.6
1.	1	10	1	20	7.9
1.5	1	10	1	20	9.7
2.	1	10	1	20	11.

9. A ship-stabilizing gyroscope of radius r feet and of mass m tons is brought up to a speed of n revolutions per second by a motor. What horsepower p must be exerted to accomplish this in t hours? Assume the weight of the gyroscope to be concentrated at the rim, and neglect friction and air resistance.

	r	m	n	t	p
9.	3	10	20	10	2.2
	4	20	17	10	5.7
	5	50	15	10	17.
	6	120	13	10	45.

	h	v_1	v_2
10.	.25	1.6	1.8
	.5	2.2	2.6
	.75	2.7	3.1
	1.	3.1	3.6

10. A thin hollow cylinder rolls without sliding down an incline of vertical height h meters. What is its speed v_1 in meters per second at the foot of the incline? What would be the speed v_2 of a solid cylinder? What effect would a change in size or mass of the cylinder have on the result?

11. A thin hollow cylinder rolls down an incline making an angle of β degrees with the horizontal. What is its acceleration a_1 in m/sec²? What would have been the acceleration a_2 of a solid cylinder? (The moment of inertia of the thin hollow cylinder about its line of contact with the incline of $2\ mr^2$ and of a solid cylinder is $3/2\ mr^2$.)

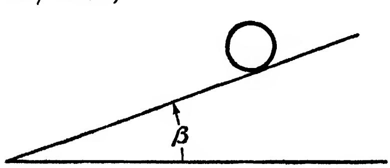


FIG. 145

β	a_1	a_2
30	2.5	3.3
20	1.7	2.3
10	.85	1.2
5	.43	.57

12. An m-kilogram spool is formed of two heavy disks joined by a cylindrical body of negligible mass. A cord wound around the body is pulled horizontally with a force of f kilograms causing the spool to roll. Which direction will the spool roll? The radius of the disk-shaped ends is R centimeters and of the body is r centimeters. What is the acceleration a of the spool in m/sec²?

m	*f*	*R*	*r*	*a*
1	.1	5	4	.13
1	.1	5	3	.26
1	.1	5	1	.52
1	.1	5	0	.65

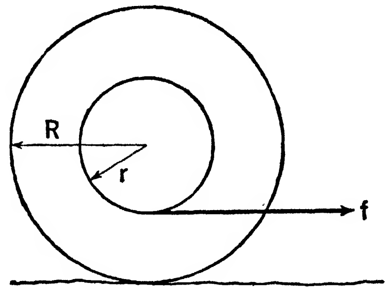


FIG. 146

13. A bicycle whose frame has a mass of M kilograms and wheels of m kilograms each is pushed forward with a force of f kilograms. How many seconds t would be required to bring it up to a speed of v kilometers per hour? Assume the mass of the wheels to be concentrated at the rims.

M	*m*	*f*	*v*	*t*
6	1.5	2	30	5.1
8	2.	3	30	4.5
10	2.5	4	30	4.3
12	3.	5	30	4.1

14. A bullet of mass m grams and velocity V meters per second strikes a pivoted solid disk in the manner shown. The disk weighs M kilograms and has a radius of R centimeters. Calculate the angular velocity n in revolutions per second communicated to the disk. Assume the impact to be inelastic and neglect the weight of the bullet in calculating the moment of inertia of the disk.

m	*V*	*M*	*R*	*r*	*n*
5	200	1.0	20	2	.16
5	200	1.0	20	6	.48
5	200	1.0	20	10	.79
5	200	1.0	20	20	1.6

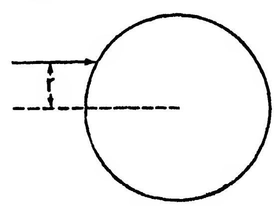


FIG. 147

15. A gyroscope weighing M grams and having a radius of r centimeters is spinning at N revolutions per second. A weight of m grams is hung on the axis at a distance d centimeters from the center. What is the precessional angular velocity n in revolutions per minute? Assume the weight of the gyroscope to be concentrated at the rim.

M	*r*	*N*	*m*	*d*	*n*
500	5	50	100	5	1.2
500	5	60	100	5	1.0
500	5	70	100	5	.85
500	5	80	100	5	.74

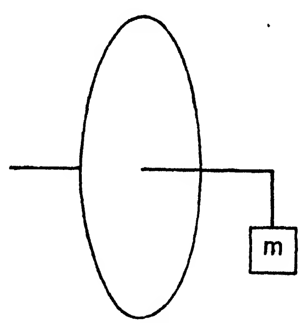


FIG. 148

16. A gyroscopic ship stabilizer of mass M tons and radius r feet is given a precessional angular velocity of ω radians per second by the precession motors. The gyroscope is spinning at a rate of N revolutions per second. What torque L in ton-feet is produced on the ship? Assume the mass of the gyroscope to be concentrated at the rim.

M	r	ω	N	L
10	3	.1	20	35.
20	4	.1	17	110.
50	5	.1	15	370.
120	6	.1	13	1100.

17. A motorcycle whose wheels weigh m kilograms each and have a radius of r centimeters rounds a curve of radius R meters at v kilometers per hour. What torque L in kilogram-meters acts to upset the machine as a consequence of the gyroscopic action of the wheels? If the center of gravity is h meters from the ground, what upsetting force F in kilograms is acting at this point? (This is, of course, in addition to the centrifugal force. It does not take account of the similar action of the fly-wheel, which in one- and two-cylinder machines has the same effect.) Compare these values of F with those of problem 13 of Chapter 11.

m	r	R	v	h	L	F
10	35	15	25	.5	2.3	4.6
10	35	20	30	.5	2.5	5.0
10	35	25	35	.5	2.7	5.4
10	35	30	40	.5	2.9	5.9

HEAT

CHAPTER 16

Temperature and Thermal Expansion

Thermometers and Temperature Scales

The science of heat, unlike that of mechanics, had no significant developments until the seventeenth century. It began when the first thermometer was devised. The first of these to survive was made by Galileo, though there are records indicating that similar instruments had been made centuries earlier. Galileo's thermometer was described as follows by one of his pupils (132:83):

> Galilei took a glass vessel about the size of a hen's egg, fitted to a tube the width of a straw and about two spans long; he heated the glass bulb in his hands and turned the glass upside down so that the tube dipped in water contained in another vessel. As soon as the ball cooled down the water rose in the tube to the height of a span above the level in the vessel. This instrument he used to investigate the degrees of heat and cold.

Such a thermometer, easily made, will be found surprisingly sensitive, though rather erratic because its readings depend upon atmospheric pressure as well as upon temperature. But neither the existence of this pressure nor its variations were recognized at the time of Galileo. These were the discoveries of Torricelli and Pascal (pages 83, 85).

There is some significance in the fact that scientific comprehension of heat phenomena began when the first instrument for measuring temperature was devised. About fifty years ago this close relation between knowledge and measurement was expressed by Lord Kelvin as follows:

> when you can measure what you are speaking about and express it in numbers, you know something about it. . . .

Galileo's crude thermoscope was not itself capable of translating temperatures into numbers, to be sure. But it was the parent of instruments that did. Prior to that the only method of judging temperature was by sensation, which is notoriously unreliable.

Between 1592 and 1742, just a century and a half, the thermometer developed into its present form. The second stage of its development, in 1632, consisted of inverting the instrument and filling it with water, which thus became the new thermometric substance. Though far less sensitive than Galileo's form, it no longer responded to fluctuations in barometric pressure, even though the tube was open at the top. Twenty-five years later, the tubes were commonly sealed at the top and alcohol was substituted for water. The first use of mercury was in 1659 (25:100). It had the advantage of being opaque, of not wetting the tube, of conducting heat readily, of having a low freezing point and a high boiling point, and, as was discovered later, of changing its volume more nearly in proportion to changes of temperature than other liquids in common use. It is little wonder that one physicist enthusiastically exclaimed, "Surely nature has given us this mineral for the making of thermometers" (25:117).

The growth of thermometric scales, the final stage in the development of the thermometer, was a chaotic process. A commentator in 1779 enumerated nineteen different scales which were in current use. Three have survived: the *Fahrenheit*, devised before 1708 by a Danish astronomer named Olaf Roemer [1] (cf. page 288) and later given publicity by Daniel Fahrenheit; the *Réaumur*, devised by a Frenchman of that name in 1730; and the *centigrade*, devised by a Swede named Celsius in 1742 which is now the standard scale of temperature in the metric system. The three scales are represented in Figure 149. The centigrade and Réaumur both use the temperature of melting ice as their zero point, while the Fahrenheit denominates that temperature as 32° above zero. The Fahrenheit zero was set arbitrarily by Roemer. The centigrade sets 100° as the boiling point of water, and the Réaumur 80°. The same temperature happens to be represented by 212° on the Fahrenheit scale. Fahrenheit's original upper point of reference was not that temperature, but was instead the temperature of the human body. He called this (38:6)

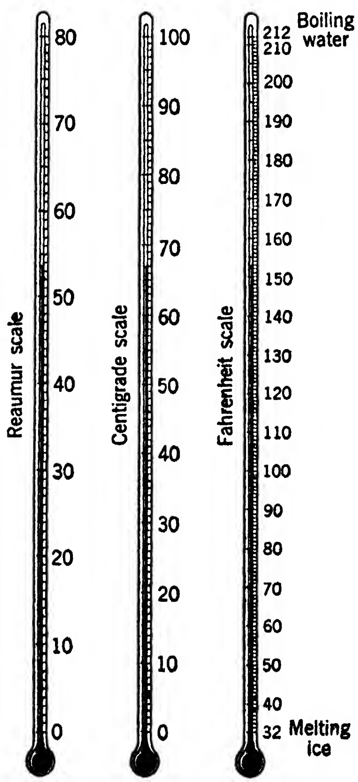


FIG. 149. COMPARISON OF THERMOMETER SCALES
(Courtesy of Taylor Instrument Companies.)

> . . . the 96th degree, and the alcohol expands to that point if the thermometer be held in the mouth or armpit of a healthy person.

It is a great pity that a scale established so awkwardly and so inaccurately should be the one uniformly used in English-speaking countries.

The point of departure in the establishment of any thermometer scale is

[1] *Annals of Science*, *2*, 133 (1937).

a pair of *fixed points*, temperatures which can be realized with maximum precision without reference to any thermometer whatever. By common practice, these are now the freezing and boiling points of water at "normal" atmospheric pressure. Roemer (27a:40) is said to have been the first to suggest that two fixed points were required to establish a thermometric scale, not merely one. When Celsius established the centigrade scale in 1742 by dividing the temperature range between the freezing and boiling points of water into 100 equal parts, he chose the *boiling point* as his zero and measured *down* to the freezing point, which he called 100°. This was soon inverted by a contemporary of his, however, producing the centigrade scale as we know it now (25:118).

The use of expansion as a measure of change of temperature has its limits, of course. It becomes inapplicable both at low temperatures where thermometric substances freeze and at high temperatures where they vaporize. Other methods, electrical and optical, take the place of expansion in these temperature ranges. The instruments used in that connection are too technical to justify description here.

The Effect of Temperature on Gases

The behavior of gases under changes of temperature happens to be much simpler than that of liquids or solids, and it is fortunate that this happened to be the first case studied. The earliest study in this field, made by Amontons (1663–1705) in 1699 (77:129), concluded

> that unequal masses of air under equal weights increase equally the force of their spring for equal degrees of heat.

The context indicates that by "increase in force of spring" the writer means ratio of change in pressure to the original pressure so that the somewhat hazy seventeenth-century phraseology means that air always shows the same value of this ratio for a given temperature change. Let the change in pressure be represented by ΔP,[1] the original pressure (conventionally taken as the pressure at 0° C.) be represented by P_0 and the change in temperature by Δt. Then "increase in force of spring for equal degrees of heat" becomes

$$\beta = \frac{\Delta P}{P_0 \Delta t}. \qquad (1)$$

This ratio, denominated by β in equation (1), is called the ***pressure coefficient.***

Evidently the value of the pressure coefficient β can be determined by measuring the pressures of air at each of two observed temperatures while maintaining the volume at a constant value. This comes out to be approximately $\frac{1}{273}$, a value which has been found to apply, not only to air, but with considerable fidelity to all "permanent" gases, that is, gases

[1] Capital delta (Δ) is frequently used to express the idea of "change of," especially when small changes are implied.

which liquefy at such low temperatures that their liquefaction was considered impossible until fairly recently. For such gases the value of β is also quite constant over a wide range of temperature. At one time hydrogen was supposed to perform more consistently in this respect than any other gas. For this reason it has been officially designated as the standard thermometric substance. All thermometers are accordingly standardized ultimately by a hydrogen thermometer in which the volume of the gas is held constant, the change of pressure measuring the change of temperature.

The Concept of "Absolute Zero"

Equation (1) may be modified into the following form. Let the upper temperature be t and the lower 0° C., whence Δt becomes t. Let the higher pressure be P_t and the lower P_0, whence $\Delta P = P_t - P_0$. Making these substitutions and solving for P_t

$$P_t = P_0(1 + \beta t). \tag{2}$$

Since β has the value $\frac{1}{273}$, equation (2) states that the pressure of a gas is augmented by $\frac{1}{273}$d of its zero-degree value for every degree centigrade that the temperature rises and is diminished in the same measure for every degree that the temperature falls. Thus, at 273° above zero the pressure of a gas should have double the zero-degree value, and at 273° below zero the pressure should disappear entirely. This latter hypothetical state of affairs has given rise to the concept of the *absolute zero*, which is a useful concept as long as one does not demand too much from it. So-called absolute temperature ("°K") then is expressible by adding 273° to the centigrade temperature. Thus 0° C. is 273° K, and 100° C. is 373° K, and, of course, −273° C. is 0° K.

It may not be difficult to imagine the pressure of a gas dropping to zero, but after this has occurred, the substance can scarcely be termed a gas. Moreover, the fact that gases behave in the ordinary range of temperatures as though their pressure would become zero at some temperature outside of that range does not necessitate the conclusion that this would actually occur. Amontons, who was the one to get the idea first (in 1703), could be excused for drawing unwarranted conclusions, but it is now known that at low temperatures the value of the pressure coefficient β departs from its conventional value of $\frac{1}{273}$ for even the most "permanent" of the gases. Nevertheless the concept of absolute zero persists. Its presumed value, extrapolated from data taken at higher temperatures, is −273.18° C. The lowest temperature actually reached as yet is within .005° of that point.[1]

The Expansivity of Gases

If, instead of holding the volume of a gas constant and allowing the pressure to increase as the temperature rises, we hold the pressure constant by allowing volume to change, the proportional change in volume that occurs per degree change in temperature could logically be termed the volume

[1] *Proceedings of the Royal Society*, *149*, 152 (1935).

coefficient, as distinguished from the pressure coefficient previously discussed. The term *expansivity* is preferable, however, since it is more general, being applicable to liquids and solids as well as gases. In terms parallel to those defining the pressure coefficient — equation (1) — expansivity of a gas may now be defined as follows: If a change of volume ΔV at constant pressure accompanies a change of temperature Δt, the volume at 0° C. being V_0, then the mean expansivity α is

$$\alpha = \frac{\Delta V}{V_0 \Delta t}. \qquad (3)$$

Corresponding to equation (2) with analogous notation,

$$V_t = V_0(1 + \alpha t). \qquad (4)$$

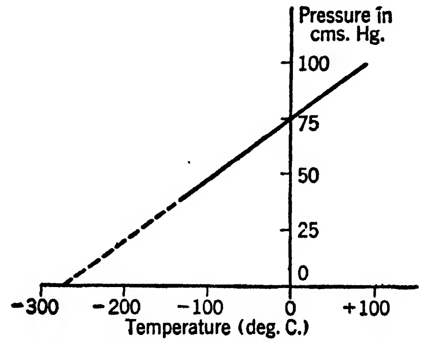


FIG. 150. VARIATION OF PRESSURE WITH TEMPERATURE FOR AIR

The first reliable measurements of expansivity of gases were made nearly a century after pressure coefficients had been identified and measured. They were made by Jacques A. C. Charles (1746–1823) about 1787 but were never published. Fifteen years later Joseph L. Gay-Lussac (1778–1850) performed the same experiments with better technique and results (106:27 ff.). He concluded (106:47)

> that all gases, speaking generally, expand to the same extent through equal ranges of heat; provided all are subject to the same conditions.

This is analogous to the fact that the pressure coefficient is also substantially the same for all permanent gases. The question arises whether the two coefficients are also equal to each other. Gay-Lussac's value for α was .00375 (106:44). This value was refined by later observers to .00366, which is precisely $\frac{1}{273}$, thus answering the above question in the affirmative.

The Law of Boyle and Gay-Lussac

When Boyle's law was considered in Chapter 7, it was noted that the law was applicable only if the temperature of the expanding or contracting gas was held constant. Now that the response of a gas to changes in temperature is known, this limitation may be removed. If we imagine a body of gas undergoing successively a change of temperature (either at constant volume or constant pressure) and a Boyle's-law change of pressure and volume, it may easily be shown that either

$$PV = P_0V_0(1 + \beta t) \text{ (temperature change at constant volume)}$$

or $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ (5)

$$PV = P_0V_0(1 + \alpha t) \text{ (temperature change at constant pressure),}$$

where t is the change of temperature in degrees centigrade, P_0 being the initial pressure, here taken as one atmosphere ($1.013 \cdot 10^5$ newtons/m^2) and

V_0 the corresponding volume in cubic meters (the initial temperature being supposed 0° C.), and P and V the final pressure and volume. It is sometimes put in the form

$$PV = RT \tag{6}$$

where T is the temperature on the absolute scale and R is $P_0V_0/273$. Equation (6) may readily be deduced from equation (5) by giving α or β its value $\frac{1}{273}$ and then changing the temperature scale.

The product P_0V_0 and hence the corresponding value of R will depend on the quantity of gas being compressed and expanded. If some specified mass of gas be considered, say one kilogram, the value of the product, and hence of R, will depend on the density of the gas. If the value of R is known for one gas, it may readily be found for any other through multiplying by the inverse ratio of the densities. But if V is taken as the volume of one gram-molecule (the number of grams of the gas numerically equal to its molecular weight), then V_0 will be the same for all gases, and R becomes a "universal constant." On this basis, the value is $R = 8.313$, pressures being expressed in newtons/m², corresponding volumes being those of one gram molecule of the gas in cubic meters.

The Expansivity of Liquids

Equations (3) and (4) cover the case of the expansivity of liquids as well as gases, except that the value of α is, in general, very much less than that of gases, is different for different liquids, and bears only approximately the same value in notably different temperature ranges even for the same liquid. The pressure coefficient β — equation (1) — has no practical significance in connection with liquids. Liquids are so incompressible that no ordinary container could withstand the pressures necessary to maintain the volume unchanged as temperature rises.

The increase in volume of a liquid with rising temperature, measurably proportional to such rise, is accompanied, of course, by a decrease in its density. Advantage of this is taken in hot-water heating systems, which depend for their action upon the expansivity of water. So-called *convection currents* are set up by the sinking of the more dense cooled water in the upper portions of heating systems into the "boilers" below, thus continuously exchanging places with the water previously heated. Hot-air furnaces work on the same principle, except when circulation is effected by motor-driven fans. The draft in chimneys is similarly induced, and most meteorological phenomena, including winds, cyclones, tornadoes, and cloud formation, are principally large-scale manifestations of convection currents set up by localized expansions and contractions of air.

The Anomalous Expansivity of Water

Though all liquids expand non-uniformly, water, the most common liquid, shows this in a particularly flagrant manner. This is indicated graphically in Figure 151. This shows, with temperatures as abscissas and volumes as ordinates, the volume of a mass of water which occupies unit volume at 4° C. The anomalous behavior of water consists in the fact that the volume is a minimum at this temperature, increasing for lower as well as for higher temperatures. This means that the value of α is negative for temperatures below 4° C., zero at that temperature, and positive for all higher temperatures. Figure 151 may be compared with Figure 150, in which the uniformity of the expansivity of air, indicated by the straightness of the line in the earlier figure, is in contrast with the non-uniformity of the expansivity of water as shown in this one. In both cases the expansivity is proportional to the slope of the curve. The expansivity of water at different temperatures may be estimated from the slope of the curve of Figure 151 and will be seen to vary from approximately $-.0001$ at $-2°$ C., through zero at $+4°$ C., to $+.0001$ at $+12°$ C.

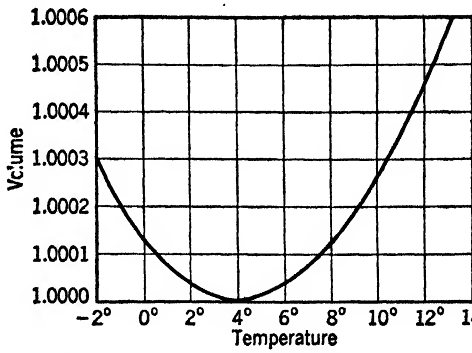


FIG. 151. CURVE OF EXPANSION OF WATER

That ice forms on the tops of bodies of water instead of below the surface is due to this peculiarity. Were it not for the negative value of the expansivity of water just above the freezing temperature, many forms of aquatic life could not exist, for the formation of ice on the bottoms of rivers and lakes would deprive marine life of its food supply during the winters.

The Expansivity of Solids

For gases and liquids, *volume* expansion is the important consideration. For solids, while this must sometimes be taken into account, greater importance attaches to *linear* expansion, that is, expansion along any one of the three dimensions of a solid. The expansivity is usually the same along one dimension as another, though values may be different along the grain of wood or the axis of a crystal than crosswise. Values for solids are less than those for liquids and, like the latter, are not necessarily the same at different temperatures. Negative values, corresponding to that of water below 4° C., are extremely rare in solids, however. Linear expansivity may be defined in terms analogous to those used in equation (1) and equation (3). Thus, if a change of length ΔL accompanies a change of temperature Δt, the length at 0° C. being L_0, then the linear expansivity λ of the material is

$$\lambda = \frac{\Delta L}{L_0(\Delta t)}. \tag{7}$$

Corresponding to equations (2) and (4), with analogous notation

$$L_t = L_0(1 + \lambda t). \tag{8}$$

Values of λ for the common metals run from about .000,025 for aluminum down to .000,000,9 for invar, a nickel steel mixture especially alloyed for low expansivity. They are thus much smaller than expansivities of liquids.

In spite of small expansivity, changes in length of structural materials exposed to seasonal extremes of temperature are sometimes considerable. A mile of railroad is a yard longer in summer than in winter. A thousand-foot bridge changes its length by seven or eight inches. A long steam line requires special appliances to allow for expansion. Unless proper provisions are made, the stresses set up by expansion and contraction would damage the structures.

Ratio Between Linear and Volume Expansivities

It is occasionally desirable to know the volume expansivity of a solid, as when changes in volume of a glass container with changing temperature are required. The volume expansivity is easily shown to be almost exactly three times the linear expansivity. To show this consider a cube the length of whose edges is L_0 at 0° C., the corresponding volume being of course L^3_0. When the cube is heated to $t°$ C. the length of the edges increases to $L_0(1 + \lambda t)$ and the volume becomes V_t. Then

$$\begin{aligned} V_t &= L^3_0(1 + \lambda t)^3 \\ &= V_0(1 + 3\lambda t + 3\lambda^2 t^2 + \lambda^3 t^3). \end{aligned}$$

But since λ is so small, the terms involving λ^2 and λ^3 are negligible in comparison with those preceding.[1] Hence, to a close approximation

$$V_t = V_0(1 + 3\lambda t). \tag{9}$$

Comparison with equation (4) will show that $\alpha = 3\lambda$. (10)

Questions for Self-Examination

1. Identify the stages in the evolution of the mercury-in-glass thermometer.
2. What was the evolution of the "fixed points" necessary in establishing any thermometric scale?
3. Tell the history of the fahrenheit scale; of the centigrade scale.
4. Define and compare the *pressure coefficient* and the *expansivity* of gases.
5. Compare the expansivities of gases, liquids, and solids.
6. Restate Boyle's law for the case where heat of compression is allowed to accumulate and discuss it.

[1] For example, with iron (λ = .000,01)

$$V_t = V_0(1 + .000{,}03\,t + .000{,}000{,}000{,}3\,t^2 + .000{,}000{,}000{,}000{,}001\,t^3).$$

Terms beyond the second in the parentheses are clearly negligible in comparison with the rest for any conceivable temperature range to be met in practice. Hence, to an abundantly sufficient degree of approximation

$$V_t = V_0(1 + .000{,}03\,t).$$

7. Discuss the expansivity of water.
8. Show why the volume expansivity of a solid has three times the numerical value of the linear expansivity.

Problems on Chapter 16

1. A meteorological balloon whose initial volume is 5 cubic meters rises from a level where the pressure of the gas in it is 75 centimeters of mercury at a temperature of 20° C. to a height where the temperature is − 30° C. and the pressure in the balloon due to expansion is 32 centimeters. Find the new volume, assuming the balloon to have expanded freely. 9.7 cubic meters.
2. A glass vessel contains 200 cubic centimeters of alcohol and is full at a temperature of 15° C. How many cubic centimeters will spill out when the temperature is raised to 40° C.? (λ = 000,008 for glass and α = .0011 for alcohol.) 5.4 cc.
3. A steel ball displaces 400 cubic centimeters of water at 0° C. How many cubic centimeters will be displaced at 30° C.? (λ = .000,0105 for steel.) 400.38
4. A wrought iron tire whose inner diameter is 70 centimeters at 0° C. is to be shrunk on a wheel whose diameter is $\frac{2}{3}$ centimeter too large. To what temperature must the tire be heated? (λ = .000,0114 for wrought iron.) 840° C.
5. A clock with a wrought iron seconds pendulum keeps accurate time at 0° C. How much will it lose per day at 20° C.? (λ = .000,0114 for wrought iron.) 10 secs.
6. Assuming that the highest temperature attained by materials exposed to the sun the year around is 50° C. and that the lowest is − 30° C., what allowance in inches must be made for extreme seasonal changes of length in the following cases?

Object	Expansivity	Length (ft.)	Allowance (in.)
steel rail	$12 \cdot 10^{-6}$	30.	.35
steel bridge span	$12 \cdot 10^{-6}$	1700.	20.
cement walk panel	$10 \cdot 10^{-6}$	4.	.04
pine deck boards, length	$5.4 \cdot 10^{-6}$	12.	.06
pine deck boards, width	$48 \cdot 10^{-6}$	.25	.011

7. An L-meter stainless steel tape is correct at 0° C. What correction l in millimeters must be applied when it is used at t° C.? (Take λ = .0000096.)

L	t	l
2	− 10	− .19
5	+ 5	+ .24
10	+ 20	+ 1.9
20	+ 30	+ 5.8

8. An iron steam pipe joins two points L meters apart. The temperature of the pipe varies between t° C. and T° C. What must be the range of motion l in centimeters of an expansion joint to be placed in the line? The expansivity of iron is .000012.

L	t	T	l
200	− 20	180	48.
150	− 10	170	32.
100	0	160	19.
50	10	150	8.4

9. The *linear* expansivities of some metals and their specific gravities at 20° C. are as shown below. Find the specific gravities at T° C.

Metal	Sp.g. at 20° C.	Lin. Exp.	T	Sp.g. at T° C.
aluminum	2.7	22.21 10^{-6}	0	2.7
copper	8.89	15.96	50	8.9
iron	7.86	11.45	100	7.8
tin	7.3	20.94	150	7.2

10. Two strips of metal whose thickness is a centimeters and which have expansivities of α and β respectively, are riveted together forming a straight bar. The temperature is raised t° C. What is the radius of curvature r in meters?

a	t	α	β	r
.1	1	.000018	.000012	170.
.1	100	.000018	.000012	1.7
.1	10	.000064	.000018	2.2
.1	10	.000026	.000007	5.3

11. A collapsible gas tank has a volume of v cubic meters at a temperature $t°$ C. What is its volume V in cubic meters at $T°$ C. if the pressure on the contained gas remains the same?

v	t	T	V
20	0	20	21.5
20	10	30	21.4
20	20	40	21.4
20	30	50	21.3

12. The linear expansivity of glass is .000008. Find the weight of water W contained in a glass flask which is rated at 100 cubic centimeters at 0° C., at the temperatures indicated.

T	Sp.g. Water	W
0°	.99987	99.99
4	1.	100.01
8	.99987	100.01
20	.99823	99.86
50	.98807	96.75

13. A steel drum contains hydrogen at a pressure of p atmospheres at $t°$ C. If the limit of safety of the drum is P atmospheres, what temperature, $T°$ C., will the drum safely withstand, assuming that its strength remains the same?

p	t	P	T
10	−20	20	230
15	0	25	180
20	20	30	170
25	40	35	160
30	60	40	170

14 The reading of a barometer is B_T centimeters at $T°$ C. Knowing that the linear expansivity of the brass scale is .00001900 and that the volume expansivity of the mercury is 0001818, find what the reading B_0 would have been if the temperature had been 0° C.

B_T	T	B_0
75	10	74.9
74	20	73.8
73	30	72.6
72	40	71.5

15. A clock pendulum is of a metal whose coefficient of linear expansion is λ_1. Its length to the bottom of the bob is l meters. A container holds mercury to produce compensation for changes in length of the pendulum with changing temperatures. At what height h in centimeters must the mercury stand? Allow for change in cross-section of the container with changing temperature as well as for change in length of the suspending rod. The coefficient of *linear* expansion of the material of the container is λ_2. The coefficient of *volume* expansion of the mercury is .0001815.

Fig. 152

λ_1	l	λ_2	h
.000011	1	.000011	14
.000011	1	.000025	17
.000018	1	.000011	23
.000018	1	.000025	27

16. A block of metal whose coefficient of expansion is λ and whose Young's modulus is Y newtons per square centimeter, is placed in a vise strong enough to hold its length l constant, and is then heated through $t°$ C. What force f in metric tons does each square centimeter of its cross section exert?

l	λ_1	Y	t	f
10	.000008	$60 \cdot 10^6$	100	4.9
10	.000011	$21.4 \cdot 10^6$	100	2.4
10	.000018	$10.8 \cdot 10^6$	100	2.
10	.000023	$6.5 \cdot 10^6$	100	1.5

CHAPTER 17

Quantity and Migration of Heat

The Distinction Between Heat and Temperature

The first man to make a clear distinction between temperature and quantity of heat was Joseph Black (1728–99) of the Universities of Glasgow and Edinburgh. About the middle of the century [1] he was commenting to his classes on the current view of the nature of heat. Though Black agreed with the prevailing theory of his time that heat was a weightless fluid which flowed from regions of high temperature to those of lower, he disagreed with some of the conclusions that his contemporaries were drawing from this theory. His dissent in one instance was expressed as follows (77:135):

> This is taking a very hasty view of the subject. It is confounding the quantity of heat in different bodies with its general strength or intensity [i.e., temperature] though it is plain that these are two different things, and should always be distinguished, when we are thinking of the distribution of heat.

Black did not coin a name for the fluid which he thought constituted heat. He usually referred to it as the "matter of heat" or, when he was concerned with its amount, as the "quantity of heat," as in the above quotation. Some time during the latter part of his life the name *caloric* was coined to denominate the "matter of heat," and Black's views on the nature of heat have accordingly come to be known as the "caloric" theory. Black himself never used the term, however.

The Concept of Specific Heat

During Black's time there was considerable speculation on how heat distributed itself among the various parts of a region which was at uniform temperature throughout. One theory was that there was the same quantity of heat in every equal part of the volume. Another was that the heat was distributed proportionally to the weights of the various parts of the region in question. Black showed that both surmises were incorrect. He referred to some experiments by Fahrenheit in which water and mercury at different temperatures were mixed. The heat yielded to the mixture

[1] The exact time when Black first formulated this point of view is somewhat in question, since he never published it himself. It was published posthumously from notes taken by his pupils at about the time indicated.

was found not to be proportional either to the volume or to the weight of whichever was the hotter of the two liquids. He stated his conclusion thus (77:136–39):

> This shews that the same quantity of the matter of heat has more effect in heating quicksilver than in heating an equal measure of water, and therefore that a smaller *quantity* of it is sufficient for increasing the sensible heat of quicksilver by the same number of degrees.... Quicksilver, therefore, has less *capacity* for the matter of heat than water has; it requires a smaller quantity of it to raise its temperature by the same number of degrees....
>
> We must therefore conclude that different bodies, although they be of the same size, or even of the same weight, when they are reduced to the same temperature or degree of heat, whatever that may be, may contain very different quantities of the matter of heat; which different quantities are necessary to bring them to this level, or equilibrium with one another.

This idea of Black's of a certain "capacity for heat" characteristic of each substance was new and proved to be correct. His terminology, however, was soon modified, the term *specific heat* being introduced by some of his contemporaries to describe essentially what Black meant by "capacity for heat" (131:183, 204).

In Chapter 8, the convenience of having some standard of reference for densities was pointed out. The density of water being accepted in this rôle, the ratio of the density of any given substance to that of water assumed some importance and was termed the specific gravity of the substance in question. Water has also come to be the standard for specific heats, the specific heat of any substance being defined as *the ratio of the quantity of heat required to warm a given mass of the substance between two temperatures to the quantity similarly required for water.*

Like specific gravity, which changes its value in consequence of the thermal expansion of the substance, specific heat also depends upon temperature. Tables of values of either entity always specify the temperatures for which the tabular values apply.

The Unit of Quantity of Heat

The creation of the concept of quantity of heat necessitates the choice of a unit. The calorists employed as this unit the quantity of heat that must enter a unit mass of water to raise its temperature one degree; or they reversed the definition to involve the heat which must be abstracted from a unit mass of water to lower its temperature one degree, experiment having demonstrated that the quantity of heat involved was the same in either case. This definition is still valid. In metric units, one *Calorie* is defined as *the quantity of heat transferred whenever one kilogram*[1] *of water changes its temperature by one degree centigrade.* As thus defined the Calorie has

[1] An older definition, now going out of vogue, provided for a calorie $\frac{1}{1000}$ as large as this. It was termed the gram-calorie, because a gram was used in place of a kilogram as here. The word "Calorie," representing the kilogram-Calorie, will be capitalized in this book to distinguish it from the gram-calorie.

slightly different values at different temperature levels, since the remark of the preceding paragraph applies to water as well as to other substances. The value of the Calorie fluctuates between extremes of about a half a per cent above and below its mean value. For accurate work a more precise definition is necessary, but it need scarcely be formulated here — partly in view of the fact that there is as yet no complete agreement on what this definition should be.

Ways in Which Heat Is Transferred

On page 176 one of the ways of transferring heat was mentioned, that of convection currents. The idea is almost absurdly simple and would scarcely merit further mention were convection not commonly included in a trilogy of ways in which transference of heat is commonly stated to occur. Convection, involving motion of the medium, is more of a mechanical than a heat phenomenon since the actual heating and cooling of the medium can take place in no other way than by conduction or radiation. Convection is, however, one of the major elements in meteorology, since terrestrial air motions of all kinds are simply convection currents on a huge scale.

Conduction has been defined (82:10) as

> the flow of heat through an unequally heated body from places of higher to places of lower temperature.

In this process, unlike convection, the heated substance does not migrate. Obviously, conduction can occur when the heated substance is in any one of its three states, solid, liquid, or gaseous, while convection is necessarily confined to the last two.

Radiation, or radiant energy, has been similarly defined (*loc. cit.*) as a process by which

> the hotter body loses heat and the colder body receives heat through some intervening medium which does not itself thereby become hot.

The most impressive example of radiation is the transfer of heat from the sun to the earth through regions of space known to be frigid in the extreme, but the process occurs in greater or less measure wherever differences of temperature exist. It is worthy of note that the entity in transit by radiation cannot properly be called heat until absorbed at the end of its journey whereupon it reassumes its ability to elevate temperature. Radiation of heat may be compared to broadcasting a radio signal; indeed, radiation actually is a broadcast, though on wave-lengths much shorter than those to which a

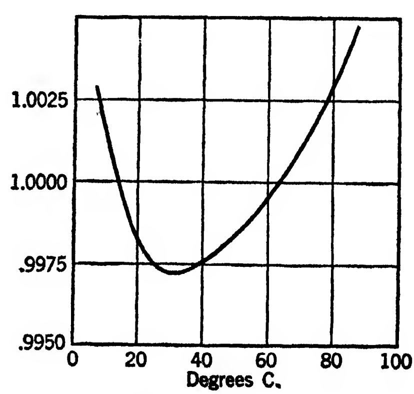


FIG. 153. RELATIVE VALUE OF THE CALORIE AT DIFFERENT TEMPERATURES

radio receiving set can be tuned. Sounds are converted at a broadcast studio into an electromagnetic disturbance which, whatever it may be, is certainly no longer sound. Captured by a receiving set, this disturbance is reconverted into sounds tolerably similar to those which had entered the microphone. In a closely analogous way, the heat of the sun, for example, is converted into radiant energy and transmitted into space. Some of it is intercepted by the earth and other astronomical bodies and absorbed. In the process of absorption it is reconverted into heat.

Thermal Conductivity

If a wooden object and a metal object are removed by hand from a hot oven, the metal object will feel much the hotter of the two. If the same two objects be picked up outdoors on a cold winter day, the metal object will this time seem to be the colder of the two. In both cases the metal could be demonstrated to be at the same temperature as the wood. There can be no question about the metal transferring heat the more rapidly to the hand in the first case, nor away from it in the second, but this is not due to any higher or lower temperature. Rather the impression of temperature is an illusion produced by the greater rate at which heat is transferred. All of this constitutes an additional reason for distrusting sensation as a means of evaluating temperature.

The proclivity of different materials for transferring heat forms the basis of a concept called thermal *conductivity.* The definition of thermal conductivity which is today generally accepted was first formulated by Joseph Fourier in 1822 in a work which has already been referred to in another connection (page 116) and which will later be seen to have inspired some of the early discoveries in current electricity (page 463). Fourier pointed out that the rate of conduction of heat would be proportional to the cross sectional area, a, of the conducting body, proportional to the difference of temperature, ΔT,[1] between the two regions delivering and receiving the heat, and inversely proportional to the distance, l, between these regions. It would also depend on the substance conducting the heat. Such dependence could be specified by a numerical coefficient, σ, termed the *thermal conductivity* in the following way.

Consider the number of calories, H, flowing in t seconds between two points having a difference of temperature ΔT. By the preceding paragraph

$$\frac{H}{t} = \sigma \frac{a\overline{\Delta T}}{l}. \tag{1}$$

Then, when a is one square meter, ΔT 1° C., and l one meter,

$$\frac{H}{t} = \sigma, \tag{2}$$

[1] The Greek capital *delta* (Δ) is frequently used to express the difference between two quantities, usually implying that the difference is small, though that is not an essential point in the present use of the symbol.

or, in words, *the thermal conductivity of any substance is the number of Calories which will flow each second between opposite faces of a cube of that substance one meter on an edge under a temperature difference of 1° C.*

Defined in this way, the numerical values of thermal conductivities of some of the metals are, in order of magnitude:

silver	.00100	tungsten	.00035
copper	.00092	iron	.00016
gold	.00070	mercury	.00002
aluminum	.00048		

It is worthy of note that this same table represents also the order and the relative values of *electrical* conductivities, a discovery of considerable theoretical significance first made in 1853.[1] (Compare page 465.)

The thermal conductivities of building materials are of practical importance, the values naturally being very much less than those of metals. They range from about 10^{-6} for brick through 10^{-7} for wood down to nearly 10^{-8} for felt.

The thermal conductivities of liquids and gases are much more difficult to measure than those of solids, for the disturbing effects of convection are hard to circumvent. Their values are all low. Water is among the highest of the liquids, being about $13 \cdot 10^{-7}$. Among the gases hydrogen and helium are the highest, each being about $3.3 \cdot 10^{-7}$, while air is $.57 \cdot 10^{-7}$. The kinetic theory of gases has utilized measured values of heat conductivity to yield information on the mean distance between molecules.

One of the classical early applications of Fourier's theory of heat, including his definition of thermal conductivity, was the first scientific estimate of the age of the earth. The temperature of the earth was known to increase about 1° C. for every hundred feet depth. Measurements of the conductivity of the igneous rock constituting the bulk of the earth then gave information on the rate at which heat was escaping from the center of the earth to outer space. It was then largely a matter of arithmetic to set the time at which the earth originally contained enough heat to cause its temperature to be that of the sun from which it had presumably been thrown off in the beginning. The age of the earth thus deduced from Fourier's theory[2] was not more than 200,000,000 years. This estimate is now considered much too low, for it is now known that the earth is cooling less rapidly than was originally supposed, there being a supply of heat from radioactive materials not known at the time.

The Discovery of Radiant Energy

The fact that radiation from the sun travels in some other guise than heat was mentioned on page 183. The fact that the paths followed by this radiation can be traced by tracing the accompanying light furnishes a

[1] Wiedemann and Franz, *Poggendorff's Annalen, 89,* 497 (1853).

[2] William Thomson, *Cambridge & Dublin Mathematical Journal* (1844).

temptation to identify heat radiation with light. In the common use of burning glasses over a period of many centuries no clear distinction was ever made until the seventeenth century between the light and the heat that were focused by the glasses. But in 1620, Francis Bacon showed the first clear recognition of the possibility that the two might be handled separately. He said (11:127):

> Let the burning glass be tried on warm objects which emit no luminous rays, as heated but not ignited [i.e., not incandescent] iron or stone or hot water, or the like; and observe whether the heat becomes increased and condensed, as happens with solar rays.

About sixty years later, Mariotte discovered another phenomenon which emphasized still further the distinction between light and heat radiation. Though the two remained together when sunlight was acted upon by a burning glass, they were separated by a burning glass when the source was a fire instead of the sun. In his experiment he put a concave metal mirror before a fire. At its focus the hand could not long endure the heat; but when a glass plate was placed over the mirror, heat could no longer be felt at the focus, though the light was substantially undiminished (112:*1*:303, 344).

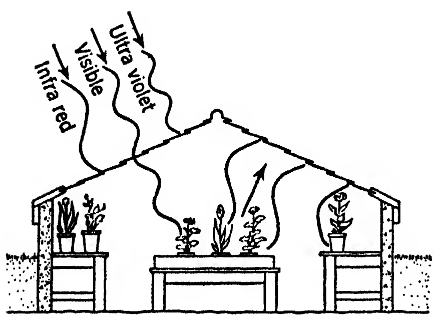


FIG. 154. GLASS ACTS AS A HEAT TRAP TO RADIATION FROM THE SUN

More than a hundred years elapsed before further steps were made toward an understanding of heat radiation. In 1777, Carl W. Scheele (1742–86) repeated and extended Mariotte's observations (112:120 ff.) not only distinguishing radiant energy from light on the one hand, as Bacon and Mariotte had done, but also distinguishing it from "fire" or ordinary heat on the other hand. He coined the name "radiant heat" and spoke of "the radiant heat which is invisible and differs from fire," that is to say, which is neither light nor ordinary heat.

Mariotte's observation furnishes the explanation for the action of the familiar "hothouse." Glass transmits radiation from the sun but is opaque to radiation from lower-temperature sources, such as stoves, and even more so to radiation from sources of still lower temperature, such as warm earth. Hence, a glass-covered box exposed to the sun acts as a "heat trap" and permits the enclosed earth to build up temperatures much above those of the surroundings. The same effect is oppressively evident inside an automobile that has stood for some time in the sun. For the same reason, a thermometer exposed to the sun will always indicate temperatures higher than that of its surroundings. The amount of the excess depends on the thermometer. It would be small if the bulb contained alcohol, since much of the radiation would simply travel through the transparent liquid without being absorbed. The excess would be greater for opaque mercury, though the silvery surface would still reflect much of the radiation. But if the bulb were painted a dead black, most of the radiation incident on it would be absorbed and the thermometer would register a temperature far higher than that of its surroundings. Hence one may see that any statement about temperature in the sunshine is largely meaningless, since every object exposed to the sun assumes a different temperature.

Subsequent studies have made it clear that radiation is much more analogous to light than to heat. The relation is, indeed, much closer than that of mere analogy. Because of this fact, it will be of advantage to defer a consideration of some of the attributes of radiation until the characteristics of light have been studied (pages 393 ff.). Others lend themselves to treatment without the concepts furnished by a study of light.

Prévost's Theory of Exchanges

As early as the sixteenth century, Giambattista della Porta, whose contributions to light at the tender age of fifteen will presently be considered, had remarked (98:264) that a concave mirror reflected sound, light, heat, *and cold*. Noting that the light of a candle set before a mirror and its heat each produced its appropriate sensation on the eye placed at the conjugate focus, he remarked

> but this is more wonderful, that, as heat, so cold should be reflected: if you put snow in that place if it come to the eye... it (the eye) will presently feel the cold.

A similar effect was identified by subsequent observers. The observation lent some weight for a time to a theory that there was a fluid carrying the attributes of cold just as there had been assumed to be the fluid called "caloric." But that theory was short-lived. It did, however, produce more substantial fruit in what has become known as Prévost's *theory of exchanges* of 1791. The correct interpretation of Porta's observation was not that the ice radiated "cold" to the eye but that the eye, being the warmer of the two, radiated heat to the ice and felt cool in consequence.

Prévost, of Geneva, asked himself a hundred and fifty years ago whether the warm eye ceased radiating whenever a warmer body was substituted for the ice. (104.) He concluded that it did not; that, on the contrary, every body radiated heat no matter what its temperature was; but that warmer bodies radiated more rapidly than cooler and that, hence, the net flow was from the warmer to cooler bodies. He considered that the ultimate state of equal temperature represented, not the cessation of radiation, but merely a state of dynamic equilibrium in which the bodies radiated and absorbed heat at equal rates.

Laws of Cooling

Having observed that rates of cooling of warm bodies were greater the higher their temperatures above the surroundings, Newton made a statement in 1701 involving the hypothesis that rate of cooling is actually proportional to excess of temperature over that of the surroundings. This is commonly called *Newton's law of cooling.* It was never intended as a law of radiation, though it is sometimes classified as such and criticized in consequence for its inaccuracy. It was merely an attempt to describe, on a purely empirical basis, the rate of cooling of hot specimens ordinarily encountered. While this cooling is due in part to radiation, conduction to the surrounding air also makes a contribution. This contribution is, of course, influenced by the convection currents thereby set up. Within its province Newton's law of cooling is a useful approximation applying to small differences of temperature, but it is nothing more.

Almost two centuries after Newton's time (1879) the real law of cooling by radiation was discovered by Joseph Stefan (1835–93). It is usually termed *Stefan's fourth power law.* He said (77:378):

> We obtain numbers which come very close to the [observed] rates of cooling if we assume that the heat radiated by a body is proportional to the fourth power of its absolute temperature.

Stefan was thinking of radiation in the same sense as Prévost had been thinking of it nearly a hundred years earlier. Thus the net radiation between a heated object and its surroundings would be proportional to the difference of the fourth powers of the respective absolute temperatures. Compare, for example, the radiation from a piece of metal at the temperature of boiling water (100° C.) with its radiation when heated to the point where it barely glows in the dark, say 500° C., the surroundings being at room temperature, say 20° C. The ratio would be

$$\frac{(273+500)^4-(273+20)^4}{(273+100)^4-(273+20)^4}=\frac{(35.7-.74)\cdot 10^{10}}{(1.93-.74)\cdot 10^{10}}=29.4.$$

Thus a piece of iron barely red hot experiences a net loss of heat by radiation about thirty times as fast as it does when at the temperature of boiling water. Though it is quite unfair to Newton's law to invoke it in such a case, the corresponding calculation is $\frac{500-20}{100-20}=6$, a notable discrepancy. It is interesting to observe, however, that if the two temperatures are taken as 22° C. and 21° C., the room temperature being still 20° C., the two results agree within a fraction of one per cent.

Questions for Self-Examination

1. Sketch the development of ideas about the nature of heat.
2. Explain the concept of specific heat.
3. Define the Calorie.
4. Define thermal conductivity and give approximate values for representative materials.
5. How was the radiant energy ("heat radiation") first identified?
6. Comment on the apparent radiation of cold and its connection with Prévost's theory of exchanges.
7. Compare Newton's law of cooling with Stefan's fourth power law and give a numerical example of your own devising.

Problems on Chapter 17

1. The product of the weight of an object by the specific heat of its material is sometimes termed its *water equivalent.* Why is the term appropriate?
2. A 50-gram slug of gold of specific heat .032 is used to measure the temperature of a small furnace. Taken from the furnace and dropped into 50 grams of water initially at 0° C., it raises the water to 29° C. What was the temperature of the furnace? 940° C.
3. The specific heat of glass is .17 and its specific gravity is 2.6. The corresponding figures for mercury are .033 and 13.6. Show that when a thermometer of given volume is dipped into a liquid, the heat absorbed by it is almost independent of the proportion of glass and mercury making up the thermometer.
4. An iron boiler is made of plate 1 centimeter thick. Its exposed surface is 8 square meters. The water inside is at a temperature of 100° C. The exposed surface is at 80° C. How much heat is lost every hour by conduction? 9200 Calories.
5. The heat loss from a certain cottage occurs almost entirely through the sidewalls, of area 1000 square meters. The walls are 30 centimeters thick, made of masonry of conductivity 10^{-6}. How much heat is lost per hour when the inside temperature is 20° C. and the outside −20° C.? 480 Calories.
6. A copper vessel of mass m kilograms contains W kilograms of water at t° C. M kilograms of copper at a temperature T° C. is dropped into the water. The final temperature is τ° C. Find the specific heat, s, of copper.

m	W	M	t	T	τ	s
.3	.4	.5	10	100	19.	.095
.3	.45	.5	10	100	18.2	.096
.3	.5	.5	10	100	17.5	.096
.3	.55	.5	10	100	16.8	.095

7. In a bomb calorimeter containing W kilograms of water m grams of coal are burned. The bomb is of steel and weighs M kilograms. The observed rise in temperature is t° C. Find the heating value H of the coal in calories per gram. Take the specific heat of steel as .133.

W	m	M	t	H
1.5	2	1	10	8100
1.75	2	1	9	8400
2.	2	1	8	8500
2.25	2	1	7	8300

8. A "cooling calorimeter" consists of a copper container weighing m grams. When it is filled with W grams of water, T seconds is required to cool from the higher to

the lower of two observed temperatures. When it is filled with M grams of a liquid of unknown specific heat, t seconds is required between the same two temperatures. What is the unknown specific heat, s? Take the specific heat of copper as .093.

m	W	M	T	t	s
100	300	225	900	341	.480
100	300	250	900	370	.471
100	300	275	900	405	.472
100	300	300	900	438	.471

9. The thermal conductivity of aluminum is .00048. How many Calories of heat H will pass in an hour through an aluminum teakettle containing boiling water? The bottom of the teakettle is d centimeters thick and the area is a square centimeters. It is set on a gas flame which maintains the outside at a temperature $T°$ C.

d	a	T	H
.3	100	125	140
.3	100	115	87
.3	100	110	58
.3	100	105	29

10. A wall is composed of plaster and brick, the plaster (conductivity p) P meters thick and the brick (conductivity b) B meters thick. If the inside temperature is $T°$ C. and the outside $t°$ C., what is the temperature, τ, at the interface and how many calories, H, escape through each square meter of wall in the course of a day? Use the principle that the quantity of heat which escapes through the entire thickness must pass through each layer in turn.

P	p	B	b	T	t	τ	H
.02	$7 \cdot 10^{-7}$	.10	$15 \cdot 10^{-7}$	20	− 20	8	36
.02	$7 \cdot 10^{-7}$	.10	$15 \cdot 10^{-7}$	20	0	14	18
.01	$7 \cdot 10^{-7}$	.10	$15 \cdot 10^{-7}$	20	− 20	13	43
.01	$7 \cdot 10^{-7}$	.10	$15 \cdot 10^{-7}$	20	0	16	21

CHAPTER 18

Change of State

The Heat of Fusion

Black's work was not confined to setting up the concept of specific heat and laying the foundation for the definition of a heat unit. These marked the beginning of the science of heat by making the first and all-important distinction between temperature and quantity; but more was to come. Armed with the new concept of quantity of heat, Black naturally turned next to the heat relations involved in transformations between solid, liquid, and gaseous states of aggregation.

These needed clarification. Black pointed out that even a casual consideration of the prevailing ideas on melting and freezing, for example, showed that they must be erroneous. He remarked (21:116 ff.):

> Fluidity was universally considered as produced by a *small* addition to the quantity of heat which a body contains when it is once heated up to the melting point.... If this common opinion had been well founded, if the complete change (of ice or snow) into water required only the further addition of a very small quantity of heat, the mass, though of considerable size, ought all to be melted, in a very few minutes or seconds, the heat continuing incessantly to be communicated from the air around....
>
> Were this really the case, the consequences of it would be dreadful in many cases; for, even as things are at present, the melting of great quantities of snow and ice occasions violent torrents, and great inundations in the cold countries, or in the rivers that come from them. But were the ice and snow to melt as suddenly as they must necessarily do, were the former opinion of the action of heat in melting them well founded, the torrents and inundations would be incomparably more irresistible and dreadful. They would tear up and sweep away everything, and that so suddenly that mankind should have great difficulty to escape from their ravages.
>
> This sudden liquefaction does not actually happen. The masses of ice or snow, after they begin to melt, often require many weeks of warm weather, before they are totally dissolved into water....
>
> *A great quantity*, therefore, of the matter of heat which enters into the melting ice, produces no other effect but to give it fluidity, without augmenting its sensible heat; it appears to be absorbed and concealed within the water, so as not to be discoverable by the application of a thermometer.
>
> In order to understand this absorption of heat into the melting ice and concealment of it in the water more distinctly, I made (among others) the

> following experiment.... I put a lump of ice into an equal quantity of water heated to the temperature 80° C.[1] and the result was that when the ice was all melted the fluid was no hotter than water just ready to freeze....
>
> I shall now mention another example, an experiment first made by Fahrenheit [*Philosophical Transactions, 33*, 78 (1724)]. He... exposed globes of water in frosty weather so long that he had reason to be satisfied that they were cooled down to the degree of the air, which was four or five degrees below the freezing point. The water, however, still remained fluid, so long as the glasses were left undisturbed, but, on being taken up and shaken a little a sudden freezing of a part of the water was instantly seen.... But the most remarkable fact is, that while this happens the mixture of ice and water suddenly becomes warmer, and makes a thermometer, immersed in it, rise to the freezing point.

Only in a field in which scientific development was long overdue could almost casual observations have been so devastating to previous views. What would seem to be the inescapable implication in the length of time required for ice to melt even in warm weather had been lost on previous observers. It would seem as though anybody might have observed that a chunk of ice would cool an equal weight of water through 80° and have drawn a correct deduction therefrom. And even as acute an experimenter as Fahrenheit had failed to grasp the significance of his sub-cooling and freezing experiment of forty years before. These observations pointed unequivocally toward the absorption of a huge quantity of heat in the process of melting unaccompanied by any rise in temperature, and the evolution of a corresponding quantity in freezing.

Black coined the term "latent heat" to describe the heat which seemed thus to become concealed when fusion occurred and mysteriously to reappear when freezing set in. The term, while still in use, is giving way to the more descriptive *heat of fusion*. One may conclude correctly from Black's observations that the heat of fusion of ice is 80 Calories per kilogram.

Farmers use the heat of fusion of ice to protect stored fruit from freezing. They provide large tubs of water, the freezing of which at a temperature slightly above that which would damage the fruits evolves heat which prevents the room from becoming colder as long as any considerable amount of water remains unfrozen.

The Influence of Pressure on the Freezing Point

Most substances occupy a larger volume in the liquid than in the solid state, the process of solidification being accompanied by contraction. Water is a notable exception. It is for this reason that ice floats. Its specific gravity is .917. Thus about 8 per cent of the total volume of floating ice projects above water.

The temperature at which freezing occurs, termed the *freezing point*, is

[1] Black's thermometers were graduated in Fahrenheit degrees, and his other measurements were in the English system. All observations are here converted to centigrade degrees and the metric system.

ordinarily stated to be 0° C. The *melting point*, analogously defined, is the same temperature. But both, while identical, may deviate from the conventional 0° C. It was not until the middle of the nineteenth century (63·*1*:156) that the discovery was made that the freezing or melting point was lowered by the application of pressure, the amount of such lowering being .0075° C. for each additional atmosphere of pressure. This is another discovery that it seems rather curious was not made earlier, for the making of snowballs is dependent on the lowering of melting point by pressure. Moreover, skating, while not dependent upon it, is considerably facilitated by it. The flow of glaciers also depends in part on this.

The reason for this phenomenon is not far to seek. Consider a mixture of ice and water at 0° C. Any further melting will decrease the volume, and, conversely, anything acting to decrease the volume will stimulate melting. Now, apply pressure. The natural decrease-of-volume response will be accompanied by the melting of a part of the ice. The remainder of the mixture will grow colder through yielding the heat of fusion absorbed by the melting ice. Melting will cease with the resulting establishment of a new state of temperature equilibrium; that is, the melting point will have become lower. One of the implications of this explanation which has been borne out by observation is that, in the case of solids which expand on melting, the application of pressure will raise the melting point and vice versa.

The Heat of Vaporization

Transformations between the liquid and solid states are not the only changes of state that can occur. Even more common, in fact, are transformations between the liquid and the vapor states. Of the heat relations involved in this type of occurrence, Black was also the discoverer. He said (21:154):

> I can easily shew, in the same manner as in the case of melting, that a very great quantity of heat is necessary to the production of vapour.... This great quantity of heat enters into the vapour gradually, while it is forming, without making it perceptibly hotter to the thermometer.... On the other hand, that when the vapour of water is condensed into a liquid, the very same great quantity of heat comes out of it into the colder matter by which it is condensed.

In support of his last assertion, Black invoked the common knowledge possessed by his hearers in consequence of their presumed familiarity with a "favorite" Scotch pursuit. He said (21:166):

> All of you know well enough how the operation of common distillation is conducted...

He pointed out as evidence of the heat yielded through the condensation of a vapor into a liquid, the heating of the steady stream of cold water which flowed over the condenser of a still in operation. He might have utilized measurements of this heating effect to secure information on the "latent heat" of condensation, but there is no record of his having done it in just that way. Instead he made a somewhat indirect measurement of the heat consumed in vaporizing water, now called the ***heat of vaporization***. It gave the value 450 Calories per kilogram. The value is now known to be 539.5 Calories per kilogram. Black's value was very creditable, considering the pioneer character of his venture.

The Influence of Pressure on Boiling Point

The temperature at which boiling sets in, termed naturally the ***boiling point***, depends upon pressure, as does the freezing point, and for a quite analogous reason. The volume of a liquid increases upon vaporization, and since pressure inhibits increase of volume, increased pressure will put an impediment in the way of vaporization, that is, raise the boiling point. From 100° C. at one atmosphere of pressure, the boiling point of water rises to 120° C. for 2 atmospheres, 152° C. for 5, 181° C. for 10, 214° C. for 20, and so forth (see Fig. 155). Below normal atmospheric pressure, the diminution of the boiling point is even more rapid. This diminution is one of the accompaniments of any increase in altitude such as in mountain climbing or an airplane flight. At an altitude of 1000 meters above sea level where the barometric pressure is .91 of the sea-level normal, the boiling point of water is 97° C. instead of 100. At 5000 meters, where the pressure is .57 of the normal, the boiling point is 87° C., and at 10,000 meters where the pressure is .32 of the normal, the boiling point is 71° C. (see Fig. 156). Under such circumstances, the boiling of meats and vegetables must be done in pressure-cookers.

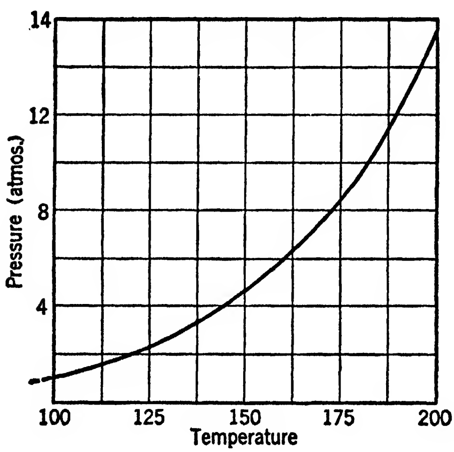


FIG. 155. PRESSURE OF WATER VAPOR ABOVE ONE ATMOSPHERE AS A FUNCTION OF TEMPERATURE

Since the boiling point of water lowers and the freezing point rises as pressure is diminished, one is led to inquire whether the two may become identical at some value of the pressure. If that condition should occur, water could be seen boiling in the presence of ice. Such a state of affairs is, in fact, possible and has been realized. It occurs at a pressure of about .006 atmosphere (4.6 mm. Hg.), and a temperature of .0075° C. It is quite appropriately termed the ***triple point*** (see Fig. 157).

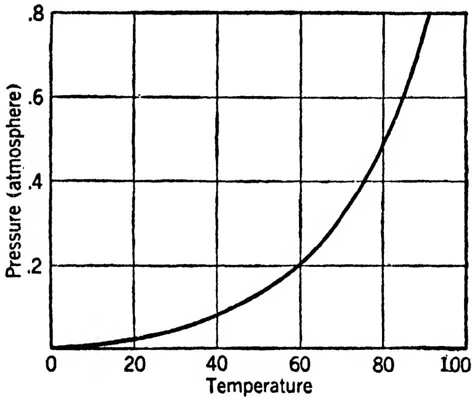


FIG. 156. PRESSURES OF WATER VAPOR BELOW ONE ATMOSPHERE AS A FUNCTION OF TEMPERATURE

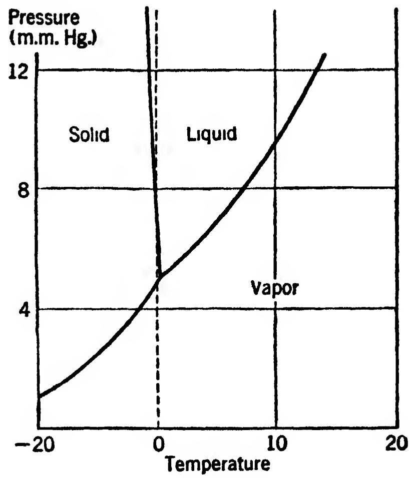


FIG. 157. THE TRIPLE POINT OF WATER

Sublimation

Though the triple-point pressure for water is a small fraction of an atmosphere, this is not true for all substances. One exception is the triple point of carbon dioxide which occurs far above atmospheric pressure. Consequently this substance cannot exist in liquid form at atmospheric pressure. Solid carbon dioxide, familiar under the name of Dry Ice, is so called because in evaporating at atmospheric pressure it passes directly from the solid to the vaporous state, a process known as *sublimation*. Several common substances, such as camphor and naphthalene ("moth balls"), act in the same way. The converse, the passage directly from the vaporous into the solid state, is well illustrated in the formation of frost and snow.

Corresponding to the existence of definite temperatures at which melting and vaporization occur, there seem to be corresponding temperatures at which sublimation occurs, the solid remaining at that temperature as heat is supplied until vaporization is complete. Also, as would be expected, the heat of sublimation is approximately equal to the sum of the heats of fusion and of vaporization.

The Caloric Theory of Change of State

The phenomena of change of state created a problem for the calorists, but under Black's leadership they rose to the occasion. The problem was, How can large quantities of heat be poured into a substance while it is melting or vaporizing, without increasing its temperature? Black had called this heat "latent," which was simply a way of saying that it did not raise the temperature. The adjective constituted no explanation, however. Black concocted an explanation which consisted in attributing the condition to a heat-absorbing type of chemical reaction, a phenomenon well known to chemists of today. He said, referring to vaporization (21:161):

> The heat, therefore, did not escape *along with* the vapour, but *in* it, probably united to every particle, as one of the ingredients of its vaporous constitution. And as ice, united with a certain quantity of heat, is water, so water, united with another quantity of heat, is steam or vapour.

The Kinetic Theory of Change of State

Nearly three hundred years ago, Robert Hooke asked (87:97), What is the cause of fluidness? and answered his own question thus:

> This I conceive to be nothing else but a certain pulse or shake of heat; for heat being nothing else but a very brisk and vehement agitation of the parts of a body, . . . [they] are thereby made so loose from one another that they easily move away and become fluid.

This picture lent itself readily to an extension covering vaporization. Indeed the kinetic concept of gas pressure had become very specific at the hands of Daniel Bernoulli in 1738, as has already been pointed out (page 140). Not only was gas pressure explainable in terms of molecular impacts, increasing in magnitude both with added compression and with increasing temperature (molecular speeds), but it was particularly easy to visualize evaporation as consisting of the breaking away of occasional molecules from the surface of the liquid (Fig. 158). If the container was open, these molecular escapes were permanent. But if the container was closed, the escaped molecules, being imprisoned in the space above the liquid, accumulated in number, and an occasional molecule would be recaptured upon impact with the liquid surface. As the density of the vapor increased through accumulation of the escaped molecules, the frequency of these recaptures became greater. Presently the number of recaptures became equal to the number of escapes, and thenceforth the molecular concentration above the liquid, that is to say, the vapor density, remained unchanged. It followed that the vapor pressure also remained unchanged.

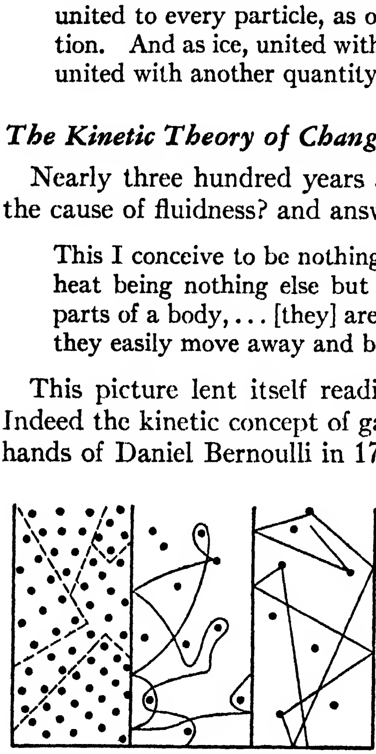


FIG. 158. SOLID, LIQUID, AND VAPOR AS STATES OF MOLECULAR AGGREGATION

It is to this condition that characteristic vapor pressures are today attributed. The vapor pressure above water in a closed container, for example, is always the same at a given temperature (about 1.75 centimeters of mercury at usual room temperature). Oils have lower vapor pressures, and volatile liquids, such as alcohol and ether, have higher pressures (4.45 and 43.3 centimeters respectively). The latter are associated with the fact that at room temperature the two liquids are much nearer their boiling points than is water. In fact, the boiling point is sometimes defined as that temperature at which the vapor pressure of the liquid becomes equal to the pressure upon its surface.

But the vapor pressure of a liquid is known to increase with rising temperature. This is clearly evident for water in Figures 155 and 156. This too is explicable on the kinetic basis. Assume, for example, that a state of equilibrium has been established for a particular temperature. If, now, the temperature rises, the rate of escape of molecules from the surface will increase, and the concentration above the surface will increase in consequence, up to the point where the new and larger number of recaptures becomes again equal to the number of escapes. A higher vapor density and vapor pressure is thus established, characteristic of the new temperature.

The Liquefaction of Gases

These concepts shed some light on the next stage in the development of knowledge about changes of state, the liquefaction of gases. In 1823, Michael Faraday, whose principal contributions were to the field of electricity, half by accident liquefied chlorine, which until then had been classified unqualifiedly as a gas. Through the incident he discovered that the circumstance that seemed to favor the liquefaction of gases was high pressure accompanied by cooling. He forthwith applied the same treatment to other gases and was successful with some of them. With the greater power and more effective refrigeration available in later years, the number of gases which could be liquefied was gradually extended until in 1908 the most refractory of all the gases, helium, was liquefied at a temperature 4.3° above the absolute zero. The progress of this work eliminated the supposed line of division between "permanent gases" and gases which were liquefiable. This classification had sprung into being in 1845, when at the conclusion of Faraday's extended work, the gases which he had not succeeded in liquefying had received the appellation "permanent gases."

The Discovery of the Critical Point

In 1863, a significant discovery was made by Thomas Andrews (1813–85) at Belfast, Ireland. He wrote (quoted 25:201):

> On partially liquefying carbonic acid (CO_2) by pressure alone and gradually raising at the same time the temperature to 31° C., the surface of demarcation between the liquid and gas became fainter, lost its curvature and at last disappeared.... At temperatures above 31° C., no apparent liquefaction of carbonic acid, or separation into distinct forms of matter, could be effected, even when a pressure of 300 or 400 atmospheres was applied.... The properties described in this communication, as exhibited by carbonic

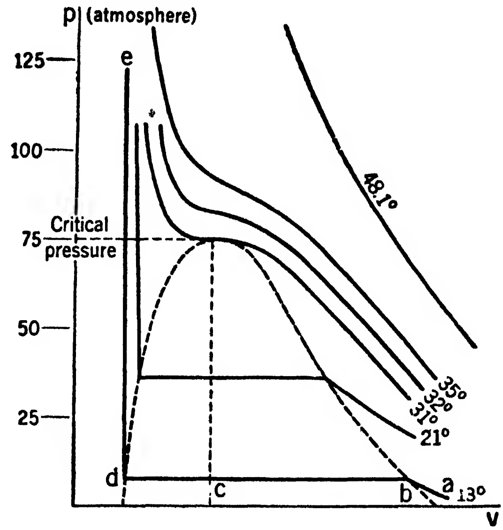


FIG. 159. ANDREWS' ISOTHERMALS ABOVE AND BELOW THE CRITICAL POINTS

acid, are not peculiar to it but are generally true of all bodies which can be obtained as gases and liquids.

Andrews gave the name *critical temperature* to the 31° point of carbon dioxide and the analogous temperatures for other substances. Below that temperature the substance could exist as part liquid, part vapor. The pressure necessary to liquefy the gas when at the critical temperature was termed the *critical pressure.* At lower temperatures, less pressure was required. At higher temperatures no amount of pressure, however great, would serve to produce liquefaction.

The circumstances can perhaps best be visualized by reference to one of Andrews' diagrams. Figure 159 represents a pressure-volume diagram, similar to the one used in the discussion of Boyle's law (Figure 88). Here, however, at the lower temperatures the behavior of the gas departs from that shown in Figure 88. The upper isothermal — as the pressure-volume line for any given temperature is called — is substantially a Boyle's-law line, which indicates that 48° C. is far enough above the critical temperature so that carbon dioxide measurably obeys Boyle's law. But deviations occur at lower temperatures, suggesting a liquefaction not yet attained. The 31° isothermal represents the bare attainment of liquefaction, and the horizontal portions of the remaining isothermals show the processes of liquefaction for progressively lower temperatures.

These horizontal portions of the curves are especially informing. The 13° curve, for example, indicates that when carbon dioxide at that temperature has been compressed to about ten atmospheres pressure (point *b*) liquefaction begins; that the pressure remains unchanged (line *bcd*) as the volume diminishes until, when all the vapor is condensed (point *d*), any further diminution of volume of the essentially incompressible liquid can be obtained only by the exertion of enormous pressure (line *de*). The sequence may be represented as in Figure 160.

The constancy of the pressure during the period of condensation, even though the volume is being sharply diminished during that interval, is quite understandable in terms of the kinetic concepts already used. If the volume of the saturated vapor above a liquid is decreased, the concentration of vapor molecules increases and, therefore, momentarily accelerates the recapture of such molecules by the liquid. The original density (and accompanying pressure) is almost instantly restored through this temporary excess of molecular recaptures over escapes. The process repeats itself upon any further diminution of volume, thus maintaining the pressure constant until the volume of the vapor has diminished to zero. It is evident that the process would be reversed if the volume of vapor were undergoing an increase.

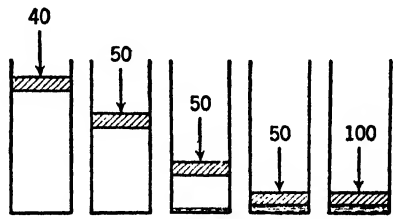


FIG. 160. LIQUEFACTION OF CARBON DIOXIDE BY PRESSURE AT 13° C.

Questions for Self-Examination

1. Contrast the facts about heat relations accompanying changes of state with early fancies. Tell what considerations led Black to investigate the subject and how he did it.
2. How did Black first measure the heat of fusion of ice and the heat of vaporization of water?
3. What is the effect of pressure on freezing and boiling points and why?
4. With a $P - T$ diagram describe the properties of the "triple point."
5. What is the difference between a vapor and a gas in its response to diminution of volume and why?
6. Sketch and interpret a set of typical $P - V$ isothermal curves such as Andrews discovered for carbon dioxide.

Problems on Chapter 18

1. If a fine wire with weights at its ends is hung over a block of ice, the wire will cut through the block yet the ice is solid behind it. A copper wire will cut through the block more rapidly than an iron wire in this way. Explain this.
2. A lump of iron of mass .75 kilogram at a temperature of 400° C. is laid onto a piece of ice. It melts 420 grams of the ice in cooling to 0° C. What is the specific heat of the iron? .11 Cals. per kg.

3. A quantity of water is sub-cooled gradually down to −10° C. before freezing begins. Freezing then proceeds rapidly without the extraction of any more heat from the water, the temperature rising in the meantime to 0° C. At the end of the process, what proportion of the water is frozen? The specific heat of ice is .5. .12

4. Suppose the ground to be covered with snow at 0° C. to an average depth of 10 centimeters, the density of the snow being .2 kilogram per liter. How much rainfall at 10° C. will be required to melt it? 16 cms.

5. Water initially at 4° C. is confined in a cast iron vessel and exposed to low temperature. Ultimately the vessel bursts. Assuming that water is substantially incompressible, would you judge that freezing occurs before or after the bursting?

6. An "artificial" ice rink is resurfaced by turning steam into pipes embedded in the ice for the purpose. If M metric tons of ice are to be melted, how many tons of steam, s, will be required, the temperature of the water being t degrees after the melting is complete? Take heat of vaporization as 539 and of fusion as 80 Calories per kilogram.

M	t	s
10	0	1.3
10	10	1.4
10	20	1.6
10	30	1.8

7. Water, if caused to evaporate with sufficient rapidity, may be frozen by the loss of its own heat of vaporization. A mass of water, originally M grams, at $T°$ C. is so treated. How many grams m will remain when freezing is just complete, assuming no radiation or conduction of heat?

M	T	m
1000	100	720
1000	50	790
1000	25	830
1000	0	870

CHAPTER 19

Heat and Mechanical Energy

Early Steam Engines and the Mechanical Equivalent of Heat

The next step toward a comprehension of the nature of heat was in part an outgrowth of the engineering developments of the eighteenth century. It constitutes an illustration of the continual interplay between the physical sciences and technology. Technology consists primarily in the adaptation of scientific principles to utilitarian ends. Its motivation is filling the requirements of daily life in a machine age, and its working materials are the discovered laws of science. Without the basic sciences it could not exist. On the other hand, the sciences frequently find in technology fertile fields for exploration. This is because technology incorporates much of the unanalyzed accumulated experience of the race along with the scientific material which it utilizes. The unrecognized mysteries of the commonplace constitute an inexhaustible store of scientific problems. One needs only to consider Archimedes and the balance, Galileo and falling bodies, Newton and the apple, Pascal and the water level, and Black and the flow of heat, to appreciate this fact.

In 1763, James Watt (1736–1819), scientific-instrument maker in the University of Glasgow, was asked to repair a model of a type of "fire engine," as steam engines were called in those days, which had been in use for about fifty years. Several hundred of them were then pumping water from deep mines in England and on the European continent. It was a very crude affair indeed, and apparently Watt considered it so, for he set about devising the improvements which have made his name immortal.

But the important thing was, not that the engine was so crude, nor even that it started Watt in the construction of a better one, but that it was a practicable device for converting heat into mechanical energy and utilizing it as a source of power. The discovery that heat could be so converted possessed enormous potentialities, not merely in unlocking the sources of power that were to create the Industrial Revolution, but also in giving emphasis to a phenomenon which was destined to shed much light on the nature of heat itself.

The reverse discovery, that mechanical energy could be converted into heat, brings into the scene one Benjamin Thompson (1753–1814). Thompson was a Massachusetts Yankee who became an officer in the British army during the American Revolution and later was made a Bavarian Count, thereafter going under the name "Count Rumford." Rumford was the name of the estate which he owned earlier near Concord, New Hampshire.

While supervising the boring of brass cannon for the Bavarian army, Rumford's attention was arrested by the amount of heat generated by friction of the boring tool with the gun. It seemed so disproportionately large to come out of the objects evolving it that Rumford's curiosity was piqued as to the source of such large amounts of heat. The caloric theory attributed heat produced in this way to the squeezing out of caloric from the brass by the action of the boring machinery. This explanation apparently did not appeal to Rumford. It occurred to him that if this were correct, there should be less caloric in the shavings than in the same weight of solid brass, because so much would have evolved in the process of converting the brass to shavings. Rumford decided to find out whether the shavings gave out less heat in the process of cooling than did the same weight of solid brass. That is, he compared the specific heats of the brass in the two states, but could find no difference between them.

The discovery that the frictional heat had not been abraded out of the metal left unanswered the original question as to where it did originate. Rumford's suspicion fell upon the horse which, by a treadmill, was furnishing the power for the boring operations. He had noticed that the quantity of heat developed in the process of boring was not at all proportional to the amount of metal removed, but that it seemed rather to be proportional to the work done by the horse, whether much or little metal was removed by the process. So he set about the measurement of the constant of proportionality, specifically the number of joules of work necessary to heat one kilogram of water 1° C.[1] This ratio between work performed and the heat generated thereby has come to be called the *mechanical equivalent of heat.* Rumford deduced a value that was too high by about one-third, the value accepted at present being 4183 joules/Calorie.

The Conservation of Energy

It would probably be correct to say that Rumford gave his energies to preliminary experimental work in a field the principal significance of which eluded him. When the field next began to assume importance, forty-four years later, the situation was exactly reversed. The man who reopened the question in 1842 was, in contrast, enormously impressed with its significance; but his approach was characterized by an almost complete lack of experimentation. He was an obscure German physician, J. R. Mayer (1814–78). His great discovery, or perhaps, in view of his lack of experimental foundations we should say his great "hunch," was what has since taken the form of the scientific doctrine of the *Conservation of Energy.* This

[1] Rumford's measurements were in the English system of units. The corresponding metric units are being used here.

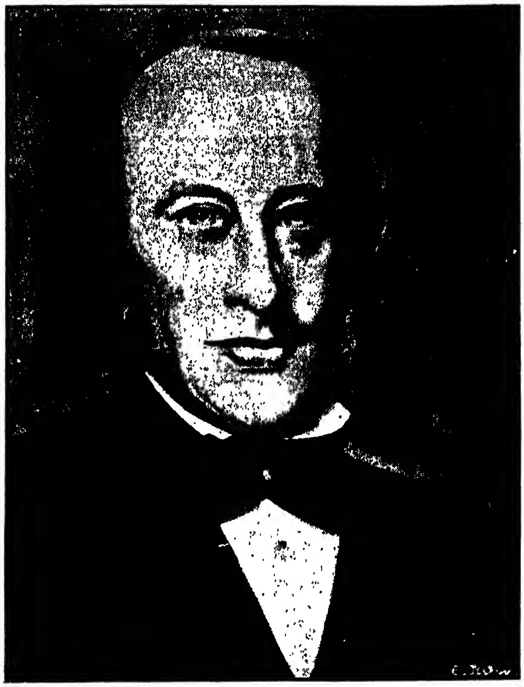


FIG. 161. JAMES P. JOULE (1818–1889)

was a concept toward which the scientific mind had been groping for two hundred years. In his 1842 paper he stated that

> force [the current term for energy] once in existence cannot be annihilated; it can only change its form.

In this and in a succeeding paper in 1845 he expanded this idea to cover what was, especially for the time, an amazing variety of phenomena. He began with inorganic manifestations of interconvertibility of work and heat, including chemical reactions, and extended it to the brand-new idea that plants and coal deposits when burned were merely yielding up the heat previously received from the sun. He included animal life — which took in man — in his idea that heat and energy output are to be equated to intake, ultimately of energy from the sun. He extended the principle to astronomical phenomena, computing the speed of fall to the earth from an infinite distance as 34,450 feet per second, and proposed the theory that the heat of the sun was due to the impacts of the myriads of meteorites known to be falling into it daily. Both of these ideas were entirely new at the time.

Joule Determines the Mechanical Equivalent of Heat

James P. Joule (1818–89) was a brewer of Manchester, England. Like many other men of independent means in the history of science, he became a scientist by avocation. He first became interested in "electric engines," the electric motor being in somewhat of a prenatal stage at that time. Before reaching his majority he had published several papers including, among other things, some measurements of the power developed by his rudimentary motors. In 1840 he presented a paper to the Royal Society on the production of heat by the electric current. It seems natural that, having measured both the power and the rate of evolution of heat manifested by an electric current, he should be led to compare the two. This he did, in 1842, the year of Mayer's original paper. He deduced the value 4510 joules per Calorie [1] for the mechanical equivalent of heat. That was his beginning. From that time on he devoted his life to redetermining, in every way that lent itself to experimental attack, the value of this constant. In 1843 he observed the rise in temperature due to the friction of water with the walls of capillary tubes through which it was driven, and deduced the value 4145. In 1845 he measured the heat developed or absorbed when air was compressed or expanded to the accompaniment of a measured amount of work (4300). In the same year he observed the rise in temperature of water churned by a paddle wheel whose work was measured (4800). Dissatisfied with this he repeated the experiment in water and in oil in 1847 (4200 and 4210). In 1849 he repeated the same measurements (4150). In 1850 he agitated mercury (4160). In the same year he utilized friction in cast iron (4165). In 1867 he heated water electrically instead of by a paddle wheel (4210). In 1878 he repeated his paddle wheel experiment and deduced his last value, 4154.

Joule concluded one of his papers with the following statement:

> I will therefore conclude by considering it as demonstrated by the experiments contained in this paper — (1) that the quantity of heat produced by the friction of bodies, whether solid or liquid is always proportional to the quantity of force [i.e., work] expended; and (2) that the quantity of heat capable of increasing the temperature of a kilogram of water (weighed in vacuo, and taken at between 13° and 16°) by 1° C., requires for its evolution the expenditure of a mechanical force represented by the fall of 424 kilograms through the space of 1 meter.

Ratio of Specific Heats of Gases

When any material is heated, the resulting expansion against surrounding pressure, atmospheric or otherwise, involves the performance of work. The heat equivalent of this work, in the ratio determined by Joule, is abstracted from the object being heated. Consequently more heat is required to produce a given change in temperature of an expanding object than would be required if it were not allowed to expand. This additional heat is so small for solids and liquids that it is ordinarily disregarded. But with gases it is considerable and results in variable specific heats for a given gas depending on the accompanying degree of expansion. Two special cases are impor-

[1] Joule's results were stated in the English system of units. The corresponding metric units are being used here, though they were not adopted until after Joule's time.

tant; one when the gas is heated in a rigid container and not allowed to expand at all, the other when it is allowed to expand at such a rate as to keep its pressure constant. Measurements in the former case yield the ***specific heat at constant volume***, in the latter the ***specific heat at constant pressure***. The latter possesses, of course, the higher value.

The ratio of these two specific heats, usually represented by γ (gamma), is a constant for a given gas, but is different for different gases. It is about 1.66 for monatomic gases like oxygen and nitrogen, 1.41 for diatomic gases like oxygen and nitrogen, 1.33 for triatomic gases like carbon dioxide, and still smaller for polyatomic gases and vapors.

The Conservation Concept

Joule, like Mayer, was animated by a conviction of the conservation of energy. Though he did not particularize on this theory as Mayer did, his conception of the inclusiveness of the doctrine was no less wide. In his 1843 paper he said (77:205):

> I shall lose no time in repeating and extending these experiments, being satisfied that the grand agents of nature are, by the Creator's fiat, *indestructible;* and that wherever mechanical force is expended, an exact equivalent of heat is *always* obtained.

The italics are Joule's own. This statement is no less inclusive than Mayer's. Indeed because he did not particularize as Mayer did, it may be said to be broader. A comparison of the inclusiveness of this statement with the very limited number of ways that the principle lent itself to verification at his hands will make it clear that his experiments constituted far from conclusive evidence on the validity of his belief. Yet because he performed experiments, even experiments which had little to do with some of the aspects of his doctrine, his doctrine was accepted in its entirety. Acceptance of Mayer's doctrine came only when it was seen to be identical with Joule's. Actually, the volume of qualitative evidence covering a wide variety of phenomena which Mayer proffered constituted better support of the broad implications of the doctrine of conservation of energy than did the close numerical agreement of Joule's results in a limited field.

It is perhaps unnecessary to emphasize the new "turn" that the doctrine of Mayer and of Joule gave to the old principle of conservation of energy. The old form was restricted to purely mechanical transformations. The phenomena of oscillation, of elastic impact, and of frictionless fluid motion are frequently treated, as in Chapters 12, 13, and 14, with the aid of the restricted form of the principle. But wherever friction or other types of dissipation of energy are involved, accompanied as they must ultimately be by the generation of heat, only the more inclusive form of the conservation principle is adequate. That form is variously phrased, but the following statement is perhaps as good as any (102:118):

> Energy is recognized in various forms, and when it disappears in one form it appears in others, and in each case according to a fixed rate of exchange.

> The total quantity of any energy, measured in terms of any one form, is constant whatever forms it may assume.

Many years were to elapse before the scientific world came to a general recognition of the full import of the doctrine of conservation of energy. Perhaps the most effective of its earlier champions was Hermann von Helmholtz, who was at that time (1847) principally known as a physiologist who was especially versed in the field of sound. It is somewhat thought-provoking to realize that the five men who were the first to comprehend the full import of the principle of the conservation of energy were all young men and were all professionally outside of the field of physics at the time that they made their contributions. These were Mayer, a German physician, aged twenty-eight; Carnot, a French engineer who preceded all the rest in the discovery and who will be discussed further in the next chapter, aged thirty-four; Helmholtz, a German physiologist, aged thirty-two; Joule, an English industrialist, aged twenty-five; and Colding, a Danish engineer who made the same discovery independently of the others and almost simultaneously, aged twenty-seven.

Perpetual Motion

It was the principle of conservation of energy, a principle which Poincaré has termed "the grandest conquest of contemporary thought," that finally set in strong relief the futility of the perennial efforts to devise a perpetual motion machine. A corollary of the principle that the sum of all the forms of energy output of a machine must always be exactly equal to the total energy input is that the output can never exceed the input. What constitutes the will-o'-the-wisp of the perpetual motionist has been the fond hope in many quarters that a machine could be devised whose output would be greater than its input. The principle of conservation of energy alone, therefore, should be enough to dispel this illusion.

Actually the case for perpetual motion is even less valid than the conservation principle, taken by itself, allows. The statement that the total energy output of a machine is always *equal to* the input includes in the energy output such items as friction and heat loss, inescapable characteristics of the operations of any machine. These forms of "output" are useless whereas the perpetual motionist is naturally interested only in *useful* output. The useful output of any machine is consequently always less than the input, the ratio of the two being termed *efficiency*, a term already introduced in the study of mechanics (page 151). The efficiency of common machines is much lower than the average individual realizes. That of a high quality internal combustion engine is about 40 per cent, of a high quality steam engine about 20 per cent, of a locomotive engine about 5 per cent. There are, moreover, certain losses in the transmission of mechanical or electrical energy to the locality of utilization and in its application at the point of consumption which diminish practical efficiency still further. When coal is burned to supply steam for the generation of electricity, only a small fraction of 1 per cent of the coal's heat energy appears in the illumination provided by the electricity. All the rest is lost along the way, dissipated in the form of heat. Even the light itself is absorbed and ultimately converted into heat.

The Degradation of Energy

The same can be said of any series of energy transformations whatever. There is a "tax," payable only in the form of heat, exacted at each step. All kinds of energy ultimately dissipate themselves in the form of heat. The tragedy of the process is that the heat thus generated can practically never be utilized. Once dissipated it can never be recovered or reclaimed. The user can only turn to the source, almost invariably the sun in the final analysis, for another handout, which will be similarly spent in its turn. Just as much energy resides in the dissipated heat as resided in the original handout from the sun, but it has been rendered unavailable. The process is technically termed the *degradation of energy.*

The degradation of energy was first identified by Lord Kelvin in 1851 and is perhaps as important a scientific generalization as is the conservation of energy. Kelvin pointed out that, as a consequence of this continuous universal process, the inescapable decreasing availability of energy indicates an ultimate stoppage of all energy flow in the universe, a "running down" of the cosmic clock. Long before that state is reached all life will have disappeared, for life depends on, and perhaps consists of, a peculiar form of energy flow. The end of the world is usually pictured as a state of utter frigidity. This is not necessarily a correct picture. The final state of the universe may be temperate or even hot. The important point will not be temperature level, but the entire absence of *differences* of temperature. Everything will be cold or medium or hot to the same degree, and consequently no energy can flow from one place to another. This state of affairs is inevitable, according to every evidence now available. Perhaps it is not necessary to take the matter too seriously, however, for even the most enthusiastic prophets of doom admit that many billions of years will elapse before that condition comes to pass, and a lot of things can happen in a billion years.

Questions for Self-Examination

1. How did the steam engines of the early eighteenth century differ from the later types?
2. Tell the story of Rumford's contribution to our knowledge of the nature of heat.
3. Tell the story of the first estimate of the value of the mechanical equivalent of heat by Count Rumford.
4. Tell the story of Mayer's prevision of the doctrine of conservation of energy.
5. Joule's scientific career was largely confined to measuring the mechanical equivalent of heat. Mention some of its episodes.

6. State the principles of conservation of energy and degradation of energy and sketch the prospect of the "heat death" of the universe.

Problems on Chapter 19

1. Coal is consumed at the rate of 56 pounds per hour under the boiler supplying steam to a 20 H.P. engine. Find the efficiency, assuming that 1 pound of coal can convert 15 pounds of water at the boiling point into steam at the same temperature. Take the heat of vaporization of water as 971 B.T.U. per pound. 6.3 per cent.
2. A horse working at 28,000 foot-pounds per minute is to be replaced by an engine of efficiency 4 per cent. How much coal must be supplied to the boiler? Take the heating value of the coal as the same as in the preceding problem. 3.7 lbs/hr.
3. The cylinder of an air compressor is cooled by water flowing at the rate of 4 kilograms per minute. If 10 H.P. is expended on the compressor, how much is the water being raised in temperature? 27° C.
4. In freezing weather ice forms in a tank at the rate of 1 kilogram an hour. What is the wattage of an electric light which, when immersed in the tank, will just prevent freezing? 93 watts.
5. A weight of M kilograms is attached by means of a cord and pulley to a paddle. The paddle is of the same material as the container in which it is mounted, the total mass of the paddle and container being m grams and their common specific heat s. A fall of the weight through a height of h meters is found to heat the container, paddle, and W grams of water through t° C. What is the mechanical equivalent of heat J in joules per calorie, assuming no loss by friction, radiation, or conduction?

M	m	s	h	W	t	J
10	1500	.093	30	1000	.62	4160
10	1500	.093	30	700	.84	4170
10	1500	.093	30	430	1.23	4200
10	1500	.093	30	240	1.85	4190

6. What speed in meters per second would be required of a lead bullet initially at t° C. to raise its temperature and just melt it on impact? Assume all the heat to remain in the lead. The specific heat of lead is .0320, its melting point is 327° C., and its heat of fusion is 5.86 Calories per kilogram.

t	v
0	370
100	330
200	290
327	220

CHAPTER 20

Heat Engines

Work as a Form of Heat

The usual conception of an "engine" as a source of power places at the center of the stage the picture of a great force: the force of expanding steam in the steam engine or of explosively burning vapors in the case of the internal combustion engine. Of course it is true that sources of commercial power do exert forces, yet those forces would be useless if their exertion did not result in motion. It is the rate at which an engine does work that counts — not primarily the force that it exerts. But whatever the immediate manifestation of work, the ultimate origin can only be a transfer of heat, a portion of which incidentally becomes converted into work. This idea was emphasized in the preceding chapter and will be further developed in this one. It suggests the reason why all devices for converting stores of heat into mechanical work are termed *heat engines*. It indicates also why James Watt's first contribution to the development of the steam engine — which was the prevention of a great waste of heat — was also his greatest contribution.

Watt's First Improvements of the Steam Engine

The old Newcomen engine which had been turned over to Watt for repair in 1763 even at its best was consuming at each stroke several times as much steam as was needed merely to fill the cylinder. This was because the steam came in contact with the cylinder walls, which were chilled from the jet of water that had condensed the previous cylinder-full, and most of it went into heating the walls up to its own temperature. So much was being lost in this way that Watt found that the engine was consuming at every stroke eight cylinder-fulls of steam. Watt, in collaboration with Black, had experimented extensively on the "latent heat" of steam, and he was acutely aware of the huge amount of heat which was being lost. It took him a long while, however, to see how it could be avoided.

After considering the possibility of making cylinders of some material of low specific heat, which would, therefore, warm up with little heat, he finally hit upon a much better idea. He is very explicit about the time and place where this came to him. He says[1] that he was walking on the Glasgow Green on a spring day in 1765 thinking of the engine,

[1] Hart, *Transactions of the Glasgow Archaeological Society*, *1*, 1 (1859).

> when the idea came into my mind that, as steam was an elastic body, it would rush into a vacuum, and, if a communication were made between the cylinder and an exhausted vessel, it would rush into it, and might be there condensed without cooling the cylinder.

Watt's idea of providing a separate *condenser* effected a greater improvement in the economy of steam-engine operation than any other invention but his second idea, that of utilizing the energy of expansion of the steam, was a close competitor to the first.

In all of the early engines, including those benefiting from Watt's first improvement, steam at full boiler pressure followed the piston clear to the end of its travel. When the exhaust valve was opened at the end of the stroke, this high-pressure steam escaped without doing any further work. This was another source of wastage to which Watt some years later turned his attention. He solved this problem by shutting the steam off early in the stroke and allowing it to expand, a principle which has characterized the operation of steam engines ever since. Under this arrangement the pressure in the cylinder falls gradually during the last portion of each stroke, almost or quite to that of the condenser. This greatly reduces the loss of power at each exhaust.

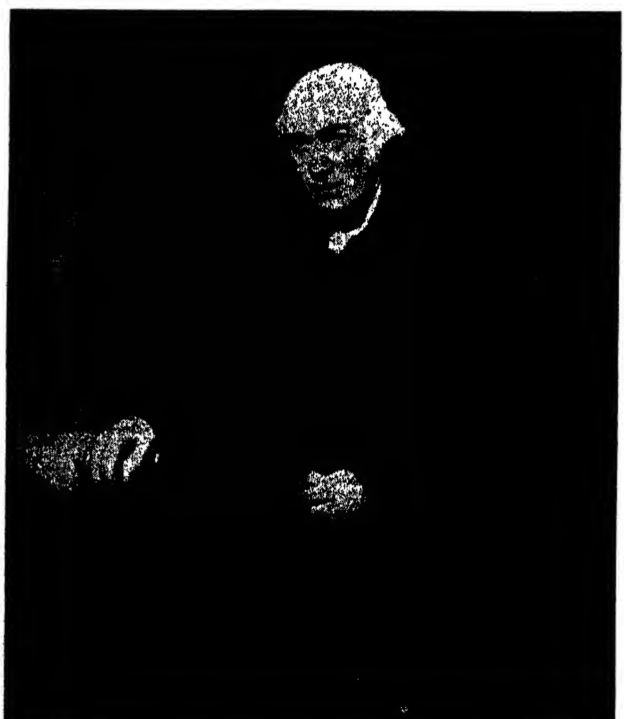


Fig. 162. James Watt (1736–1819)
(Courtesy of The Science Museum, London.)

As a current example of little or no expansion, we may consider a locomotive engine starting under a heavy load. The loud sound of its exhaust under these circumstances shows the wasted energy. Such sound, on the other hand, is almost entirely absent from a high-quality stationary engine. This performance is approached, indeed, by the locomotive engine as it picks up speed and the cut-off is progressively shifted to points earlier in the stroke.

The Pressure-Volume Diagram

Perhaps an understanding of the function performed by expansion will be facilitated by resorting to graphical representation. Let us take pressure-volume co-ordinates as was done in representing Boyle's law in Figure 88. Figure 163 represents an idealized case of zero expansion. Ordinates represent the pressure of steam in the space to the left of the piston. Abscissas represent the distance of the piston from the extreme left of its travel. Two simplifying assumptions are made, namely, that there is zero clearance — space behind the piston at the extreme left of its travel — and zero back-pressure — pressure in the space (for example, the condenser) into which the spent steam is exhausted. If steam from the boiler flows into the cylinder during the entire power stroke, as in the case of the old Newcomen engine and the modern locomotive engine under starting conditions, the cycle of events is represented by the rectangular p–v pattern of Figure 163. If, however, the steam is allowed to expand in the cylinder in consequence of cutting off the supply before the power stroke is complete, the cycle of events is represented by the p–v pattern of Figure 164. It will be evident that only about one half as much steam is consumed at each stroke as in the preceding case, but that much more than one half as much work is done.

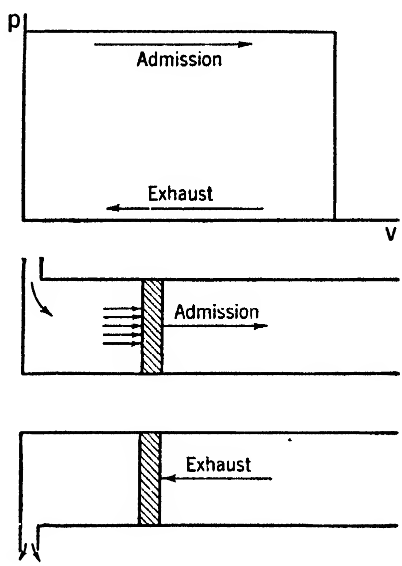


FIG. 163. CASE OF ZERO EXPANSION

The work done by the engine at each stroke is proportional to the area of the diagram. This is perhaps most immediately evident in Figure 163, where the area is simply the product of the two dimensions of the rectangle, that is, the product of the steam pressure by the volume swept out by the piston in one complete stroke. But this may be stated

$$(\text{pressure} \times \text{area of piston}) \times \text{length of stroke}$$

This, being force times distance, is obviously the work done by the steam at each stroke. The proportionality between the area of the diagram and the work done at each stroke still applies to Figure 164, though the determination of the area would be less simple for that case.

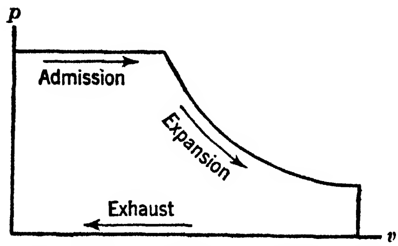


FIG. 164. UTILIZATION OF EXPANSION

Another of Watt's great contributions to the development of the steam engine was the devising of an instrument called a steam engine indicator, which would record just such diagrams for engines in actual operation. *Indicator diagrams*, so called, have been of immense utility to engineers ever since. They not only give information on the power developed by an engine, that is, work per stroke multiplied by number of strokes per second, but — what is even more useful — disclose the faults in adjustment of the valve gear and give other types of information on the working condition of an engine.

Of Watt's other numerous improvements of the steam engine, the most notable is making engines *double acting*; that is, applying the steam first to one side and then to the other of the piston, instead of limiting it to one side as in all engines up to 1782. All in all, Watt's work was so far-reaching that for a century after his death further improvements in the steam engine could only be made in details.

The Steam Engine as a Heat Engine

But the impossibility of surpassing Watt in improving the practical operation of the steam engine did not preclude the possibility of deepening the scientific comprehension of the principles involved. This was the next step, and it was taken by a brilliant young French engineer, S. N. L. Carnot (1796–1832), in 1824. By that time the steam engine had become very common in industrial practice, though it had not yet been applied extensively to transportation. The Atlantic had been crossed only once under steam power (1819), and the locomotive had not yet come into existence. Carnot sensed that the theory of the steam engine was comprehended vaguely or not at all, and set himself the task of correcting this.

Carnot's great contribution was in directing attention to the fact that the real source of "motive power" was difference of temperature. The exertion of pressure on a piston was an incidental detail of a complicated process, the heart of which was the flow of heat. He said (76:7–9):

> The production of motion in the steam engine is always accompanied by a circumstance which we should particularly notice. This circumstance is the passage of caloric from one body where the temperature is more or less elevated to another where it is lower. What happens, in fact, in a steam-engine at work? The caloric developed in the fire-box as an effect of com-

> bustion passes through the wall of the boiler and produces steam, incorporating itself with the steam in some way. This steam carrying the caloric with it, transports it first into the cylinder, where it fulfills some function, and thence into the condenser, where the steam is precipitated by coming into contact with cold water. As a last result, the cold water in the condenser receives the caloric developed by combustion. It is warmed by means of the steam as if it had been placed directly in the fire-box....
>
> Everywhere where there is a difference of temperature, and where (a flow of) caloric can be effected, the production of motive power is possible. Water vapor is one agent for obtaining this power, but it is not the only one; all natural bodies can be applied to this purpose, for they are all susceptible to changes of volume, to successive contractions and dilations effected by alternations of heat and cold; they are all capable, by this change of volume, of over-coming resistances and thus of developing motive power.... The vapors of all bodies which are capable of evaporation, such as alcohol, mercury, sulphur, etc., can perform the same function as water vapor.

Carnot's extension of the list of substances capable of utilization in power production is interesting in view of subsequent developments. He did not cite gasoline, a substance unknown in his day, but he did mention alcohol, the use of which is on the increase, and mercury, which was first used for such a purpose only a little over a decade ago.

The emphasis which Carnot thus placed on the central rôle played by the transfer of heat in the operation of what he termed "heat engines" was of the utmost significance. He thereby put his finger on the crucial element in all engines, an idea which seemed to have occurred to no one before him. He was most explicit about it (77:20–21):

> The motive power of heat is independent of the agents employed to develop it; its quantity is determined solely by the temperatures of the bodies between which, in the final result [i.e., upon the completion of a cycle of operations] the transfer of the caloric occurs.

The idea was of even broader import than Carnot realized at the time, for he specifically excepted machines "worked by men or animals, by waterfalls or by air currents." Not possessing the concept of the mechanical equivalent of heat, he was in no position to realize that these also, "in the final result," were subject to his great generalization that "motive power" was always attributable to heat transfer.

The Conservation of Energy

Carnot made another error connected with his association of motive power with heat transfer. It was much more serious than his simple error of omitting engines which were not apparently of the heat-transfer type. The more serious error arose, as did the lesser error, because he did not possess the concept of the mechanical equivalent of heat. Out of his preoccupation with the flow of heat in connection with engines he jumped to

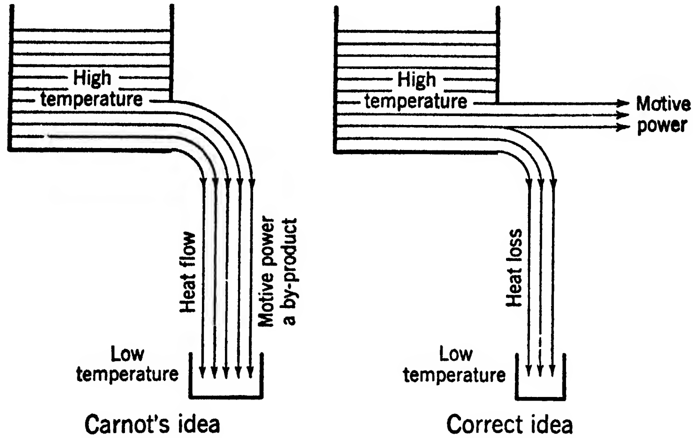


FIG. 165. CARNOT'S IDEA AND THE CORRECT IDEA OF THE HEAT FLOW IN AN ENGINE

the conclusion that the theoretically perfect engine would simply transfer *all* the heat that it received from the source of higher temperature, intact to the region of lower temperature. He used an analogy which was quite plausible.

> The production of motive power in the steam-engine is therefore not due to a real consumption of the caloric, *but to its transfer from a hotter to a colder body*. . . . We may with propriety compare the motive power of heat with that of a waterfall. . . . The motive power of falling water depends on the quantity of water and on the height of its fall; the motive power of heat depends also on the quantity of caloric employed and on that which might be named, which we, in fact, will call *its descent* — that is to say, on the difference of temperature of the bodies between which the exchange of caloric is effected.

The italics in the foregoing equation are Carnot's own. His idea was perfectly clear and his analogy rather convincing. But it was wrong. We now know that a part of the heat supplied to any engine is diverted and changed to mechanical energy, so that less heat is delivered to the exhaust of even a perfect engine than is supplied from the source. Subsequent measurements (47a) showed that the heat delivered at the exhaust of a steam engine was less than that supplied to it, the difference being *the heat equivalent of the work done by the engine* (see Fig. 165).

The principle of conservation of energy makes this conclusion mandatory for us, but that principle was unknown in 1824. It was, in fact, the apparent contradiction between Carnot's theory and the doctrine promulgated twenty years later by Mayer and by Joule which was primarily responsible for the doubts with which the latter was regarded. This error of Carnot's is another misdemeanor which must be chalked up to the discredit of the caloric theory of heat. It was his commitment to this theory which caused Carnot to be impressed by the analogy between the waterfall and the heat-engine and which led him to identify his idea of a sort of "conservation of heat" with the doctrine of *conservation of matter*, which had been experimentally established twenty years before.

Though this was Carnot's influence on the development of scientific doctrine, he apparently changed his opinion before his death in 1832. Some of his later papers, published posthumously in 1872 in connection with a second edition of his *Reflections*, make it unmistakable that he had acquired a clear prevision of the principle of conservation of energy; that he had even made a surprisingly accurate determination of the mechanical equivalent of heat; and that he had laid out for himself a program of investigation which included all the important developments in this field made by others in the ensuing thirty years, a truly amazing foresight. If he had not been cut down during an epidemic of cholera at the age of thirty-six, he would probably have developed into one of the greatest men of science of all time.

The Reversibility of an Engine Cycle

Carnot discovered another attribute of engines which was only second in importance to his emphasis on their heat relations, namely, their reversibility. This does not refer to anything as trivial as the direction of rotation. What it refers to is the possibility of actually reversing the operation which an engine normally performs. In normal operation an engine produces mechanical energy by taking heat from a region of high temperature and delivering some of it to one of low. Therefore, when "reversed," it should be capable of absorbing mechanical energy to effect a transfer of heat from a region of low temperature to one of high. Carnot envisioned this with his remark that

> wherever there is a difference of temperature the production of motive power is possible.

He added,

> Conversely, wherever this power can be employed, it is possible to *produce* a difference of temperature.... To effect this the operations which we have just described could have been performed in a reverse sense and order,

and he proceeded to describe this reversal in detail.

The process (represented in Figure 166) is simply the reverse of that pictured in the right-hand portion of Figure 165. Energy is furnished to the system at the point marked "motive power input." This "pumps" heat from the cold region to the hot region, adding its own heat equivalent to the quantity thus delivered to the region of high temperature. This is simply the working principle of the ordinary electric refrigerator. The cooling unit is the "low temperature" part of Figure 166. Heat is pumped out of it by the compressor, "motive power input," into the condenser, "high temperature," and thence escapes into the room outside the refrig-

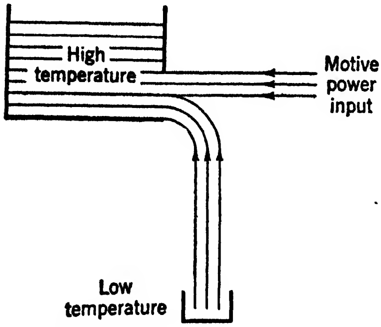


FIG. 166. THE ACTION OF AN ENGINE REVERSED

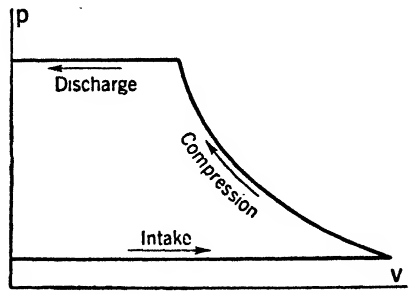


FIG. 167. THE ACTION OF A COMPRESSOR

erator. Thus the compressor of such a refrigerator is simply an engine with its action reversed.

While minor alterations might be required in the mechanism, an engine and the compressor of a mechanical refrigerator are in principle identical. The cycle of their operations can be the same, except that it is performed in reverse order. (Compare Figure 167 with Figure 164.) The area of the indicator diagram is proportional to the work done during the cycle, as before, except that it is now work absorbed instead of work evolved.

The Mechanical Refrigerator

Figure 168 shows the operation of one of the two types of refrigerator now in common domestic use, the *compression type*. As in all types, the refrigeration is produced by the evaporation of a volatile liquid in the evaporator or "cooling unit." The resulting vapor is then pumped up to high pressure by a motor-driven compressor. The high-pressure vapor is condensed back into a liquid, usually by an air-cooled "condenser." It is at this point that the heat removed from the interior of the refrigerator plus the heat equivalent of the work done by the electric motor on the compressor is delivered to the room. The liquid then enters the cooling unit, is vaporized, and the cycle is repeated.

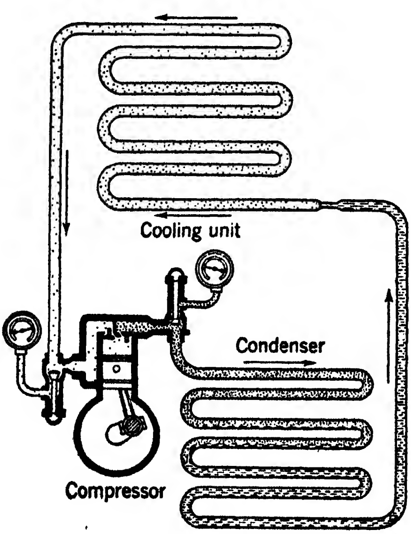


FIG. 168. COMPRESSION TYPE OF MECHANICAL REFRIGERATOR ("ELECTRIC REFRIGERATOR")

Figure 169 shows the operation of the other type of mechanical refrigerator in common domestic use, the *absorption type*. This type differs from the former principally in the agency used to remove the vapor from the evaporator, condense it back to a liquid, and then complete the cycle. The two processes are identical as far as what occurs in the cooling unit is concerned, the evaporator being the actual point of application of refrigeration, whatever the system.

The absorption type of refrigerator uses a refrigerant — ammonia — which is readily soluble in water. A convection current, maintained by a flame — usually of gas, whence the name "gas refrigerator" — carries the ammonia to the *absorber*, where it is dissolved in water. The solution then flows to the *generator*, where the ammonia is boiled out of the water again. The water, raised by the agency of the heat through a process made familiar in coffee percolators, flows back into the absorber. The hot ammonia vapor rises to the condenser where, after being converted into a liquid, it flows by gravity back to the evaporator, thus completing the cycle. The cycle of events is aided by the presence of a gas — hydrogen — under a pressure of several hundred pounds, which is sealed into the system, along with the ammonia and water, in the process of manufacture of this type of refrigerator.

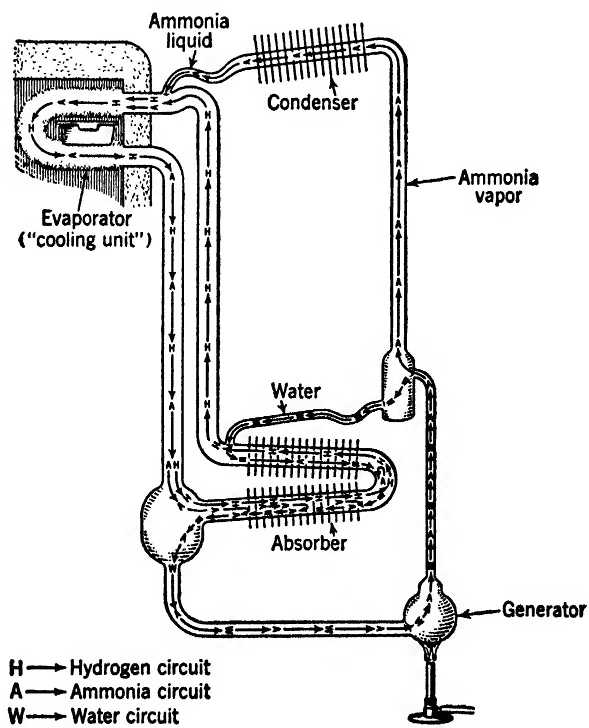


FIG. 169. ABSORPTION TYPE OF MECHANICAL REFRIGERATOR ("GAS REFRIGERATOR")
(Courtesy of Servel, Inc.)

To the uninitiated the spectacle of "cooling by heat" appears paradoxical. It is a "paradox," however, which is equally characteristic of the "electric" and the "gas" types of refrigerator. In both cases energy is required to lift the heat out of the refrigerator and deliver it to the higher temperature of the surroundings. Whether the energy is furnished in the form of electricity or heat is an incidental detail. If the system is of a type which requires heat, heat could be furnished equally well by electricity or gas. If it requires mechanical energy, that too could be provided either by an electric motor or, say, by a steam engine supplied from a boiler fired by gas. The central point in any type of refrigerator is the element foreseen by Carnot, the absorption of energy involved in the engine cycle when reversed.

Refining the Concept of Reversibility

Yet the concept of reversibility in its full sweep involves more than merely reversing the *function* of an engine. It requires that the engine shall be theoretically perfect, be without friction, and work without any heat losses or other dissipation of energy. For energy so lost cannot be recovered, and a machine which incurred such losses could never be truly reversible. A steam engine cycle would not necessarily be reversible in this way, even if the engine were perfectly insulated against heat losses. In fact special precautions would have to be observed to make it reversible. For example, if at the end of its stroke the pressure of the expanded gas is greater than that in the condenser, as is usually true in actual operation, it escapes with a characteristic "puff" as the exhaust valve is opened, and this constitutes an irreversible event. A moment's thought will make it clear that the attempt to reverse such an event would be unsuccessful. For when the intake valve of a compressor closes and compression is about to begin, the gas could not be expected to raise its own pressure spontaneously before compression actually begins, to the extent that it dropped its pressure at the same point previously (Fig. 170). There are other ways in which a cycle can be rendered irreversible and most of them get in their work in practice, especially in the innumerable ways in which heat can be lost to the surroundings in effecting a cycle of such operations. But in theory, the reversible cycle is a possibility of which

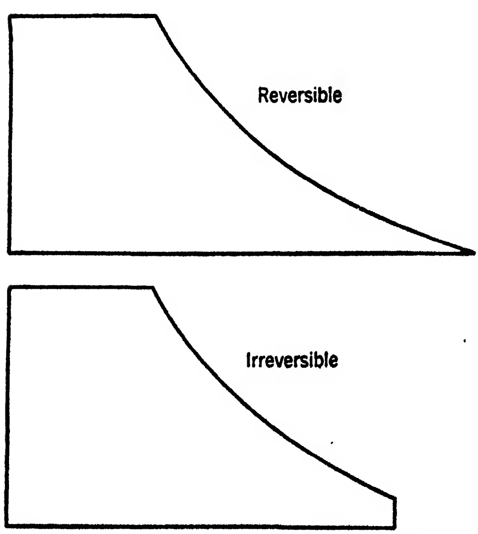


FIG. 170. ONE WAY TO MAKE A CYCLE IRREVERSIBLE

much has been made, entirely aside from its utility in the mechanical refrigerator. It will be utilized in the present chapter.

The Efficiency of a Heat Cycle

The efficiency of any operation has been defined (page 151) as the ratio of useful energy yielded to total energy absorbed. Originally formulated to apply to merely mechanical operations, it is still valid when heat and other forms of energy are included. Specifically, it is applicable to the engines which have been under discussion. A certain quantity of heat, Q_1, leaves the high temperature source in connection with the operation of an engine. Part of it, Q_2, passes through the engine and goes mainly out of the exhaust. The difference $(Q_1 - Q_2)$, represents the quantity of heat converted into mechanical energy by the engine. The efficiency of the process is

$$\text{efficiency} = \frac{Q_1 - Q_2}{Q_1}. \quad (1)$$

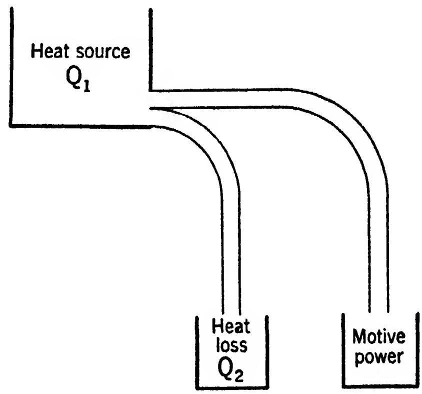


FIG. 171. HEAT PATH IN A REVERSIBLE ENGINE

A somewhat different expression for the efficiency of an engine was deduced by Carnot. He derived it from his water-power analogy (page 214), picturing motive-power as depending on

> the quantity of caloric and on the . . . difference of temperature,

considering it proportional to the product. Hence, taking Q as the quantity of heat descending — on Carnot's view — from a region of absolute temperature T_1 to a region of absolute temperature T_2, the work actually done was proportional to the product $Q\,(T_1 - T_2)$ and the maximum work theoretically obtainable was QT_1. Hence, the efficiency was

$$\text{efficiency} = \frac{Q\,(T_1 - T_2)}{QT_1} = \frac{T_1 - T_2}{T_1}. \quad (2)$$

Though this expression, like Black's concepts of specific and "latent" heats, was deduced on a false premise about the nature of heat, it was subsequently found, also like Black's results, to be nevertheless correct. If, now, there is no loss of heat, all the rejected heat, Q_2 going into a perfectly insulated reservoir at temperature T_2, and being capable of being "pumped" back to the source undiminished by reversing the action of the engine, then expressions (1) and (2) represent the efficiency of a reversible engine and can be equated to each other. Hence, for a reversible engine

$$\frac{Q_1 - Q_2}{Q_1} = \frac{T_1 - T_2}{T_1}. \quad (3)$$

Either side of equation (3) represents the maximum efficiency any engine can have. The assertion that *no engine can be more efficient than a reversible engine* is known as *Carnot's theorem.*

Improving the Efficiency of the Steam Engine

Equation (3) shows that the efficiency of a reversible cycle increases with increase of difference of the temperatures between which it operates. This is also true of the non-reversible cycles encountered in daily engineering practice. Efficiency is thus increased with any rise in T_1 or diminution in T_2. Carnot recognized and commented on this fact (76:51). A practical lower limit is set on T_2 by the prevailing temperatures of the air into which the exhaust usually occurs or of the cooling-water in the case of engines equipped with a condenser. It is only in connection with T_1 that any latitude occurs. Hence, the tendency in advanced engineering practice is toward higher and higher steam-pressures, with accompanying increases in steam temperatures. Still higher temperatures are sometimes secured, without the extra hazard of excessive pressures, by superheating the steam. This consists of applying extra heat to the steam after it has been separated from the water, thus causing it to lose the properties of a vapor and assume more unequivocally the properties of a gas. All the "permanent" gases are simply the very highly superheated vapors of substances that can exist as liquids only at extremely low temperatures. Equation (3) shows that the maximum efficiencies attainable in a steam engine working between atmospheric pressure and boiler pressures of 50, 100, 200, and 300 pounds per square inch (temperatures 138° C., 164° C., 194° C., 214° C.) are respectively 9, 15, 20, and 24 per cent. The corresponding efficiencies when a condenser is added, providing an exhaust temperature of, say, 30° C., are 26, 31, 35, and 38 per cent. With the steam superheated 100° C. above the normal temperature in addition, the corresponding efficiencies would be 41, 44, 47, and 49 per cent. In practice, the best operating efficiencies seldom exceed one half of those indicated here as the highest theoretically obtainable.

The Steam Turbine

The difficulty of providing effective lubrication at such high temperatures is one of the reasons why the steam turbine, a steam "windmill" which has crashed the gate into high engineering society, has largely displaced the reciprocating steam engine. The last word in turbines is one using mercury vapor instead of water. It works over a temperature range of 450° C., to which its high theoretical efficiency is due (60 per cent). The exhaust mercury vapor is used to produce steam which is supplied to a steam engine, thus adding materially to the efficiency of the process. It is interesting to note in this connection that Carnot said (76:58):

> It would no doubt be preferable if there were an abundant supply of a liquid which evaporated at a higher temperature than water, the specific

> heat of whose vapor was less for equal volume, and which did not injure the metals used in the construction of an engine; but no such body exists in nature.

Mercury answers these specifications today, though it did not in Carnot's time, when it was not "abundant" enough to be used for such a purpose. Moreover, all engine parts can today be made of steel which mercury does not attack. Many of them in Carnot's time had to be made of brass which disintegrates almost immediately when placed in contact with mercury at high temperature. The great difficulty with mercury, in addition to its high cost, is the violently poisonous nature of its vapor.

The Internal Combustion Engine

An account of heat engines would not be complete if it did not include what is now the most common type of all, the internal combustion engine, most common in automobile and aviation practice. Electric power plants are using Diesel engines (page 222) quite commonly; Diesels are beginning to displace steam turbines in marine practice; and they are even becoming evident in the most advanced railroad practice.

As the name implies, the chief characteristic of the internal combustion engine is the fact that combustion of its fuel takes place inside of the cylinders instead of outside as is the case with steam. In effect the internal combustion engine uses air as its working substance instead of steam, the air being heated by the burning of the vaporized fuel with which it is impregnated. In this connection, too, Carnot was prophetic. He remarked (76:56):

> Water vapor can be formed only by the aid of a boiler, while atmospheric air can be heated directly by combustion within itself. Thus a considerable loss is avoided, not only in the quantity of heat, but also in its thermometric degree.

The internal combustion engine which Carnot thus envisioned did not come into existence for more than forty years and did not begin really to compete with the steam engine for nearly a century.

The Otto Cycle

The details of operation of the automobile type of engine are so familiar that it is scarcely worth while dwelling upon them. The accompanying illustration (Fig. 172) shows the nature of the four events constituting the cycle of operations, namely, intake, compression, power, exhaust. These events occur on successive strokes, and each requires substantially the whole of its stroke. This is termed the Otto cycle, first actually effected by a German engineer of that name in 1876, though it had been proposed fourteen years before. The cycle itself is shown in Figure 173. The ignition, accompanied by a large and almost instantaneous increase in pressure

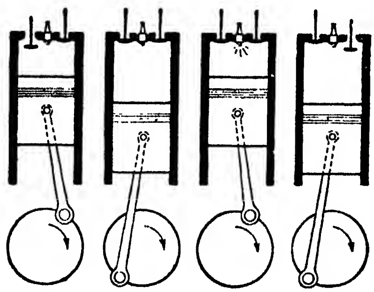


FIG. 172. THE SEQUENCE OF EVENTS IN THE OTTO CYCLE

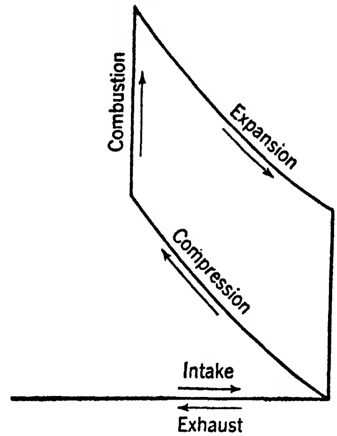


FIG. 173. THE OTTO CYCLE

while the piston is near the inner end of its stroke, might be termed a fifth event in this cycle. It is effected by an electric spark.

The engine described in the preceding paragraph requires four successive strokes (two revolutions) to complete its cycle. Small engines, such as those used as the outboard motors of small pleasure boats, complete the cycle of four events in only two strokes (one revolution). The former engine is called a *four-stroke-cycle* ("four-cycle") engine and the latter a *two-stroke-cycle* ("two-cycle"). The two-cycle performs the power and exhaust on the forward stroke, the intake and compression on the return. It has the advantage of a power stroke every revolution instead of every alternate revolution as in the four-cycle, and can, therefore, be made lighter for the same power. But it is far less efficient.

The Diesel Engine

Though the efficiency of the Otto cycle is higher the greater the compression ratio, there is a practical limit. The explosive mixture which experiences the compression is also heated by it. For compression ratios greater than about 6:1, the mixture, if the fuel is gasoline, reaches its ignition temperature before the end of the compression stroke. The explosion thus occurs before the spark, that is, too early in the cycle, and thus administers an impulse tending to reverse the engine. This produces a "knock" and causes loss of power. The sole function of "ethyl" is to defer the ignition and make it take place more slowly, thus facilitating higher compression ratios.

But even "ethyl" is only a palliative. In 1898, Rudolph Diesel built an engine which solved the compression dilemma. He allowed only air in the cylinder during the compression. In the absence of a combustible mixture, no pre-ignition could occur, and there was, therefore, no limit to the attainable compression except the structural strength of the engine In practice the Diesel engine uses compression ratios about two to three times as great as the highest used in automotive practice. The fuel is injected into the cylinder in a high-pressure spray. It ignites instantly on account of the high temperature of the compressed air, no electric ignition system being required. The spray continues during an appreciable fraction of the power stroke, at such a rate as to maintain the pressure substantially constant. The remainder of the stroke constitutes the expansion of the products of this combustion. The form of the Diesel cycle is shown in Figure 174. The Diesel engine is about 20 per cent more efficient than engines operating on the Otto cycle. They are rapidly displacing the steam engine for industrial use. They possess the disadvantage of great weight of the material required to withstand the stresses due to high compression, and of increased friction due to tighter fit between piston and cylinder walls to prevent leakage at the high pressures used.

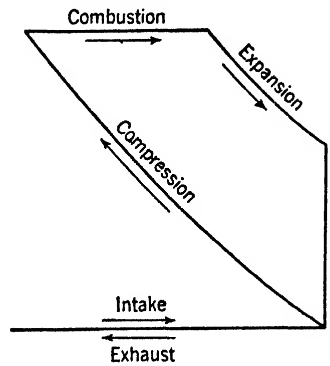


FIG. 174. THE DIESEL CYCLE

The Diesel engine is being adapted to automotive use through a modification termed the "semi-Diesel." In this the compression ratio is reduced and the cylinder wall artificially heated to raise the temperature above the ignition point. It shares with the Diesel the advantage of being able to use crude oils as fuel, by virtue of the high temperature at which it operates. It sacrifices some of the theoretical efficiency of the Diesel in the interests of portability. As soon as the economic resistance from manufacturers and vendors of gasoline can be overcome, the semi-Diesel bids fair to displace the gasoline engine at least for heavy automotive and airplane use.

Questions for Self-Examination

1. What were James Watt's principal contributions to the development of the steam engine?
2. Draw an idealized indicator diagram for a steam engine, identify its branches and show how it may be made to yield information on the power being developed by the engine.
3. Outline Carnot's contributions to the theory of the heat engine. Identify the famous error of his theory.
4. Discuss the concept of reversibility of the heat engine and its bearing on theoretical efficiency.
5. State the principle of the mechanical refrigerator. Compare the "electric" type with the "gas" type.
6. What is the basis of Kelvin's so-called "thermodynamic" scale?
7. Describe the Otto cycle and the Diesel cycle, with the aid of indicator diagrams.
8. Compare the "four-cycle" engine with the "two-cycle" engine.

Problems on Chapter 20

1. A steam engine indicator is so geared that each centimeter of abscissa of the diagram corresponds to $\frac{1}{2}$ foot of stroke and each centimeter of ordinate corresponds to steam pressure of 25 pounds per square inch in the cylinder. The diameter of the piston of the engine is 15 inches. The engine speed is 60 revolutions per minute. If the area of an indicator diagram is 25 square centimeters, what is the "indicated horsepower"? 200.

2. If an automobile engine working at 25 H.P. uses 3 gallons of gasoline per hour, what is its efficiency? The heat of combustion of gasoline is 29,300 Cals/gal. 18 per cent.

3. In measuring the specific heat of a gas, one applies heat and observes the number of Calories required to heat one kilogram of the gas through 1° C. The value comes out to be less if the gas is confined to a constant volume while being heated than when it is allowed to expand, maintaining a constant pressure. Why should these two values of the specific heat be different?

4. Ten grams of air (density 1.3 kg/m^3 at 0° C.) are heated from 0° C. to 100° C. under atmospheric pressure. What is (*a*) the change in volume, (*b*) the work done by the expanding air, (*c*) the heat equivalent of this work? .0028 m^3; 280 joules; .068 Cals.

5. The specific heat of air at constant pressure is .237 Cal/kg. How many Calories were required to heat the air under the conditions of the preceding problem? What is the specific heat of air at constant volume? .18 Cal; .17 Cal/kg.

6. A steam engine working between the temperatures of 250° C. and 40° C. has an efficiency of 20 per cent. How does this compare with the efficiency of a Carnot engine working between the same temperatures? About $\frac{1}{2}$.

7. The volume swept out in one stroke of a certain pump is .1 cubic meter. The average steam pressure is 10 kilograms per square centimeter. Find the work done per stroke. If one stroke consumes .5 kilogram of steam, how much heat is needed to produce it? What therefore is the efficiency? 100,000 joules. 270 Cals. .99 per cent.

8. A locomotive engine uses steam at a pressure of P pounds per square inch (absolute) and exhausts it at p. Corresponding temperatures of the steam in degrees C. are given below. What is the maximum theoretical efficiency of the engine?

P	T	p	t	e (%)
250	205	30	126	17.
200	195	45	135	13.
150	181	60	145	7.9
100	164	75	153	2.5

9. Calculate the horsepower P and the thermal efficiency e of an engine from the following details:

diameter of cylinder	d inches
length of stroke	s feet
speed	n revolutions per minute
mean steam pressure	p pounds per square inch
consumption of coal	c pounds per hour.

One pound of coal will provide 3000 Calories in burning. One horsepower is the equivalent of 746 watts. The mechanical equivalent of heat is 4183 joules per Calorie.

d	s	n	p	c	P	e
15	3	80	75	300	190	.14
15	3	70	60	250	140	.12
15	3	60	45	200	87	.09
15	3	50	30	150	48	.07

SOUND

CHAPTER 21

The Nature of Sound

The Production of Sound

The most casual observation shows that sound always originates in some kind of motion. The sound of an explosion, the thud of a heavy fall, the impact of a hammer, the tone of a violin string, or the raucous warning of an automobile horn, all are clearly produced by rapid motion. In some instances motion is not so evident. The voice of a singer, the shrill sound of a whistle, the majestic tone of an organ pipe — in these we must be content with less direct evidence of the existence of motion, either because the moving object is concealed, as are the vocal cords of the singer, or because the moving object consists of a fluttering fin of air and, hence, is not easy to identify. But in every case the origin of a sound can be traced to motion of some kind.

Since energy must be expended to maintain the motion and acceleration associated with origins of sound, we are led to consider whether at least a part of that energy may not go into the sound itself. It is, in fact, possible to demonstrate that mechanical energy is associated with sound. As would also be suspected, the energy content becomes greater with increased loudness of sound.

But the energy content of ordinary sounds is exceedingly small in comparison with that involved in most of our activities. The volume of sound produced by a brass band is moderately large. Yet to produce sufficient sound to contain one horsepower, it would require a band of ten million pieces, playing fortissimo. If all the energy of normal speech could be converted into heat and used to warm a cup of tea, it would require steady conversation by a million people for an hour and a half to raise the tea to the proper temperature. The power emanating in sound from a group of ten thousand people cheering their loudest would be about the same as that required to maintain an ordinary electric light.

The energy content in sound of any kind is entirely incommensurate with the work required to produce the sound. In other words, sound-producing devices, regarded as machines, are uniformly inefficient in the extreme. One of the most efficient sound-producing instruments is the "loud speaker" familiar in radio. Even the best of these converts into sound only about 5 per cent of the electrical energy which it receives. In most musical instruments, the mechanical efficiency is probably but a tiny fraction of 1 per cent. Very recently certain types of "giant speakers," used for special kinds of sound projection, have been devised which convert into sound somewhat more than 50 per cent of the energy which they receive. These are marvels of sonic efficiency.

Sound Requires a Material Medium

Sound requires a material substance as the medium of travel. Any substance will do, whether it be gas, liquid, or solid. Without some medium to bear it, sound cannot travel from place to place. A bell in an evacuated glass jar is almost silent even though its clapper can be seen to beat against the gong

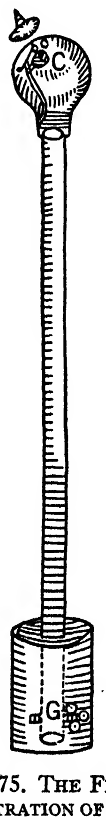


FIG. 175. THE FIRST ILLUSTRATION OF THE BELL *in vacuo* EXPERIMENT
(From Kircher's *Musurgia Universalis* of 1650.)

Figure 175 shows the first attempt to perform the bell *in vacuo* experiment. It is from a work of 1650 by Athanasius Kircher, a voluminous encyclopedist of the seventeenth century. The iron clapper of the bell inside the glass globe at the top was actuated from outside by a "vigorous lodestone." This was long before the day of air pumps, and the requisite vacuum was produced by a Torricellian mercury column (page 83); hence, the long vertical tube was necessary.

Neither this experiment nor similar ones that followed during the next ten years led to the correct conclusion. All indicated that sound would traverse a vacuum. Robert Boyle in 1660 improved the air pump recently invented by von Guericke (page 85). With its aid he repeated the experiment and interpreted it correctly. He said (84:21):

> The experiment was repeated with the suspending in the receiver a watch with a good alarum, which was purposely set that it might, before it should begin to ring, give us time to cement on the receiver very carefully, exhaust it very diligently, and settle ourselves in a silent and attentive posture.... We silently expected the time when the alarum should begin to ring. . and were satisfied that we heard the watch not at all.

The nature of sound was at that time a question largely without an answer. The importance of the bell *in vacuo* experiment, and especially of its correct interpretation, lay in the indication that it gave of the necessity for a medium for the transmission of sound and the consequent implications about the nature of sound itself.

The Speed of Sound in Air

That sound requires time to travel is indicated by many common observations. The lapse of time between a stroke of lightning and the ensuing thunder, between the puff of smoke emerging from a starter's pistol and the report of the explosion, between the stroke of a distant hammer and the sound of the impact: all these are occurrences of common observation. In each case the lapse of time may be taken as a measure of the distance, since it is now known that sound travels through air somewhat over a mile in five seconds (1087.13 feet or 331.36 meters per second at 0° C.).

Perhaps the most significant thing about the speed of sound in air is the fact that there is a definite value for it. One might be justified in asking, for example, whether loud sounds do not travel more rapidly than weak ones, or whether high-pitched sounds do not travel at different speeds from low-pitched sounds. The answer to the first question would be in the negative, though with a qualification to be noted later. The answer to the second is also negative, this time without any qualification. If the large and small instruments of a band are in time with each other at the place where the music originates, they will also be in time with each other as far as the band is heard. This would not be true if sounds of different pitch traveled at different rates. If they did, the piccolo might be heard a measure or two before the bass horn, or perhaps after it. Though the speeds of some varieties of wave do depend on what for them is the equivalent of pitch in sound (e.g., water waves), this is not the case for sound waves.

It has long been known that the speeds of extremely loud sounds were greater than those of sounds of ordinary intensity. The abnormal speeds of the sounds of concussions produced by heavy artillery were studied in 1918 by D. C. Miller (85a:121 ff.). He found in a representative case that the speed of the sound produced by a 10-inch rifle was nearly 2500 feet per second near the muzzle, but that when the sound wave was 100 feet away from the muzzle, its speed had diminished to 1700 feet per second. At 200 feet it had diminished to 1300 feet per second, and at 300, to about 1100 feet per second. By the time the wave had traveled 650 feet, which was half a second after the explosion, its speed had fallen to within $\frac{1}{100}$ of 1 per cent of the normal value and remained at the normal value thenceforth. This makes it clear that any increase in the speed of sound due to the loudness is a decidedly abnormal condition, encountered only in extreme cases. Ordinarily it is safe to assume that all sounds travel at the same speed.

Speed of Sound in Other Media

The speed of sound in other media than air is usually not a matter of great practical importance. In other gases, the speed depends upon the density (inversely as the square root), so that in hydrogen, for example, the density of which is about .07 that of air, the speed of sound is $\frac{1}{\sqrt{.07}}$ or 3.8 times its speed in air. More specifically, speed in gases is related to density by the relation

$$V = \sqrt{\frac{\gamma p}{d}}, \tag{1}$$

where d is the density in kg/m^3, p the pressure in newtons/m^2, γ is the ratio of the specific heat at constant pressure to that at constant volume (see page 204), and V is the speed in meters per second. At first glance it would appear that the speed depended upon pressure as well as density. The contrary is the case, however, speed of sound in gases remaining the same at all pressures. The reason is that slow changes in pressure (such that heat of compression is dissipated and Boyle's law thus approximated) produce proportional changes in density, so that the value of the fraction remains unaffected. If, however, an increase of temperature occurs, then pressure rises if the gas is confined, or density diminishes if it is not. In either case the ratio p/d increases in proportion to the ratio of the absolute temperatures, resulting in an increased value for the speed of sound in the gas. This is treated further on page 231.

In liquids and solids the speed of sound depends also upon the modulus of elasticity, directly as the square root. The numerical values of elastic moduli for liquids and solids are very high, so that as a practical matter they dominate the value of the speeds of sound in such media rather than density which dominates in the case of gases. For both liquids and solids the relation is

$$V = \sqrt{\frac{E}{d}}. \tag{2}$$

E represents the volume modulus of elasticity in the case of liquids and Young's modulus in the case of solids. The remaining notation is the same as for equation (1).

In water, the speed of sound is nearly a mile per second (4800 feet at 70° F.) while in steel it is more than three miles per second (about 16,300 feet, depending on quality of steel). A manifestation of this high speed in steel may be observed on a railroad track. If the rail is struck, an observer a hundred or more feet distant will hear two reports: the first coming through the rail, the second through the air.

Recently sonic methods have been developed for measuring ocean depths. An electric oscillator, incorporated in the hull of a ship, sends out signals which, after reflection at the bottom, are received again at the ship. The time elapsing is a measure of the depth. In this way a continuous record of ocean depth can be kept by a ship traveling at normal speed, with considerably greater accuracy than was possible with the old method of the lead line, which necessitated a complete stop every time a depth measurement was made.

Similar methods of sounding are being developed in what is known as geophysical prospecting. Oil-bearing or ore-bearing layers are characterized by different density than the adjacent layers. A sound wave traveling through the earth will experience partial reflection upon passing from one of these layers to another, and these reflected waves may be made to yield information to observers on the surface. Accordingly, one of the methods of geophysical prospecting is to produce a sound wave in the body of the earth by exploding a buried charge of dynamite. Properly located instruments record both the time of the explosion and the time of the arrival of waves reflected from the various subterranean levels. These may betray the location of the deposits being sought.

The Effect of Temperature on Speed

The speed of sound is affected somewhat by temperature, being increased in air by about 1.1 feet per second for every degree rise in temperature on the Fahrenheit scale (.6 meter per second for each degree centigrade). It is for this reason that temperature is always specified when a value for the speed of sound is stated. One result of the effect of temperature is that ordinary sounds may frequently be heard in a still atmosphere for abnormally great distances over water, especially after sunset. This happens when the higher levels of air are distinctly warmer than the lower. The greater speed at the upper levels causes the sound pulses to "tip forward" as shown in part (*b*) of Figure 176, and thus to bring to earth some of the sound energy that would otherwise have been dissipated in the upper atmosphere. The contrary temperature distribution, warm below and cool above, has the opposite effect. Sound which initially was moving horizontally is deflected upward and lost as shown in part (*c*) of Figure 176. Hence, hearing at a distance is bad under such circumstances. Similar conditions are produced by a steady breeze, a breeze "with the sound" bringing about the condition of part (*b*) of Figure 176 and one "against the sound" producing that of part (*c*) of Figure 176. In the case of poor hearing conditions, whether produced by temperature levels or by adverse wind, a moment's consideration will make it evident that there is an advantage in having the source of sound at a considerable altitude. This is the reason for the custom of placing bells, whistles, and sirens as high as possible.

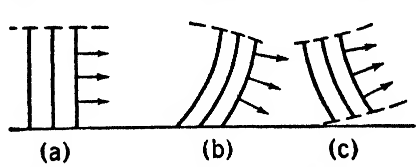


FIG. 176. THE CURVING OF A WAVE FRONT OF SOUND DUE TO VARIATION OF SPEED WITH ALTITUDE

In liquids, the velocity of sound also increases with rising temperature, whereas in solids it generally decreases. These changes of velocity of sound with temperature influence the pitches of musical instruments of all kinds. The pitch of a wind instrument rises with rising temperature, while that of reeds, bells, and tuning forks falls with rising temperature. In stringed instruments the pitch is affected, not only by changes in the velocity of sound in the strings with fluctuating temperature, but also by changes in tension due to expansion and contraction of the strings and their supports. It is impossible to make any general statement about the effect of temperature on pitch that will cover all kinds of stringed instruments.

Sound as a Type of Wave

The propagation of energy from place to place in the form of sound raises the question of the mechanism by which the transfer occurs. There are only two ways for this to take place. Either the vibrating source "kicks" a spray of air molecules in all directions, much as an intermittent lawn sprinkler sends its vibrant shower into the air, or the disturbance created by the vibrating source spreads through a relatively stationary air, as a water wave spreads from the point where a stone is dropped. It requires but little consideration to show that the first assumption is untenable. For while there is nothing positively absurd about air molecules, impelled from a sound source, reaching the ear of a listener, it is much harder to picture the analogous thing as happening in water, and impossible to account on this basis for the propagation of sound in solids.

The alternative picture is that of a series of waves emanating from the source of sound. This presents no difficulty in connection with gases and liquids. If the imagination balks at picturing waves in solids, it is only necessary to recall that the mere fact that all solids are elastic constitutes a sufficient basis for attributing to them the ability to transmit waves.

Naturally, sound waves in solids (and, in fact, in liquids and gases as well) are not at all identical with the water waves with which we are familiar on rivers, lakes, and oceans. Acquaintance with the latter, however, may furnish some concepts that will be useful in studying the former.

It is a matter of common observation that, though a wave moves steadily along the surface, the water itself does not partake of this motion.[1] A floating chip will betray the motion of the water itself. It maintains its position almost unchanged except that it bobs up and down while successive waves pass under it. Careful observation of such a chip will show its motion to be elliptical. Thus the motion of the water is entirely different from the motion of the wave passing over it.

[1] The reference here is to deep-water waves, not the surf waves most common along the shore line.

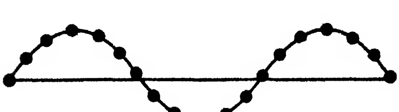


FIG. 177. A TRANSVERSE WAVE FORM

FIG. 178. A LONGITUDINAL WAVE FORM

Figure 177 conventionalizes and simplifies the concept of a wave. This shows successive stages in the progress of a wave horizontally over a row of particles which individually oscillate in a vertical direction. This is an example of a so-called *transverse* wave. The name originates in the fact that the oscillation of the particles constituting the wave is transverse to the motion of the wave itself, since the particles move vertically, whereas the wave moves horizontally. It is common to picture sound waves in this way, for reasons that will appear presently, though sound waves are not, in fact, transverse waves.

Figure 178 illustrates another kind, the *longitudinal* wave. The wave in this case consists of a compression followed by a rarefaction, instead of a crest followed by a trough. In spite of the absence of a sinuous outline, this is just as much a wave as was the preceding. The particles oscillate as they did before, but the direction of their motion is along the line followed by the wave instead of perpendicular to it and is, therefore, called "longitudinal." This picture is a representation of the principle of the sound wave.

The production and subsequent travel of an air wave in a pipe is represented in Figure 179. Sudden withdrawal of the plunger at the left creates a partial vacuum (rarefaction). Layers of air move in from the right, thus progressively shifting the rarefaction along to the right. Air also moves in from the left, thus producing a reversal of the motion of the layers that had previously entered the partial vacuum from the right. The resultant "piling-up" constitutes a condensation which follows the preceding rarefaction along the pipe toward the right. The process is repeated time after time until the energy originally imparted by the withdrawal of the plunger is dissipated and the waves cease. It will be noted that in a train of longitudinal waves, here represented as traveling to the right, the motion of the particles themselves is alternately to the left and to the right. Rarefactions are characterized by motion of the particles in a direction opposite to that of the progression of

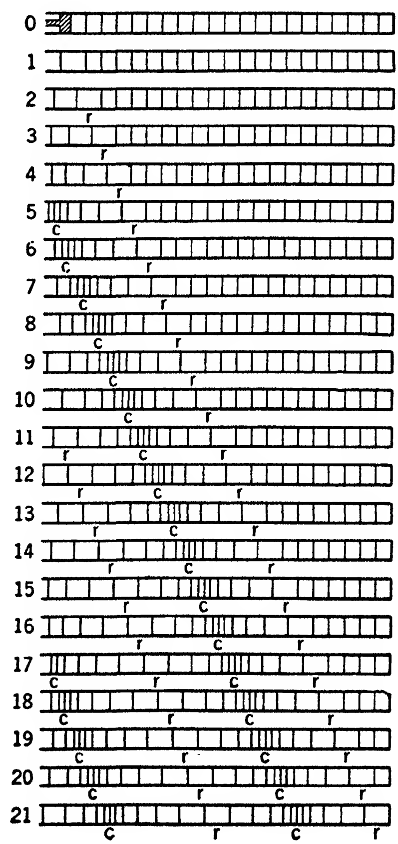


FIG. 179. THE PRODUCTION OF A LONGITUDINAL WAVE

the wave; and condensations, by motion of the particles in the same direction as the wave.

If sound is produced by a vibrating body in the open instead of being confined to a pipe as represented in Figure 179, the waves will not be confined to a particular direction of travel, but will spread out in spherical form, much as a wave on the surface of water spreads in a circle regardless of the direction of the disturbance producing it. It is a section of this kind of wave which upon entering the ear stimulates the sensation of sound. In such waves all the properties of sound, both musical and non-musical, reside.

Questions for Self-Examination

1. Describe some phenomena associated with the fact that sound requires an appreciable time to travel ordinary distances.
2. Does the speed of sound vary with (*a*) the medium, (*b*) pitch, (*c*) loudness, (*d*) temperature? Expand your answers, giving examples when possible.
3. Describe with the aid of a diagram the curving direction in which sound is sometimes propagated and account for it.
4. Describe the wave-properties of sound.
5. Distinguish between (*a*) motion of a wave and motion of the medium in which it travels, (*b*) transverse and longitudinal waves.

Problems on Chapter 21

1. Two observers operate a "speed trap." The first sounds a horn as the suspected motorist passes him. The second thereupon starts his stop-watch, stopping it when the motorist passes *him*. The speed limit is 40 miles an hour, a speed which the motorist is suspected of having attained. What is the percentage error? Is it favorable or prejudicial to the motorist? Take the speed of sound as 1133 ft/sec. 5.5 per cent.
2. The lowest pitch detectable as sound by the average ear consists of about 20 vibs/sec, and the highest of about 20,000. What is the wave-length of each? 17 m, 1.7 cms.
3. A handclap in front of a stadium is echoed as a musical tone. If the steps are 30 inches wide, what will be the frequency of the reflected sound? 220 vibs/sec.
4. How far away from a band playing at a tempo of 2 per second will marchers be exactly out of step with members of the band? 170 m.
5. The bevel on the bullet-hole in a glass plate was 10° with the surface of the plate. What was the speed of the bullet? Take the speed of sound in glass as 5000 m/sec. 880 m/sec.
6. The speed of sound in a certain metal is V meters per second. If one end of a pipe of that metal l meters long is struck, an observer hears two sounds: one from the wave which travels along the pipe, the other through the air. What interval t in seconds elapses between the two sounds?

Metal	V	l	t
steel	4990	100	.28
brass	3500	100	.27
tin	2500	100	.26
lead	1227	100	.22

7. Find the speed of sound v in meters per second in air at the temperature $t°$ C., taking the speed at 0° C. as 331.5 meters per second and the increase as .6075 m per sec per deg C.

	t	v		d	t		t	d
7.	− 40	310	8.	5	1.	9.	1	4.8
	− 20	320		20	2.1		2	19.
	20	340		40	3.		3	41.
	30	350		70	4.		4	70.

8. A stone is dropped into a well of depth d meters to the water surface. In how many seconds, t, will the splash be heard at the top?

9. A stone is dropped into a well; t seconds later the sound of the splash is heard. What is the depth d of the well in meters?

10. The density of a certain metal is d kilograms per cubic meter. Its Young's modulus is Y newtons per square meter. Calculate the velocity V of sound in it in meters per second.

Metal	d	Y	V
steel	7.830	19.6×10^{10}	5000
brass	8.400	10.3×10^{10}	3500
tin	8.000	5.0×10^{10}	2500
lead	11.340	1.7×10^{10}	1200

11. An organ pipe normally sounding middle C of frequency 261.63 is blown in and with another gas than air having a specific gravity d referred to air and γ as the ratio of its specific heats. Find the frequency n (γ for air = 1.405).

Gas	d	γ	n
hydrogen	.06952	1.407	990
methane	.5544	1.316	340
oxygen	1.105	1.398	240
chlorine	2.486	1.323	160

CHAPTER 22

The Acoustics of Rooms

The Beginnings of Architectural Acoustics

Ever since the human race started the construction of dwelling places, the acoustic properties of rooms have been left to chance. Provisions for heating were mandatory from the beginning, in all except tropical regions, as a mere matter of preservation of life. Control of lighting was a necessary condition of effective activity. Sanitation was presently found necessary to health, and more recently ventilation has received attention, particularly in public and semi-public buildings. It is only since the beginning of the twentieth century that substantial progress has been made in acoustic design. It is now possible for architects to meet acoustic specifications almost as dependably as they can meet specified conditions of heating, lighting, and ventilation. There is little excuse for defective acoustic conditions in any building erected since the beginning of the twentieth century.

As in all technical problems, when it becomes necessary to consider all the details involved in the application of acoustic design, there are complications. The ability to cope with these complications constitutes the technical equipment of the acoustician, a new type of engineer, whose services are now necessary in modern architectural design. But, also as in most technical problems, the essential principles of architectural acoustics are simple.

Nearly all of even the smaller subtopics of physics are steps in more or less continuous developments, the entire courses of which usually cover one or more centuries. The present subtopic is a notable exception to this rule. It sprang into being almost entirely *de novo* beginning in 1895 at the hands of W. C. Sabine of Harvard University. Very little had been done in this field prior to the time of Sabine's contributions. The principal work — and it was slight indeed — had been done by Joseph Henry of the Smithsonian Institution forty years before. Henry's principal contributions to physics were in the realm of electricity and will be discussed later.

When in 1895 the lecture room of a new building at Harvard was first put into use, audiences found it almost impossible to hear a speaker because of excessive reverberation. Professor Sabine was requested to undertake a study aimed at correcting the difficulty. He did his work so thoroughly that it established a new branch of the science of sound. Sabine's investigations were so completely and carefully carried out that subsequent workers have done but little in the way of extending the foundations of the subject. Subsequent additions have largely been confined to the superstructure.

Reverberation

Since excessive reverberation was recognized at the outset as the principal defect to be corrected, it was natural that reduction of reverberation should have been Sabine's first objective. What there was no reason to expect, however, was his discovery that reverberation — when clarified by a precise definition, which will be given presently — constitutes the cardinal principle of architectural acoustics. The necessity for a certain amount of reverberation, surprising though it may appear at first, may be seen as rather plausible, at least in the case of music. In music every successive chord is intimately related to those chords which precede and to those which are to follow. It is possible that the unity of a musical phrase is accentuated by the right amount of "sonic cement" in the form of reverberation as an aid in relating successive notes or chords. However that may be, musicians often complain of the "deadness" of a hall in which there is insufficient reverberation.

For speech, the advantage of reverberation is probably of a different kind. Inasmuch as the energy from a speaking voice is usually less than that involved in musical performances, reverberation unquestionably aids in bringing to the ear of an auditor a sufficient volume of sound. The degree of confusion produced by a certain amount of reverberation is apparently not sufficient to counteract the advantage of a greater volume of sound. However, as in the case of music, the advantage of the right amount of reverberation is an experimental fact, independent of any speculations that may be offered as to the reason therefor.

The full significance of reverberation in architectural acoustics cannot be developed until the term is more precisely defined. We are already sufficiently acquainted with it to realize that reverberation owes its origin chiefly to the reflection of sound. It should not be confused, however, with an *echo* which is also a phenomenon of reflection. The acoustician reserves the term echo for the case in which a short, sharp sound is distinctly repeated by reflection, either once from a single surface, or several times from two or more surfaces. In reverberation, on the other hand, there is no distinct repetition but rather a mass of sound, filling the whole room and incapable of analysis into distinct echoes. Reverberation is a consequence of multiple reflection, and persists as a continuous, though rapidly decaying, sound for a measurable time after the original sound has ceased. A reverberation measurement — and such measurements constitute a vital part of acoustic practice — consists of timing the interval

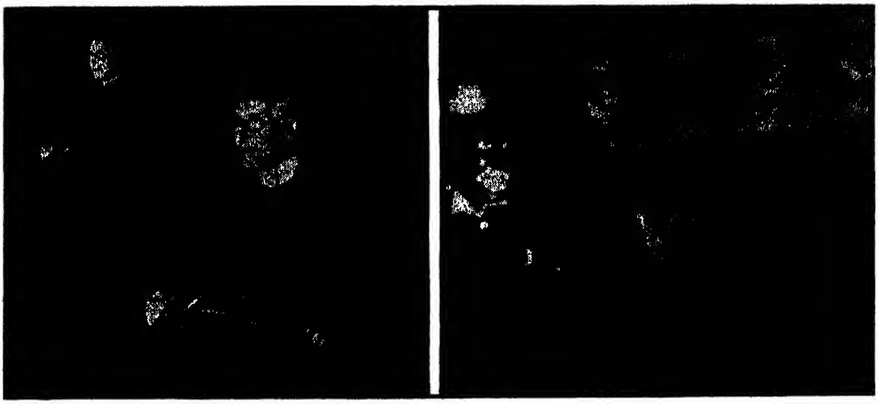


FIG. 180. A VISUAL ANALOGY ILLUSTRATING THE DISTINCTION BETWEEN ECHO AND REVERBERATION
(Courtesy of Johns-Manville Company.)

from the cessation of the original sound to the complete dying-out of the reverberation.

Time of Reverberation

It is now possible to establish a definition of the term "time of reverberation." As used heretofore, the concept has been too vague to be of much use. The reverberation of a loud sound would obviously persist longer than that of a weak one. In addition observers having different acuities of hearing would disagree as to the time when the reverberation had completely died out. These two sources of uncertainty are eliminated as follows.

Observation is made of the power of the source of a sound that is just barely audible to the experimenter in the room under test. This may be done practically by measuring the electrical energy fed into a loudspeaker. Then a tone is produced — in the same loudspeaker — that possesses exactly one million times the energy of the original barely audible tone. This is not excessively loud, for the ear can accommodate a very wide range of intensities. The time required for this to die away until it becomes inaudible to the same experimenter is then the *time of reverberation.* This is a time which can be shown to be substantially the same for all observers whose hearing is not seriously deficient. For if the experiment were repeated by a slightly deaf observer, the error that he would make by failing to hear the reverberation until it became as faint as the reverberation his predecessor was able to hear would be balanced by the fact that the slightly deaf observer had had initially to make the original tone correspondingly louder.

The time of reverberation is thus completely defined. Its importance lies in that fact already implied, that *it has been found to be the principal* *factor affecting the acoustic properties of rooms.* Not only is it an objective index of acoustic properties, but it can be controlled and, hence, the acoustic properties of a room can be adjusted to meet the requirements of good hearing conditions.

One would suspect that the time of reverberation might suffer a change if the speaker or performer should move to an entirely different part of the room or if the auditor should do likewise or if the pitch of the sound should change. Sabine found, however, that neither of the first two changes materially affects the reverberation time and that change of pitch, except for extremes, need be taken into account only under exceptional circumstances.

In ordinary small auditoriums (school- and classrooms and the like) the best time of reverberation is one second or less. With increasing size, the proper time of reverberation rises until in an auditorium having a volume of one million cubic feet the optimum time is two seconds. This is as large an auditorium as would commonly be encountered.

Reverberation and Sound Absorption

Sound is one form of energy and, when once released, will continue until it either escapes or is converted into some other form. Some of it escapes through open doors and windows. The rest of it, being to a greater or less degree absorbed at each successive reflection, must ultimately suffer complete absorption and conversion into heat. The rate at which this progresses will depend on the absorbing power of the surfaces and upon the frequency with which the sound waves encounter these surfaces. Thus, the softer and more porous the surfaces and the smaller the room, the shorter will be the time of reverberation. In a room of a given size, the time of reverberation can thus be governed by the introduction or removal of absorbing materials.

That there was a connection between reverberation and amount of absorbing material present in a room was not Sabine's discovery. It had been pointed out both by Henry (54:227) and even more explicitly just before Henry by a doctor in Boston who had done extensive experimentation with acoustical problems (66:6). But Sabine was the first to formulate mathematically the relation between time of reverberation and the amount of sound absorption provided by a given room. As in so many other instances, it was the new possibility of stating acoustical problems in numbers and, hence, of deducing quantitative answers that just before the turn of the twentieth century brought architectural acoustics into the fold of the sciences.

An Absorption Unit for Sound

Before a numerical relation can be stated between the time of reverberation of a room and the amount of sound absorption provided by the room, a unit for the latter entity must be defined. Sabine at first expressed the amount of sound absorption in terms of the number of seat-cushions brought into or taken out of the room in the course of his early experiments. Figure 181 is one of his graphs, showing times of reverberation as ordinates plotted against the numbers of seat-cushions in the room as abscissas. He soon recognized that a seat-cushion was scarcely an appropriate unit to use for absorbing power and cast about for something that could be more readily standardized. It occurred to him that an aperture, such as an open window, was the most complete "absorber" of sound that could very well be devised. The fact that it did not really absorb sound but only allowed it to escape from the room under test was not important. Its principal feature was that substantially none of the sound incident on an open window was reflected back into the room. Sabine introduced his "open-window unit" in these words (110:23):

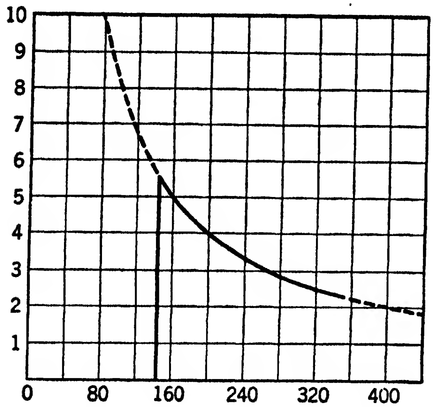


FIG. 181. ONE OF SABINE'S CURVES SHOWING THE RELATION OF TIME OF REVERBERATION TO THE AMOUNT OF ABSORBING MATERIAL PRESENT (110: 22.)

> For the purpose of the present investigation, it is wholly unnecessary to distinguish between the transformation of the energy of the sound into heat and its transmission into outside space. Both shall be called absorption. The former is the special accomplishment of cushions, the latter of open windows. It is obvious, however, that if both cushions and windows are to be classed as absorbents, the open window, because the more universally accessible and the more permanent, is the better unit. The cushions, on the other hand, are by far the more convenient in practice, for it is possible only on very rare occasions to work accurately with the windows open, not at all in summer on account of night-noises — the noise of crickets and other insects — and in the winter only when there is but the slightest wind; and further, but few rooms have sufficient window surface to produce the desired absorption. It is necessary, therefore, to work with cushions, but to express the results in open-window units.

Sabine accordingly adopted as the metric unit the sound absorption furnished by an aperture one square meter in area. The unit has never formally been given a name. Sabine referred to it as the "open-window unit." Now it is usually termed merely an *absorption unit.* It is scarcely surprising to learn, as Sabine did by careful experiment, that two square meters of aperture absorb sound at twice the rate that one square meter does, and other areas in proportion.

Absorption of Various Materials

It is, of course, impossible to find any material which is more effective than an aperture in absorption of sound, area for area. The surfaces ordinarily presented in completed and occupied structures absorb sound in varying degrees. They are rated by comparison to equal areas of aperture. The ratio of the rate of absorption of sound by a given material to that of an equal area of aperture is termed the *absorptivity* of the material in question. Absorptivity can have any value from substantially zero up to unity. Sabine found that his seat cushions had an absorptivity of .80. Commercial sound-absorbing materials seldom exceed this value even when designed explicitly for maximum absorption. The absorptivities of some materials commonly encountered in occupied buildings are listed below.

ABSORPTIVITIES OF COMMON BUILDING MATERIALS

Adapted from Official Bulletin, Acoustical Materials Association

Material	Frequency (cycles/sec) 125	500	2000
Brick, painted	.012	.017	.023
unpainted	.024	.030	049
Carpet, unlined	.09	.20	.27
lined	.11	.37	.27
Drapes, light	.04	.11	.30
heavy	.10	.50	.82
Floors, concrete or stone	.01	.015	.02
wood	.05	.03	.03
linoleum, cork, etc.		.03–.08	
Glass	.035	.027	.02
Plaster, smooth	.015	.025	.02
rough	.04	.06	.05
Wood panelling	.08	.06	.06

ABSORPTION OF SEATS AND AUDIENCE

in absorption units

Audience, per person	.10–.20	.28–.40	.33–.54
Chairs, metal or wood, each	.014	.018	.010
Pew cushions, per person	.070–.10	.135–.175	.13–.16
Pews, wood, per person		.037	
Seats, auditorium			
wood veneer, each		.023	
upholstered, leatherette, each		.15	
upholstered, plush, each		.24–.28	

From the definition of absorptivity it will be evident that the number of absorbing units a presented by s square meters of material having an absorptivity α is

$$a = \alpha s. \tag{1}$$

Thus the number of absorbing units in an empty room $4 \times 4 \times 3$ meters with smooth plaster walls and ceiling, wood floor, 8 square meters of glass window, and a wooden door 1×2 meters would be computed as follows.

ceiling	$4 \times 4 \times .025$	=	.40
walls	$[(16 \times 3) - 10] \times .025$	=	.95
floor	$4 \times 4 \times .03$	=	.48
windows	$8 \times .027$	=	.22
door	$2 \times .03$	=	.06
	Total absorption units		2.11

The Reverberation Equation

During the early stages of Sabine's work the only way to arrive at a numerical index of the acoustic acceptability of a room was to observe its time of reverberation. Any departure from the optimum value corresponding to its size gave a measure of the acoustic defect in the room. Actual measurements of time of reverberation proved, however, to be laborious and expensive undertakings, and Sabine undertook to find a less cumbersome way by which it might be determined. A cue to his solution of the problem is furnished by the graph of Figure 181. The mere fact that the time of reverberation of the room in question was affected by the number of absorbing units present indicated the existence of a relation between the two. This in itself was a major discovery. For such a relation might make it possible to calculate the time of reverberation if the number of absorbing units present in the room were known. The foregoing example indicates that a determination of the number of absorbing units in a room is a simple undertaking. Thus a way to avoid the difficulties of an experimental determination of time of reverberation is presented.

The form of the curve of Figure 181 suggests the type of relation being sought. It appears to be an equilateral hyperbola such as was encountered when Boyle's law was graphically represented (page 71). Hence, if T is time of reverberation and A the total number of absorbing units in the room,

$$AT = \text{a constant} \tag{2}$$

would be the form of the desired relation. Sabine tested this relation between T and A on rooms of many different sizes, the largest having 150 times the volume of the smallest, and found it to hold with them all. The value of the constant, however, he found to be proportional to the volume of the room. This too is analogous to Boyle's law, in which the value of the constant is proportional to the mass of gas involved. In this case the constant of proportionality was found empirically to be .164. Thus *Sabine's law* may be stated

$$AT = .164\ V. \qquad (3)$$

The Utility of Sabine's Law

Sabine's law opens up enormous possibilities. It is in fact the very heart of the science of architectural acoustics. It enables the architect to deduce immediately the time of reverberation of a room if its volume and the number of absorption units present are known. Thus, in the example on page 242, V has the value 48 cubic meters, and A has the value 2.11. The time of reverberation would, therefore, be found to be nearly 4 seconds. This would clearly be an unacceptable room acoustically.

In this way an architect can tell not only what the acoustics of existing rooms are, but also what the acoustics of rooms still in the blueprint-and-specification stage will be. More important still, it enables him to determine how much acoustic material must be provided to render a defective room acceptable. In the foregoing case, for example, the total number of absorption units necessary to make the time of reverberation of the room 1 second would be 7.87. Since 2.11 units are already present, 5.76 additional units must be provided. If the furniture and the absorptive power of the occupants do not furnish the required number of units, the extra may be provided in the form of acoustic surfacing, usually mounted in requisite quantity on the ceiling of the room. The area required is immediately determinable if the absorptivity of the surfacing is known.

Additional Acoustic Factors

While reverberation is the most important, other considerations bear on the problem of architectural acoustics. In a general way it may be stated that the acoustic properties of a room depend on three factors; shape, size, and building materials By a study of reverberation, the last two factors have been disposed of, and only the first is left.

Two main principles govern shape; avoidance of disproportionate relative dimensions and avoidance or discriminating arrangement of curved interior surfaces. It is not necessary to point out the difficulties inherent in extremely long and narrow rooms, extremely low rooms, or, in fact, almost any extreme of relative dimension. The difficulty with curved surfaces is their tendency to focus sound at certain points, thereby producing non-uniform distributions of sound that are often serious. If curved ceilings are used at all, the centers of curvature should be either far below the floor or far above it, so that the focal points will not lie anywhere near the level of the audience.

Figure 182 (110:233) delineates the defective distribution of intensity in an auditorium having a curved ceiling. The sound is concentrated in the regions numbered 10, 11, 12 at the expense of the regions 0, 1, 2. The latter would be called *dead spots*.

Auditoriums which suffer from faulty design in this respect may be treated either by putting recessed panels on the curved ceiling in such a way as to break up the regularity of the pattern in the reflected sound or by crowding absorbing material on the curved surfaces so as to reduce the amount of reflection.

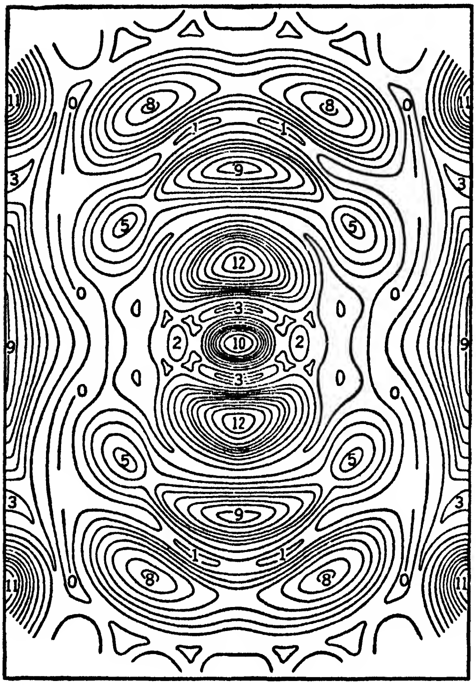


FIG. 182. DISTRIBUTION OF INTENSITY AT HEAD LEVEL IN A ROOM WITH BARREL-SHAPED CEILING WITH THE CENTER OF CURVATURE AT FLOOR LEVEL

(182:233.)

Dead spots may also come about from another cause, not connected with arched ceilings or curved walls. It may result from reflection by perfectly plane surfaces. The phenomenon called *interference* between the direct and reflected sounds is capable of producing it. This phenomenon will be treated more fully in another connection. It was discovered early in the nineteenth century by Thomas Young, who studied it principally in connection with light. In one of his papers he described it as it might be observed with water waves (quoted 133: 32).

> Suppose a number of equal waves of water to move upon the surface of a stagnant lake with a certain velocity, and to enter a narrow channel leading out of the lake. Suppose then another similar cause to have excited another equal series of waves which arrive at the same channel with the same velocity and at the same time as the first. Neither series of waves will destroy the other, but their effects will be combined; if they enter the channel in such a manner that the elevations of the one series coincide with those of the other, they must together produce a series of greater joint elevations; but if the elevations of one series are so situated as to correspond to the depressions of the other, they must exactly fill up those depressions, and the surface of the water must remain smooth.

Interference as a General Phenomenon

The effects that Young describes are represented graphically in Figure 183. The additive effect is shown first. The two component waves, really coinciding (dotted curves), are slightly displaced from each other so that each may be seen. The second part shows the mutual cancellation produced when the crests of one train lie on the troughs of the other and vice versa. The resultant of the two waves is the straight horizontal line, representing the condition in which Young says that "the surface of the water must remain smooth." If the waves do not combine in either of the ways described above, the resultant will be a train of waves with crests of a lower height than that of the first case. The height of the crests may, of course, be anything from zero up to the maximum, depending on the relative displacement of the component trains. The case illustrated is that of a quarter wave-length displacement, half way between the two extreme cases illustrated previously.

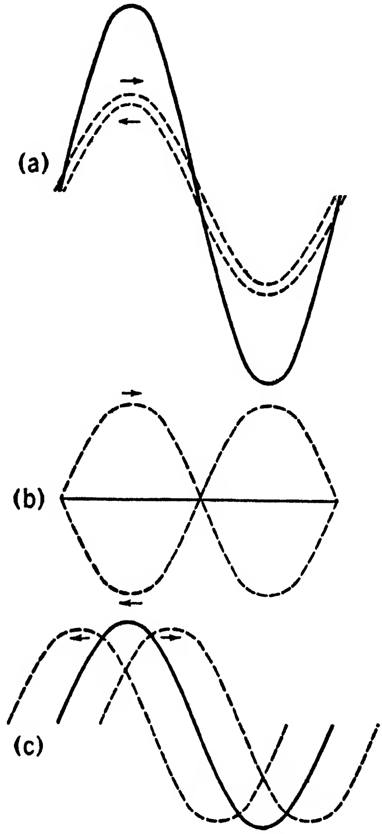


FIG. 183. ASPECTS OF INTERFERENCE BETWEEN TWO IDENTICAL WAVES TRAVELING IN THE SAME DIRECTION

The application to acoustic dead spots is clear. Since sound is composed of waves, two identical sounds may combine in such a way as to produce silence. This can occur by interference between a direct sound and a reflected sound from the same source. When it does so at a particular region, a dead spot results. The cure is the same as that for dead spots produced by curved surfaces. Either break up the regularity of the reflecting surface or cover it with absorptive material.

Questions for Self-Examination

1. On what factors do the acoustic properties of a room depend?
2. Given an acoustically defective auditorium, tell what you would do to diagnose and correct it.
3. Draw Sabine's reverberation curve, state his reverberation equation, and interpret them.
4. Define and discuss the unit of sound absorption.
5. Describe the general phenomenon of interference.

Problems on Chapter 22

1. Calculate the number of absorption units, A, in a vacant classroom of length l, width w, and height h meters. The walls and ceiling are of smooth plaster, the floor of wood, and there are W square meters of window and glass door. The room is equipped with n wood veneer seats. (Make the calculations for frequency of 500 and use the mean of spread values.)

l	w	h	W	n	A
10	8	4	12	50	9.2
15	12	4	20	125	18.
20	15	5	25	225	30.
30	25	6	30	500	69.

2. Calculate the time of reverberation, T_1, of the classroom of problem 1 when empty, T_2, with an audience of half the seating capacity, and T_3 with a full audience.

n	T_1	T_2	T_3
50	5.7	2.9	2.0
125	6.4	3.0	1.9
225	8.1	3.6	2.3
500	11.	4.8	3.1

3. Calculate the number of additional absorption units, U_1, that must be supplied if the time of reverberation of the room of problem 1 is to be reduced to 1 second for an empty room, U_2 for a half audience, and U_3 for a full audience.

n	U_1	U_2	U_3
50	43	35	26
125	100	78	57
225	220	180	140
500	670	580	500

CHAPTER 23

The Pitch of Musical Tone

Pitched and Unpitched Sound

Up to this point attention has been confined to properties of sound not necessarily associated with music. Musical tone is, however, precisely the aspect of sound which has received the most study through the ages and which, quite naturally, has a correspondingly large place in the science of sound.

A clear line of distinction between musical and non-musical sound cannot be drawn. The distinction depends too much on the musical ear and education of the listener. Less controversial would be the distinction between sound possessing identifiable pitch (*pitched sound*) and that which does not (*unpitched sound*). Though it would not be hard to find sounds which would be difficult to classify as between the two, the distinction seems to be useful and will be invoked here. While not all pitched sounds can be termed musical, music is the field in which the idea is most useful and in which the greatest number of illustrations are to be found.

Pitch and Frequency

Experiment shows that if the frequency of vibration of a source of tone increases, the pitch "rises" as the musician puts it. From many observations of this kind we have learned to associate a certain pitch with a corresponding frequency, regardless of the source. This association is not always justified. As long as the wave is "simple," the corresponding tone being "pure,"[1] it is correct. But if the wave is at all complicated in shape, inference of frequency based on observation of pitch is likely to be inadequate if not actually misleading. Stated in different terms, a knowledge of frequency gives reliable information about pitch; but observation of pitch does not necessarily give any information about frequency. If we confine attention to pure tones, however, the correspondence between frequency and pitch is complete. For the present this limitation will be accepted.

Some representative relations between commonly encountered pitches

[1] The terms "simple" and "pure" will be adequately defined later in the discussion of tone quality. The reader can gain from the context a sufficient comprehension of their meaning for present purposes.

and corresponding frequencies are presented in the following table. Many of the frequencies are averages or approximations.

	Frequency	Wave-length (feet) in air
"Middle C" (C_8)[1]	261.6	4.37
Highest note from piano (C_{12})	4,186.	.273
Lowest note from piano (A_4)	27.5	41.5
Range of orchestral music (C_5–C_{12}) approx.	30–4,000	40–.25
Representative masculine speaking voice (C_7) approx.	120.	7.5
Representative feminine speaking voice (C_8) approx.	250.	4.5
Highest audible frequency (E_{14}) approx.	20,000.	.057
Lowest audible frequency (E_4) approx.	20.	56.5

Often pitch is associated with wave-length instead of frequency. This is justifiable if attention be confined to one particular medium. Since the sound reaching our ears almost invariably comes through the air, the restriction to a single medium is a natural one. The wave-lengths in air of the representative pitches of the table above are listed in the final column. On entering a different medium, however, the wave-length suffers a change. If, for example, a sound of wave-length 25 feet in air should enter hydrogen, the increase in velocity would stretch out the wave-length to about 95 feet. In water, the wave-length would become 116 feet, and in steel 361 feet. In none of these cases would either the pitch or the frequency be affected. While it is often necessary to know the wave-length of a tone, it is more natural to associate pitch with frequency than with wave-length.

Extremes of Audible Frequency

The last two lines in the above table indicate that a musical sound, to be audible as such, must possess a frequency greater than 20 vibrations per second and less than 20,000. These limits are only approximate. There is wide variation between individuals at both limits. At frequencies below the lower limit, the sound is that of the separate impulses, individually distinguishable. As the frequency increases, these merge and produce the first traces of an extremely low tone. The establishment of the exact frequency at which the transition occurs is attended by considerable uncertainty. The experimental basis for the upper limit is rather better, partly because it is easier to produce pure high tones than it is low ones, and partly because the ear can more readily discriminate between the presence and absence of a tone at the upper frequency limit.

Not only is there variation between individuals at the upper limit, but the upper limit will be found different at different times for the same individual. In addition to erratic variations due to colds-in-the-head, general state of health, and so forth, there is a general trend downward with increasing age. The following table shows this trend for a group of sub-

[1] Subscripts indicating octaves in which designated pitches occur in this text follow the practice recommended by the Acoustical Society Committee on Standardization. See *Journal of the Acoustical Society*, *2*, 318 (1931).

jects tested at two times separated by somewhat more than twenty years. All were trained observers, being either musicians or professional men familiar with music.

Subject	Age at first observation	Upper frequency limit 1907	1929
E. A. Miller	41	19,500	16,000
F. J. Lehman	40	26,000	17,000
C. H. St. John	52	17,000	13,500
H. W. Matlack	31	25,000	21,000
E. J. Moore	34	18,000	19,500

Observations on very young children have shown the upper frequency limit to be much higher than the normal for later years; in some cases it approaches 30,000.

Even though these extremely high tones can be heard, discrimination between different pitches in this range is largely or entirely lacking. Changes in pitch which are many times the smallest observable change in the ordinary musical range are entirely undetected. The auditor can quite dependably decide on the presence or absence of pitched sound in this high-frequency region, but finds it almost impossible to say whether one pitch is higher or lower than another.

Musical Intervals and the Tempered Scale

Musical themes customarily proceed by definite steps or *intervals*. A continuous glide from one pitch to another through all the intervening pitches has not yet become good form in the more conservative types of musical performance. The system of intervals in common use has varied in successive musical eras. By a long, unconscious process of trial and error, presumably covering a large part of the development of the human race, a series of intervals has been developed and now predominates, at least in the western world. Any such series of intervals is termed a *scale*, and the particular series now in vogue is called the *tempered scale*, first brought to popular notice by J. S. Bach (1685–1750).

All scales consist of subdivisions of the basic interval called the *octave*. The octave consists of two tones having a frequency ratio of 2:1. It is called the octave because, during the era when modern musical terminology was crystallizing, the scale then in vogue consisted of eight subdivisions or *tones*. The tempered scale now in use, however, provides twelve tones to the octave instead of eight. The ratio of the two frequencies constituting each of the twelve tones [1] is the same. The higher frequency is always 1.0594 times the lower. This is the number which when multiplied by

[1] Common usage sanctions several quite different meanings for the word *tone*. One is as above, a musical interval expressible by the ratio of two frequencies. Another involves the idea of a single pitch or frequency, as when it is said that one "tone" is higher than another. The situation is complicated by still other uses of the term. This is regrettable. In this treatment, however, the context can be relied upon to clarify any doubt about the sense in which the word is being used.

itself twelve times gives the final product 2, the frequency ratio of two pitches constituting an octave.

Standard Pitch

While it is possible to construct a scale starting with any frequency whatever, the musical world has belatedly recognized the desirability of standardization. Recognized standards have existed only for two generations and been respected through common acceptance only within the last ten years. The world's standard frequency is now taken as 440 vibrations per second. It is the pitch commonly called "violin *A*." It is considered important enough so that it is being broadcasted by short-wave radio continuously twenty-four hours a day from the United States Bureau of Standards. Any player or orchestra leader who is properly equipped can be sure that his performance is based on the world's standard of frequency.

The Principle of Tonality and the Scale in C

One of the chief characteristics of modern music is the existence of a predetermined pitch, from which a melody starts, to which it repeatedly returns during its progress, and upon which it almost always ends. This pitch is usually determined by the frequency range of the composition and of the instrument that is to play it. This basic pitch is termed the *tonic*. The demand for a tonic on the part of the musically educated ear constitutes the *principle of tonality*. The collection of sharps or flats at the beginning of a composition tells what pitch is to be regarded for the time as the tonic.

The standard frequency *440* represents the tonic of only one scale, that based on *A*. Musical convention, however, has set apart another scale, based on a tonic called *C* as in some sense a unique point of departure for all scales. *C* has the frequency of *261.63*, or, of course, any multiple or submultiple of this by the factor *2*. The pitch represented by the frequency 261.63 is commonly called *middle C*. It lies near the middle of the piano keyboard and represents approximately the upper pitch limit of the male voice and the lower pitch limit of the female voice.

Fig. 184 shows a section of an organ or piano manual covering one octave beginning and ending with *C*. The names of the white digitals are shown in terms of the usual letters. If the *A* in this octave is that of the standard frequency 440, then the *C* below it is middle *C*, of frequency 261.63. Thirteen digitals are shown, as required to provide twelve tones. Each of the five black digitals lies between two white ones and is designated by either of its white neighbors. The one to the left, for example, lies between *C* and *D*. Its frequency would be higher than *C* and lower than *D*, in either

FIG. 184. SECTION OF ORGAN OR PIANO MANUAL COVERING ONE OCTAVE

case by the factor 1.0594. It is called either *C sharp* (*C*♯) or *D flat* (*D*♭). This use of two names for the same pitch is quite inexcusable in the twentieth century. It is a vestige of an archaic musical notation which came into being before the tempered scale was devised, at a time when *C*♯ and *D*♭ were not of the same pitch.

The frequencies of all the pitches appearing in this octave are shown in the appended table, beginning with that of middle *C*. Those of pitches above or below this range would be given by multiplying or dividing the corresponding frequency given in the table by 2 or the appropriate multiple of 2.

C	261.6	*G*	392.0
C♯ or *D*♭	277.2	*G*♯ or *A*♭	415.3
D	293.7	*A*	440.0
D♯ or *E*♭	311.2	*A*♯ or *B*♭	466.2
E	329.7	*B*	493.9
F	349.3	*C*	523.3
F♯ or *G*♭	370.1		

Though the above series represents the *de jure* frequencies of the tempered scale, the *de facto* series is something quite different though rather indeterminate. Though organ and piano tuners do not all follow the same system of tuning, they all agree in following some rule of thumb which gives a more or less distant approximation to the series of frequencies to which instruments are supposed to be tuned. In view of the tolerance of even the musically trained ear, perhaps the extra exertion that would be involved in really correct tuning is scarcely worth making.

Effect of Velocity of Source or Observer

Every traveler is familiar with the apparent drop in pitch of the gong at a railroad crossing when heard by a passenger on the train. If the passenger could know the normal pitch of the gong, he could realize that the pitch seemed higher than normal while the train was approaching, and lower while it was receding.[1] The reason is simply that during approach the observer is moving *toward* the source of sound at the same time that the sound is coming *from* the source. The effective speed of the sound as it reaches the observer is, therefore, the sum of the speed of sound and the speed of the train. Naturally, a larger number of waves reaches the observer every second, that is, the frequency is increased, and, hence, the pitch is higher. Just the opposite condition holds while the train recedes, the pitch being lowered thereby. The effective speed of the sound with reference to the observer is in this case the difference of the two speeds. In both these cases, the phenomenon is due to the apparent increase or decrease of the velocity of sound relative to the moving observer.

[1] Note the way this is stated. There is a persistent misapprehension that pitch *is rising* during approach and *is falling* during recession.

The apparent frequency of the sound will be related to the actual frequency in the same proportion as the apparent speed of the sound to the real. Hence, if the apparent frequency is n', the real frequency n, the speed of the observer v, and the actual speed of sound V, then

$$\frac{n'}{n} = \frac{V \pm v}{V}. \tag{1}$$

The positive sign would, of course, apply to approach; the negative, to recession.

The other case — that of a fixed observer and a moving source — is similar in its effect, but somewhat different in principle. This case, too, is familiar. The sudden drop in the apparent pitch of the horn of an automobile, the whistle of a train, or the gong of a fire-truck as any of these pass a stationary observer, is common experience. In this case, however, the sound reaches the observer with a speed which is both actually and apparently the normal speed of sound. The speed of sound through air is not affected by the motion of the source. What really happens in this case is illustrated in Figure 185. A moving source "follows up" its own sound wave; hence, the distance between the source and a given wave is less at the end of one second than it would have been if the source had been stationary. Consequently the waves are crowded up ahead of the moving source, and an observer receives a larger number of waves per second than he would if the source had been stationary, and, hence, the pitch is higher than normal. The contrary condition holds *behind* the moving source, with a consequent lowering of the pitch.

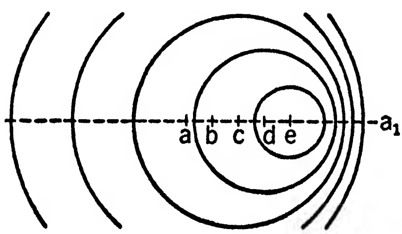


FIG. 185. THE WAVE PATTERN EMANATING FROM A MOVING SOURCE

In this case it is the wave-lengths rather than the frequencies that will be in proportion to the speeds of sound, apparent and real, as noted by the moving source. But since wave-lengths are inversely proportional to frequencies, the following equation expresses the case, u being the speed of the source:

$$\frac{n'}{n} = \frac{V}{V \mp u}. \tag{2}$$

The inversion of the sign in the denominator indicates that the negative sign would this time apply to approach; the positive, to recession.

An interesting feature of the foregoing is the difference in magnitude between the two cases. This may be made evident by considering the extreme case when the speed of the moving element (whether observer or source) becomes equal to the speed of sound. Also, the foregoing discussion is limited to the case of direct approach or recession. The case of oblique encounter is commonly observed in the passage of an airplane. The lower it is flying as it passes overhead, the more abrupt the passage from substantially direct approach to direct recession and the smaller the transition time when oblique encounter is involved.

These phenomena go by the name of the *Doppler effect* after Christian Doppler (1803–53), who first worked out the theory in 1842 (95: no. 161). A famous experimental test of equations (1) and (2) was carried out in Holland in 1845.[1] The use of a locomotive and a flat car was secured for two days. Trumpeters furnished the sound, and musically trained observers took the data. The two groups were respectively stationed on the ground and on the flat car and then exchanged positions. The changes in pitch were estimated by ear, there being no means of measurement of actual frequencies at that time. Within the limitations of the method employed, the equations were found to hold. One wonders what the stolid Dutch burghers living alongside the railroad must have thought of the performance.

Though the Doppler effect was originally studied as a phenomenon of sound, Doppler foresaw that it could be utilized in the realm of light. Change in wave-length of the light from approaching or receding stars could be made to give information on the speed of the star relative to the observer. This has actually been done and has furnished an indispensable source of astronomical information.

Questions for Self-Examination

1. Tell about the relation between pitch and frequency. Include answers to these questions: (*a*) does a given frequency necessarily determine the pitch? (*b*) does a given pitch necessarily determine the frequency?
2. State the relations between wave-length, frequency, and pitch for representative tones (high, average, and low) in air and other media.
3. Describe the even tempered scale and tell how it came into being.
4. State the principle of tonality and describe the common musical scale in *C*.
5. Describe the effect of velocity of source and observer on the pitch of a musical tone. What would be the effect if the speed of source or observer respectively should become that of sound?

Problems on Chapter 23

1. Two similar wires of the same length are stretched — one by a weight of 4 kilograms and the other by a weight of 9. What is the musical interval between the notes which they produce? See page 129 for the laws of vibrating strings and the table of intervals at the bottom of page 269 to identify the musical interval. A fifth.
2. A stretched string 3 feet long gives the pitch *C* when vibrating. What note will be given by a string one foot long stretched by the same weight and made of the same material but of one quarter the cross-sectional area? *G* $2\frac{1}{2}$ octaves above.

[1] *Poggendorff's Annalen*, *66*, 321 (1845).

3. A vibrating string is found to give C when stretched by a weight of 6 kilograms. What weight must be added to produce E? G? 3.4 kg. 7.5 kg.

4. An observer listens to a whistle sounded on a locomotive which is approaching him. The pitch appears to be C, but after the locomotive has passed the pitch is A. What was the speed? 94 ft/sec.

5. A tuning fork making 200 vibrations per second is moved toward a wall with a speed of 4 meters per second. How many beats per second will be heard between the direct and the reflected sounds? 4.8

6. What is the length λ in meters of a wave constituting a tone whose pitch is n vibrations per second? Take 331 meters per second as the speed of sound in air.

	n	λ		n	v	N		n	v	N
6.	20,000	.017	7.	500	10	520	8.	500	10	520
	5000	.066		500	20	530		500	20	530
	250	1.3		500	30	550		500	30	550
	20	17.		500	40	560		500	40	570

7. A stationary locomotive emits a whistle whose frequency is n vibrations per second. What is the apparent frequency N as heard by an observer in a train which is approaching at v meters per second? Take 340 m/sec as the speed of sound.

8. A locomotive moving with a speed of v meters per second emits a whistle of frequency n vibrations per second. What is the apparent pitch N as heard by a stationary observer whom the locomotive is approaching?

CHAPTER 24

The Intensity of Sound

Loudness and Intensity

The application of the scientific method to any field of knowledge involves, among other things, the redefining of old terms in a way which permits of the application of precise measurement. In this way, as shown in the preceding chapter, physics has redefined pitch, an inexact term, into frequency, a concept which can be measured with almost limitless precision. There is even more occasion for a similar substitution of concepts in the case of loudness. For while pitch could be determined with a certain degree of dependability before the idea of frequency was introduced, there is inherent and inescapable difficulty in specifying the magnitude of one's sensation of loudness. The physicist, therefore, simply disregards loudness, which he cannot measure, and as a substitute invokes another concept which he *can* measure. This new concept is not necessarily a faithful representation of loudness, but in so far as loudness can be specified seems to be a fair measure of it.

The idea that a spreading sound involves a corresponding spread of energy has already been introduced. This is involved in the concept of intensity of sound. It is at least plausible to assume that there is some connection, doubtless a close connection indeed, between the rate at which energy in the form of sound flows into the ear of a listener and the strength of the resulting sensation which is termed loudness. But rate of flow of energy has already been identified as power. Hence, the unit of sound *intensity* will simply be a unit of power. But since the energy content of ordinary sounds is so small, the ordinary unit of power, the watt, is too large to be usable. A submultiple one millionth the size, the microwatt, is commonly used. Accordingly, intensity of sound in a given region may be defined as *the power in microwatts passing through every square meter normal to the wave-front in the region in question.* The measurement of intensity of sound becomes thus simply a matter of the measurement of a form of power. This makes it possible to adapt to sound a technique of measurement already highly developed in other fields. Little attention will be given here to the details of this adaptation, however, because the methods of measurement are extremely varied and technical. They are chiefly electrical.

It will be informative to observe the power content of a few common sources of sound. The flow of power in the form of sound in an ordinary conversation is about 10 microwatts. Only a tiny fraction of this reaches the ear of any one listener, of course. For public speech, the power content of the sound averages about 20 times this amount and for some speakers occasionally reaches 200 times this amount. For a loudly bowed violin the power is about 60 microwatts; for certain types of organ pipe, from 140 to 3200. A sound that is barely audible carries about one millionth (10^{-6}) microwatt per square meter. One that is as loud as the ear can tolerate carries about one million microwatts. Thus the ear is an instrument of wide range. Not only can it hear over a frequency range of a thousand (20–20,000), but it is found to respond to an intensity range of something like a million-million fold. In addition, it can distinguish delicate shades of musical quality over much of these frequency and intensity ranges. The ear is by far the most versatile of our sense organs.

Limits of Audible Intensity

The lower limit of audible intensity, technically termed the *threshold of hearing,* has already been described as consisting of a power flow of 10^{-6} microwatts per square meter. This statement is somewhat arbitrary and will be clarified shortly. But, first, the existence of an upper limit of audible intensity should be noted. This seems curious until the concept is clarified. It seems to imply what is clearly absurd, that as a sound grows louder it finally reaches an intensity level so great that it can no longer be heard. This would be a misapprehension, though perhaps a natural one. When the intensity of a sound reaches one million (10^6) microwatts (1 watt) per square meter, the ear, in addition to hearing the sound, begins to experience a tickling sensation. If sound of this intensity were made incident on the skin it could actually be felt. Any considerable increase beyond this causes a sensation of pain in the ear. The intensity at which the tickling sensation begins is commonly taken as the upper limit of audible intensity, better described by its technical name, the *threshold of feeling.*

A Unit for Relative Sound Levels

Though sound intensities were stated in microwatts in the preceding section, in most cases the significant point was not the actual power content of a sound in mechanical units, but the *ratio* of those intensities. Almost always the question is, How much louder is one sound than another? While in answer to this question, the intensity of each sound could be stated in microwatts per square meter and the ratio computed, common usage now sanctions a unit of relative sound levels, in addition to the mechanical unit of power content of sound.

Two sounds are said to differ in sound level by one unit when their intensities have the ratio *10.* This unit is called the *bel* in honor of Alexander Graham Bell, the inventor of the electromagnetic type of telephone transmitter. Thus, the clatter in a fairly quiet restaurant during lunch hour has an intensity of about one tenth microwatt per square meter. This is about ten times as great as the intensity of the average sounds in a quiet home. The sound level is then said to be one bel higher in the restaurant than in the home. Similarly the sounds of a busy department store have ten times the intensity of those in an average restaurant. The sound level in the department store is then one bel higher than in the restaurant and two bels higher than in a quiet home. The sounds in a quiet garden, consisting mainly of the rustling of leaves in a gentle breeze, have only one per cent of the intensity of the average sounds in a quiet home. Hence, in such a garden the sound level is two bels lower than in the home and four bels lower than in a department store. Readers with a mathematical background will have observed that the relative sound level in bels is simply the logarithm of the ratio of relative intensities.

The Decibel as a Derived Unit

The sound levels described in the preceding paragraph differ by rather large increments. Noises casually encountered would not always differ in intensity by as large a factor as ten or more. The comparison of ordinary sound levels would, therefore, usually involve fractions of a bel. Since this is somewhat awkward, a smaller unit is usually used, the first decimal subdivision of the *bel*, called the *decibel* (often abbreviated to *db*). One bel is then 10 decibels. The sounds of the five sources thus far described could then be summarized in the different units as follows:

Sound	Intensity in microwatts per square meter	Sound level in bels	Sound level in decibels
Threshold of hearing	.000,001	0	0
Quiet garden	.000,1	2	20
Average home	.01	4	40
Average restaurant	.1	5	50
Busy store	1.	6	60

A somewhat more extensive table of sound levels follows, stated only in terms of decibels above the threshold of hearing.

Sound	Level, db	Sound	Level, db
Threshold of hearing	0	Street noise, large city	75
Whisper	15	Truck, unmuffled	80
Rustle of leaves in gentle breeze	20	Noisy factory	85
Purring cat	25	Newspaper press room	90
Turning page of newspaper	30	Noisiest spot at Niagara Falls	95
Quiet home or private office	40	Inside subway car	100
Average restaurant	50	Loud thunder	110
Noisy office or store	60	Threshold of feeling	120
Average radio	70		

The Audiogram

The lower and upper limits of audibility have been stated to lie respectively at 10^{-6} and 10^{6} microwatts per square meter (0 and 120 db). These figures are rather arbitrary, not only because different individuals vary in their ability to hear, but also because even for a given individual, the limits of audibility depend upon the pitch of the sound. Ordinary noises consist of a conglomeration of pitches, and the lower limits of audibility in such a case will be for those frequencies in the conglomeration to which the ear is most sensitive. The frequency band to which the normal ear is most sensitive has been found to lie between 2000 and 4000 vibrations per second. For both higher and lower pitches an intensity greater than 10^{-6} microwatts per square meter is required for a sound to be audible. The variation of the lower limit of audibility is indicated by the lower limit of the enclosed area of Figure 186, called an *audiogram.*

The audiogram will repay considerable study. On it, sound levels are represented as ordinates against frequencies as abscissas. The "area of hearing" extends horizontally from a frequency of *20* to a frequency of *20,000* as indicated in the preceding chapter. It extends vertically from the threshold of hearing (zero decibels) to the threshold of feeling (about 120 decibels). But the horizontal frequency limits are functions of the intensity, and the vertical intensity limits are functions of the frequency.

To see how intensity affects the upper and lower audible frequency limits, note that the frequency limits can only be realized when the sounds are loud. A horizontal line cuts the enclosed area at 20 and 20,000 only if it is well up on diagram. If it is lower, say, at 20 decibels, the lowest pitch which can be sensed is a frequency of about 250, or middle *C*. The highest is near 10,000 vibrations per second instead of 20,000. Another interesting fact may be illustrated by following the horizontal 20-decibel

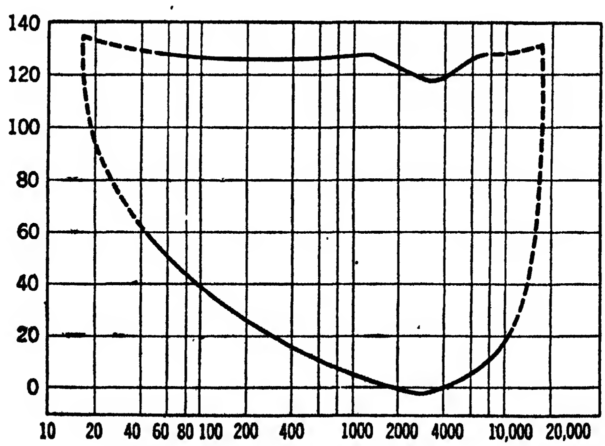


FIG. 186. THE AUDIOGRAM OF A NORMAL EAR

line across the diagram. The left portion of the line is in a region which corresponds to inaudibility. Raising the pitch while maintaining the sound level unchanged corresponds to moving to the right on this line. At a frequency of 250, the sound becomes audible and continues to become more so as the pitch-rise continues until a frequency of about 3000 is reached, whereupon the tone becomes dimmer, finally becoming inaudible again when a frequency of a little over 10,000 is reached. This illustrates excellently the statement on page 255 that intensity "... is not necessarily a faithful measure of loudness," for here we have the loudness first increasing and then decreasing while the sound-level remains unchanged.

To see how frequency affects the threshold of hearing (lower boundary line) note, as has already been pointed out, that a lower limit of 10^{-6} microwatts per square meter (taken as the zero sound level) can be heard only if the frequency is between 2000 and 4000 vibrations per second. If the frequency is either 100 or 16,000, the sound is not audible until a frequency of 40 decibels is reached. This is an intensity of 10^{-2} microwatts or 10,000 times that corresponding to the lowest audible sound as ordinarily stated. The threshold of feeling (upper boundary line) is more nearly at a constant level, varying only slightly with changes of frequency.

Deafness

The general effect of deafness is to raise the curve of minimum audibility of Figure 186. For in such a case an intensity greater than the normal is required in order to stimulate the sensation of hearing. But it does not necessarily follow that because a greater intensity is required in order to hear low-pitched tones, a similar increase is necessary for the high, or vice versa. Figure 187 illustrates a case of partial deafness for high tones, and Figure 188 one of deafness for low tones. Through tests of this kind otolo-

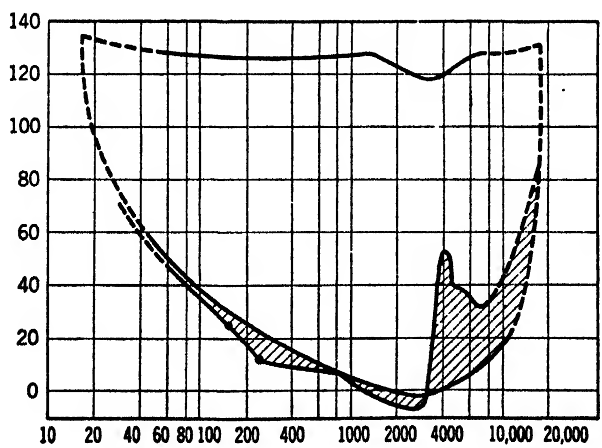


FIG. 187. AUDIOGRAM OF AN EAR DEFICIENT IN HEARING OF HIGH PITCHES

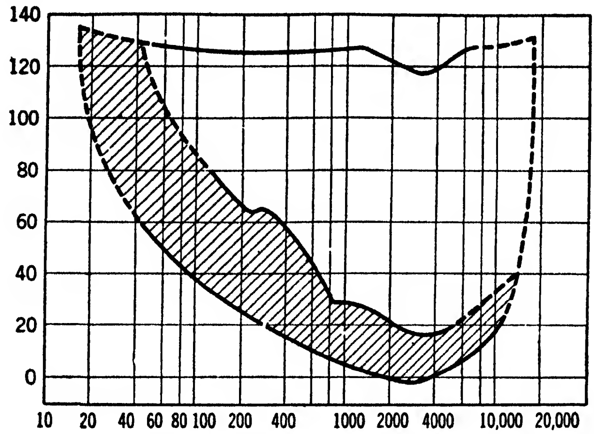


FIG. 188. AUDIOGRAM OF AN EAR DEFICIENT IN HEARING OF LOW PITCHES

gists are able to diagnose the nature of diseases of the ear that would otherwise remain unknown and to prescribe much more effectively the correct remedy — whether it is a case for medical or surgical assistance or for artificial aids to hearing.

These curves make clear the limitations on artificial aids to hearing, for it is evident that there is no virtue in so amplifying the intensity as to raise it above the threshold of feeling. Unless the curve of minimum audibility lies well below this threshold, sound amplifying devices of any kind are of questionable utility.

The audiogram is a method of representing graphically an enormous amount of information about the performance of the normal and abnormal ear. It was devised in the Bell Telephone Laboratories by Dr. Harvey Fletcher and his associates during the 1920's (41:141 ff.).

Beats

The interference of the sounds from two sources was described on page 245 ff. Interference was seen to involve the possibility that the combination of two sounds might possess either greater or less intensity than that of the individual component sounds. One of the common manifestations of these two effects is especially interesting and important. When two sources of nearly but not quite the same pitch act simultaneously, the result is a sound which alternately increases and diminishes in intensity. If the two interfering sounds have the same intensity, the resulting intensity will sink to zero between successive maxima. The resulting alternate swells and lulls are termed *beats*. Beats produced by two sources with a frequency ratio of 24:25 are illustrated in Figure 189.

It will be evident that beats constitute one manifestation of the general phenomenon of interference of sound, yet the two are usually distinguished

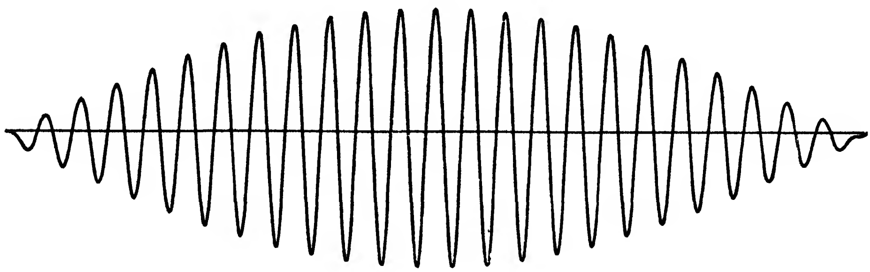


FIG. 189. BEATS AS THE RESULTANT OF TWO PITCHES HAVING THE FREQUENCY RATIO 25:24
(Curve supplied by T. E. Soller.)

from each other. The term *interference* as applied to sound is usually reserved for the combination of equal frequencies, while the term *beats* refers to the combination of frequencies which differ, though only slightly,[1] from each other. The principal distinction is between a stationary and a moving sound pattern. In other words, interference consists, *throughout time*, of *places* of maximum and minimum intensity, whereas beats consist, *throughout space*, of *times* of maximum and minimum intensity.

Tuning by Beats

The frequency of the waxing and waning of intensity which constitutes beats is simply the difference between the frequencies of the beating sounds. This would be evident from the very cause of the beats. One maximum intensity occurs when the two sounds are exactly in phase at the ear of the observer; the next maximum occurs when, having passed an interval of being out of phase, with accompanying waning of intensity, the two sounds again come "into step." Successive maxima can only occur when one of the sounds has gained or lost one vibration with respect to the other. Two pitches of 440 and 439 v.p.s.[2] will accordingly, when sounding simultaneously, produce one beat per second. Two pitches of 440 and 438 will produce two beats per second, and so forth.

The utility of beats in comparing the pitches of two sounds of nearly the same frequency will be evident. If the frequency of one of the sources is known, that of the other may be inferred by counting beats when the two are sounded simultaneously. This does not tell whether the frequency of the unknown is greater or less than that of the standard, however. This may be learned by slightly raising or lowering the pitch of the unknown and repeating the observation. If when the pitch is lowered, the frequency of the beats increases, then the unknown was originally of lower pitch than the standard. If the frequency of the beats decreases, then the unknown was originally of higher pitch than the standard. Similar results obtain —

[1] See, however, page 265 for the treatment of an important apparent exception to this limitation.

[2] The term "vibrations per second" will usually be abbreviated to v.p.s.

though with appropriate changes — when the pitch of the unknown is raised. Though a particularly well-trained musician can bring two pitches into coincidence within about a half a per cent without counting beats, anybody, even a tone-deaf person, can tune an instrument much more accurately than that by counting beats, provided that the original mistuning was not so great as to prevent the detection of beats.

Combination Tones

One of the many major contributions of Hermann von Helmholtz (1821–94) to the science of music was his explanation of what are commonly termed *difference tones*, the discovery of a similar phenomenon called *summation tones*, and a development of the theory of both, called by the general term *combination tones*.

In 1714, an Italian musician named Tartini discovered that when two especially loud tones were sounded simultaneously, a third tone became evident.[1] Later observers established the interesting fact that the frequency of Tartini's "third tone" was simply the difference of the frequencies of the two components.

Difference tones are easily observable. Whenever two persons engage in a whistling duet, soprano and alto, especially if the rendition is energetic, they will become conscious of what seems at first like a curious buzzing in the ears. More careful observation will disclose that this "buzzing" consists of a sequence of tones of rather peculiar quality not uttered by either performer. The tones possess identifiable pitches, much lower than the frequency range of the main performance. Double-barreled whistles, of the type frequently used by police, exhibit the same phenomenon even more prominently. The effect may be studied most readily by the use of an array of small organ pipes mounted on the same wind chest and capable of being blown in pairs.

Suppose that five pipes are selected for such an experiment, namely, C_8 (261.6), E_8 (329.7), F_8 (349.3) G_8 (392.0), and C_9 (523.3).[2] They are represented in pairs in the upper score of Figure 190. When sounded in pairs as shown, the difference tone produced by each pair may be identified with a little practice as the tones represented in the lower staff. These are, for C_8 with E_8, C_6 (65.4); C_8 with F_8, F_6 (87.3); C_8 with G_8, C_7 (130.8); C_8 with C_9, C_8 (261.6); E_8 with G_8, C_6 (65.4). The corresponding differences in the frequencies of the successive pairs are 68.0, 87.6, 130.4, 261.7, and 62.4. Within the approximations which constitute the tempered scale, the frequencies of the difference tones are thus the differences of the frequencies of the respective pairs of primes. Helmholtz discovered that there was another series of tones produced by such combinations, the frequencies of

[1] See A. T. Jones, *American Physics Teacher*, *3*, 49 (1935), for a discussion of priority of discovery of difference tones.

[2] The experiment can be performed with somewhat more satisfactory results if the pipes used are two octaves higher than this. The frequencies which are used here are chosen merely for convenience of representation on the staff of Figure 190.

FIG. 190. PRIMES AND DIFFERENCE TONES

which were the *sums* of the frequencies of the primes, instead of the differences as in the preceding. They are, however, very hard to produce and identify, and seem not to play any considerable practical rôle.

The Nature of Difference Tones

For a long time the cause of these difference tones (and of combination tones in general) was not understood. Helmholtz, who first gave the explanation now generally accepted, formulated it in terms that are too mathematical to warrant quoting in this connection (53:156, 411). The case can be stated somewhat more simply than he does as follows:

Nearly all sources of sound involve the vibration of elastic bodies, the air being classified as an elastic body for this purpose. As long as moderate intensities obtain, the resultant of two superposed sound vibrations is simply the algebraic sum of the components, as in Figure 189. This is true whether the vibrations are superposed in the actual vibrating object constituting the source of the waves or in the membranes of the ear which receives the two sounds. But if in any of these stages the magnitude of the vibrations becomes such as, in actuality or in effect, to exceed the elastic limit of the vibrator, the resultant can no longer be the algebraic sum of the component vibrations. The amplitude of the resultant of the two identical vibrations producing Figure 189 would no longer be twice that of either component. When Helmholtz investigated this condition mathematically, he found that the vibration resulting from such a condition possessed, not merely the frequency pattern characteristic of Figure 189, but also other frequencies that were identifiable, among them one equal to the difference of the component frequencies ("difference tone") and another equal to the sum ("summation tone").

The combination tones are found experimentally to be especially pronounced when the oscillator upon which the component vibrations are being impressed is asymmetrical, so that the successive opposite phases of the resultant vibration possess different amplitudes. One example of an asymmetric vibrator would be a rubber diaphragm with a metal plate covering a large part of its area, fastened to it at the center point. This is illustrated in Figure 191. It happens that the membranes of the ear are asymmetrical vibrators, as are the vibrating members of some musical instruments. A representative form of an asymmetrical resultant vibration is shown in Figure 192.

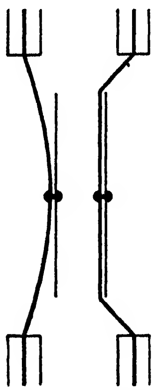


FIG. 191. AN ASYMMETRICAL VIBRATOR

From the fact that the frequency of beats, like that of a difference tone, is equal to the difference of the two combining frequencies, there is some temptation to identify difference tones with beats. Indeed difference tones are sometimes called beat tones. The parallel is somewhat misleading. Mere beats could never produce a third tone, even of the difference-tone pitch, not to mention the summation-tone pitch. One can see, however, how a third tone, of the difference-tone pitch, could arise from the asymmetrical vibration shown in Figure 192. For this case the mean displacement in oscillating positions could be said to lie along the dotted curve of Figure 192. This curve represents a frequency equal to the difference of the combining frequencies. No such frequency exists in the curve of Figure 189, for the mean displacement there is zero.

The Use of Difference Tones in Tuning Instruments

On page 260 the consideration of beats was limited to tones of nearly equal frequencies. That limitation may now be relaxed by virtue of the new element introduced by difference tones.

For example, the difference tone between two frequencies an octave apart has the same frequency as the lower of the two primes (Fig. 190, $C_8 - C_9$). If the octave is slightly mistuned, the difference tone will beat with the lower prime tone. From this arises the phenomenon of beats between two tones an octave apart. Similarly for $C_8 - G_8$ (an interval somewhat ingenuously called a "fifth" because it spans five white keys), the difference tone is C_7. This forms a so-called second order difference tone with C_8, which is also C_7, and beats result if the original chord is mistuned. Again, for $C_8 - F_8$ (a "fourth"), the first order difference tone is F_6. The frequency of F_6 is 87.3. The frequency difference between this

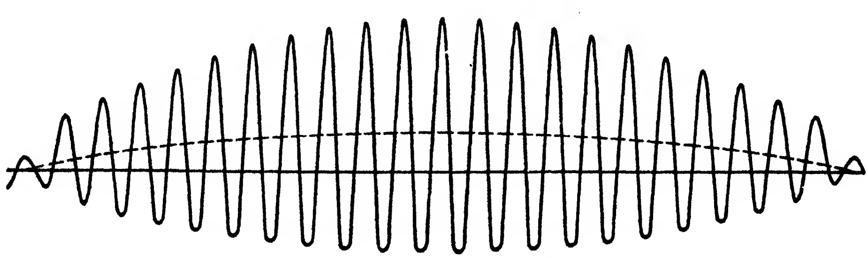


FIG. 192. REPRESENTATIVE FORM OF RESULTANT WHEN TWO IDENTICAL SIMPLE VIBRATIONS ARE IMPOSED SIMULTANEOUSLY ON AN ASYMMETRICAL VIBRATOR

and F_8 is 261.9, which is C_8. The second order difference tone will thus beat with C_8 if the original chord is mistuned.

Hence, for mistuned octaves, fifths, and fourths, beats may be heard, due to the effect of difference tones. Hence, beats may be utilized not only to tune unisons, as described previously, but also to tune other intervals. This is regularly taken advantage of by piano and organ tuners, who find just these three intervals especially useful.

"Laying the Scale"

After bringing the appropriate string or pipe into agreement with the standard tuning fork, by eliminating beats between the unison, the tuner proceeds to "lay the scale" by fifths and fourths. Anyone who has attentively listened to a tuner at work will recognize the following sequence commonly used to establish all the frequencies throughout one of the octaves (Fig. 193). After this, the frequencies outside of this octave can be established by octave beats.

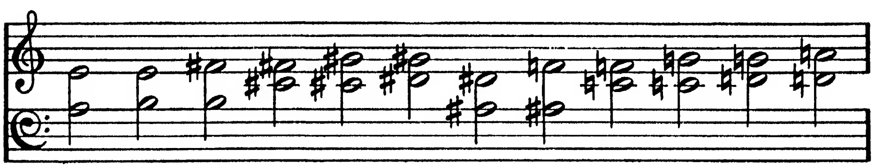


FIG. 193. THE SUCCESSION OF FIFTHS AND FOURTHS USED BY TUNERS TO "LAY THE SCALE"

As will appear in the following chapter, the frequency ratios established between fifths and fourths when beats have been eliminated are respectively 3:2 and 4:3. The corresponding intervals in the tempered scale do not have quite exactly these frequency ratios. Hence, if the tempered scale is desired, tuners should not tune out the beats entirely, but should retain a certain number, which is different for each pair of frequencies used in the tuning. Problem 9 at the end of this chapter develops this sequence. Practically, however, tuners almost never work as accurately as this. The result of the usual job of tuning is therefore a more or less distant approximation to the tempered scale.

Questions for Self-Examination

1. Discuss the relation of loudness to intensity of sound.
2. Define the decibel and state the approximate intensity levels of several common sounds.
3. Draw and interpret the audiogram of the normal ear.
4. Draw and interpret the audiograms of an ear deaf to low pitches and of an ear deaf to high pitches.
5. What are difference tones and how are they produced?

Problems on Chapter 24

1. When twin babies both cry at the same time, how many decibels higher is the intensity of the sound than when one cries alone? Similarly for quintuplets? 3, 7

2. How many times as intense is the sound in a subway as that at the noisiest spot of Niagara Falls? Compare similarly a subway with an average restaurant. 3, 100,000.

3. Helmholtz set 30 beats per second as the rate characterizing maximum dissonance. At what portion of the keyboard is the maximum dissonance for a semitone? for a tone? What musical interval in the region of C_7 produces maximum dissonance? in the octave beginning with C_6? C_9, C_{10}, a third, a fifth.

4. An organ manual covers the range from C_6 to C_{11}. The difference tone produced by depressing a certain pair of keys is near the lowest key on the manual and the summation tone near the highest. What are the two keys? B_9 and C_{10}.

5. Beats often afford a convenient means of determining the frequency of an unknown pitch. A certain pipe of an organ is found to beat 33 times in 10 seconds with a tuning fork which has a frequency of 260 cycles per second and to increase this rate when small pellets of wax are attached to the fork. What is the frequency of the note on the organ? 263.3

6. A tuning fork making N vibrations per second is moved toward a wall. An observer in the line of motion of the fork, the fork being between him and the wall, hears n beats per second due to the Doppler effect. How many meters per second v is the fork moving?

	N	n	v		N	v	n
6.	200	1	.83	7.	200	1	1.2
	200	2	1.7		200	2	2.4
	200	3	2.5		200	3	3.6
	200	4	3.3		200	4	4.8

7. The tuning fork of problem 6 is known to be moving v meters per second. How many beats n per second will the observer hear?

8. Two musical instruments sound the same note simultaneously. One plays in just temperament, the other in even temperament. How many beats n per second will be produced in each pitch of the octave containing the standard A (440)?

8.	C	2
	D	3
	E	0
	F	3
	G	4
	A	0
	B	1
	C	5

	Base frequency	Interval	Magnitude	Discrepancy (beats/min.)
9.	220.	$A - E$	fifth up	45
	329.62	$E - B$	fourth down	67
	246.94	$B - F\#$	fifth up	50
	370.	$F\# - C\#$	fourth down	75
	277.18	$C\# - G\#$	fifth up	56
	415.31	$G\# - D\#$	fourth down	84
	311.13	$D\# - A\#$	fourth down	64
	233.09	$A\# - F$	fifth up	47
	349.23	$F - C$	fourth down	71
	261.63	$C - G$	fifth up	53
	391.99	$G - D$	fourth down	80
	293.66	$D - A$	fifth up	60

9. Calculate the frequency of a perfect fifth above "cello A" (220 v.p.s.). Compare with the frequency of the nearest E, and from this verify the fact that the first step of "laying" the tempered scale is to establish E lower than a perfect fifth above "cello A" by 45 beats per minute. Verify the corresponding facts for the rest of the scale as shown in the adjoining table. (Refer to page 251 and use the frequencies indicated for the first of each pair of pitches.)

CHAPTER 25

The Quality of Musical Tone

Quantity and Wave Form

Such degree of understanding as we now possess of the physical attributes of musical sounds stems in very large measure from the work of Helmholtz. His contributions to the science of sound were only a part of his total work, but even that part was of major significance. Not the least of it was his identification of the nature of what is called the timbre or *quality* of musical tone. He outlined the problem at the beginning of his discussion of the subject in the following words (53:18–19):

> *Loudness* and *pitch* were the first two differences which we found between musical tones; the third was *quality of tone* which we have now to investigate. When we hear notes of the same loudness and same pitch sounded successively on a pianoforte, a violin, clarinet, oboe or trumpet, or by the human voice, the character of the musical tone of each of these instruments, notwithstanding the identity of loudness and pitch, is so different that by means of it we recognize which of these instruments was used....
>
> On inquiring to what external physical difference in the waves of sound the different qualities of tone correspond, we must remember that the amplitude of vibration determines the loudness, and the period of vibration the pitch. Quality of tone, therefore, can depend upon neither of these. The only possible hypothesis, therefore, is that the quality of tone should depend upon the manner in which the motion is performed within the period of each single vibration.

Examples of wave outlines which possess different forms, though they are of the same amplitude and wave-length (wave-length being the graphical analogy to Helmholtz's "period"), may be found in Figure 194. Wave-forms corresponding to tones of three different qualities are represented there. The physical attribute of different tone qualities superficially appears thus to be wave-form. While it is true that different qualities do involve different wave-forms, it is also true that different wave-forms are sometimes to be associated with one and the same quality. This does not, however, require development. For the present, attention will be confined to cases in which different wave-forms do represent different qualities.

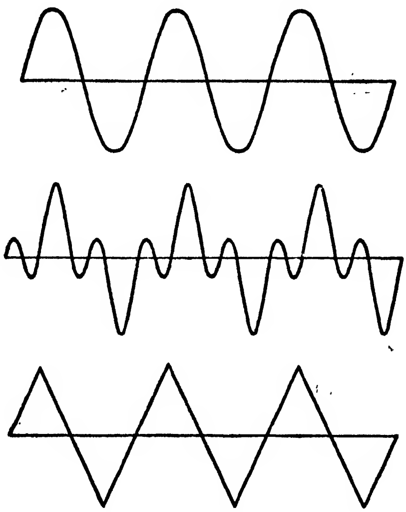


FIG. 194. THREE WAVE FORMS, REPRESENTING TONES WHICH, THOUGH OF THE SAME PITCH AND INTENSITY, ARE OF DIFFERENT QUALITY.

Pitches, Lengths, and Frequencies of Vibrating Strings

To prepare the way for the next development in the study of tone quality, the scene must be shifted back more than two thousand years. In the sixth century B.C., Pythagoras is said to have studied vibrating strings experimentally. He observed that if a string was subdivided, the pitches of the vibrating segments depended on the lengths. Moreover, harmonious combinations could be secured by adjusting the lengths of the vibrating segments to certain ratios.

The law of vibrations of strings (page 129) tells us that frequencies are inversely proportional to lengths. Hence, the foregoing ratios of lengths corresponding to standard musical intervals will also be the ratios of frequencies, paying due attention to inversion. The comparison between frequency ratios on the foregoing Pythagorean scale and on the scale of even temperament will then give some idea of the degree of "perfection of standard musical intervals" as supposedly produced by keyed instruments. The following table summarizes a few of these.

	Frequency ratios			
	Pythagorean		Tempered scale	
Interval	fraction	decimal	fraction	decimal
Octave	2:1	2.	2:1	2.
fifth	3:2	1.500	392./261.6	1.498
fourth	4:3	1.333	349.3/261.6	1.335
major third	5:4	1.250	329.7/261.6	1.260
minor third	6:5	1.200	311.2/261.6	1.189

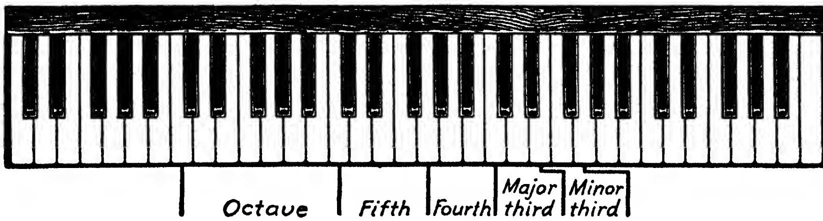


FIG. 195. COMMON MUSICAL INTERVALS BASED ON MIDDLE *C* REPRESENTED ON A KEYBOARD

The third numerical column represents the frequency ratios in the tempered scale on a foundation of middle *C* (261.6). They could, of course, be built on any other pitch, with the same result.

Figure 195 shows a keyboard with intervals as they might be laid out when based on *C*.

Comparison of the two columns headed "decimal," taking the Pythagorean ratios as the standard, gives a measure of the departure of some of the intervals of the tempered scale from this standard. The greatest discrepancy is about one per cent. This would be detectable by a well-trained musical ear under laboratory conditions, but probably not under the conditions ordinarily obtaining in a musical rendition.

Composite Vibrations in Strings

In 1636, Mersenne, who had been the first to make an accurate measurement of the speed of sound, published the first observation of what are now popularly termed "overtones." In studying the vibrations of strings, he detected other pitches higher than ordinarily associated with the vibration of a given string under specified conditions. He identified five pitches in addition to that expected from the string, now called the *fundamental*. These were the first and second octaves of the fundamental, the fifth above the first octave, the major third above the second octave, and the "major second" above the third (unheard) octave (135:140). This being his first observation of this phenomenon, it is not surprising that Mersenne missed some of the "harmonics" which we now know every string exhibits. Figure 196 shows the first twelve of the pitches, termed *partial tones*, emitted by a vibrating string. The fundamental is assumed to be C_6 (65.4), two octaves below middle *C*. The sequence is represented both on a staff and on a keyboard. The six partials which Mersenne heard, including the fundamental, were the first, second, third, fourth, fifth, and ninth. There is no record that Mersenne completed the series, nor even that he identified these higher pitches with the vibrations of the string in segments as re-

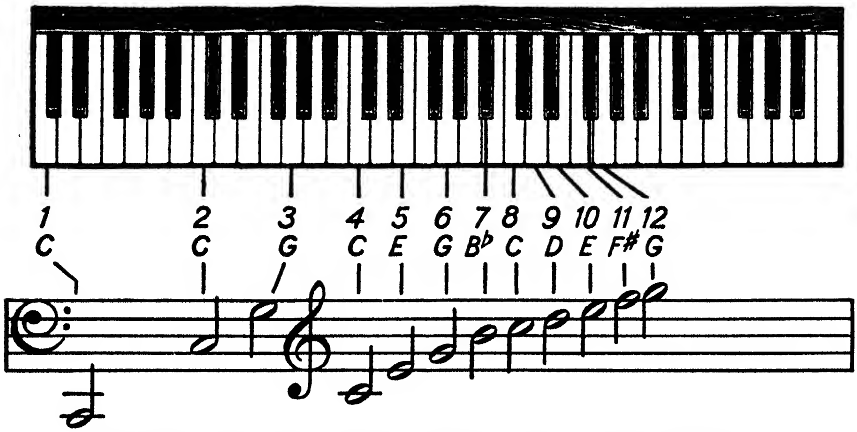


FIG. 196. THE FIRST TWELVE PARTIAL TONES OF A VIBRATING STRING WHOSE FIRST PARTIAL (FUNDAMENTAL) IS *C*

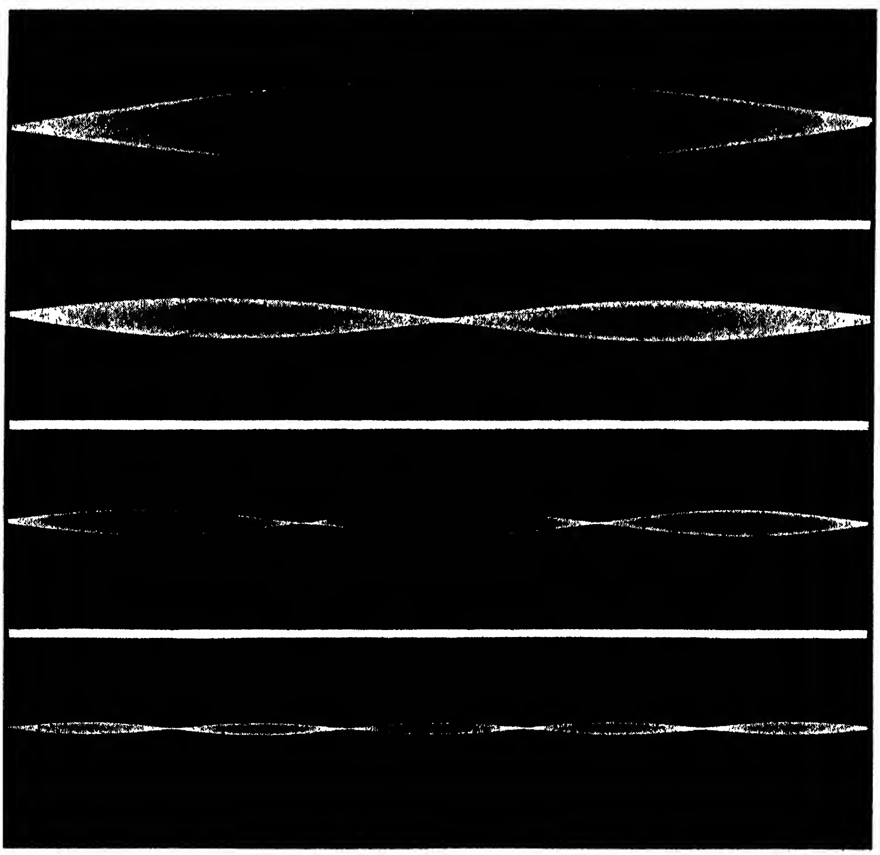


FIG. 197. VIBRATION OF A STRING IN VARIOUS SIMPLE MODES
(From *The Science of Musical Sound*, by Dayton C. Miller. The Macmillan Company, publishers.)

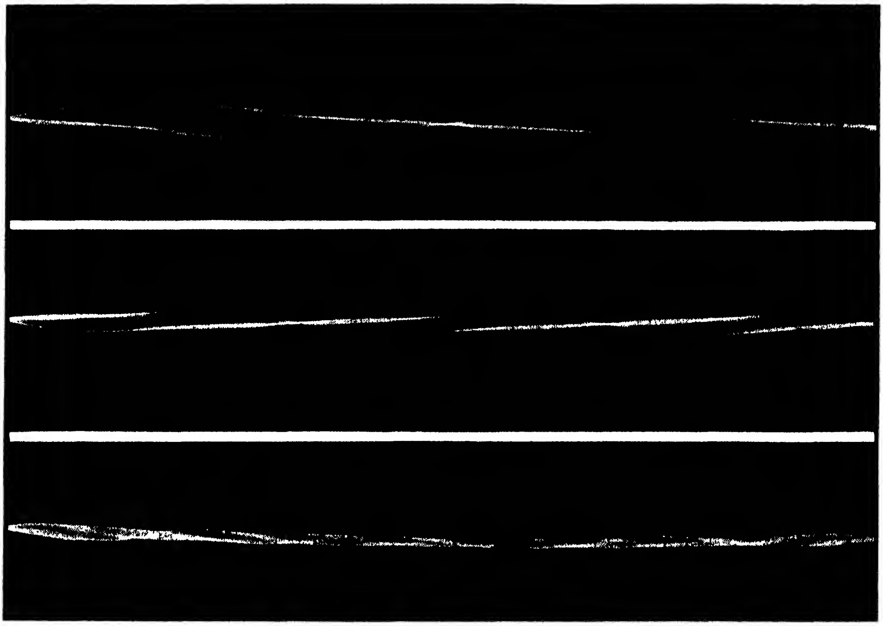


FIG. 198. COMPOUND VIBRATIONS OF A STRING MADE UP OF COMBINATIONS OF SIMPLE MODES
(From *The Science of Musical Sound*, by Dayton C. Miller. The Macmillan Company, publishers.)

corded by Pythagoras. That was first done in 1677 by John Wallis,[1] who has appeared before in connection with the study of impact.

A somewhat more adequate identification of partial tones was made in 1701 by a Frenchman named Sauveur. After discussing the vibrations of strings in segments, he stated that vibrations in these subdivisions occurred at the same time that the string was vibrating in a single segment. As he phrased it (quoted 135:141):

> Each half, each third, each fourth part of a string has its own special vibrations, while at the same time the string vibrates as a whole.

This was the first clear recognition of the composite nature of the vibration of strings. The phenomenon is not easy to visualize. Figure 197 shows time exposures of a string vibrating in a single segment and in two, three, and five segments. Figure 198 shows in the same way strings vibrating, not in simple modes as before, but in combinations of such modes. The irregularity of the outlines is testimony to the complexity of such vibration. The vibrations of the strings of musical instruments in actual performance are characterized by some such complexity of motion.

The capstone to all this was established by Helmholtz, as has already been stated. Viewed in perspective, the recognition seems natural and rather inevitable that tone quality is determined by the form of sound

[1] *Philosophical Transactions* (abridged), *2*, 380 (1677).

waves generated by compounding of partials, these partials arising, in the case of strings, from simultaneous vibration in several modes. Actually in this as in most other fields of science, the final discovery was the climax of several stages, each slowly and even painfully evolved, the whole process covering many centuries.

Recording the Forms of Sound Waves

Effective experimental study of tone quality required, not only a device capable of giving a faithful record of the forms of the sound waves to be studied, but the development of a whole new experimental method. Several major attempts during the fifty years after the time of Helmholtz's great contribution failed to provide either an adequate instrument or an effective method. The need was first supplied by D. C. Miller, of Case School of Applied Science, through his famous *phonodeik*[1] and the experimental technique developed in its use. This gave way some years later to the immensely more versatile *cathode-ray oscillograph.* It is not feasible to describe either of these instruments here, but comprehension of their product requires a brief account of the basic process of recording sound waves, which is common to both.

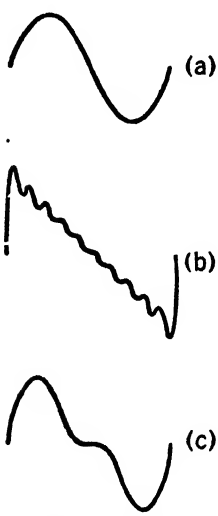


FIG. 199. REPRESENTATIVE FORM OF SOUND WAVE FROM (*a*) TUNING FORK, (*b*) VIOLIN, (*c*) OPEN ORGAN PIPE

If a pencil were attached to one prong of a vibrating tuning fork and a strip of paper were drawn across the pencil perpendicular to the line in which it was vibrating, the resulting trace, after enlargement, would in its essentials be the same as part (*a*) of Figure 199. The same experiment performed upon a violin string would give a variety of results depending on which string was chosen, how it was bowed, and so forth. But among others a curve resembling that of part (*b*) might appear. Again, recording the pressure changes in one type of organ pipe (though it would require something more refined than an ordinary pencil) might produce a curve resembling that of part (*c*). In each case the sound waves themselves, at a distance from the sounding body, would register, through the agency of a properly designed recording device, forms resembling those produced by the vibrating bodies.

The phonodeik and later the cathode-ray oscillograph were capable of thus recording the forms of sound waves from all kinds of sources. Such records constituted the first material upon which an exhaustive study could be made of the relation of tone quality to wave-form.

[1] For a description of the phonodeik, together with an account of earlier attempts to solve the problem, see 85: Lecture III.

The Rudiments of Wave Analysis

But successful recording of wave-forms is only the beginning. The heart of the study is the "breaking down" or analysis of the recorded wave into its "harmonic components." What this process means may be illustrated by further consideration of the curves of Figure 199.

The curves of Figure 199 have the same wave-length and the same amplitude, only the wave-forms being different. Part (*a*) is of particular interest. It is the simplest of the group. It lacks any complications in the way of kinks or angles. It is, in fact, a *simple harmonic wave.* We imagined it to have been produced by a tuning fork, as indeed it might have been. The tuning fork is one of the very few musical instruments which normally produces this kind of curve. A tone corresponding to this wave form is called a *pure tone.*

Now consider part (*c*) of Figure 199. It is only a degree less simple than part (*a*). It can be constructed out of two simple curves, both like that of part (*a*) except for size. The first is identical with part (*a*), and the second has half the wave-length and half the amplitude of the first (Fig. 200). Superposing the second on the first gives the curve of part (*c*) of Figure 199. It appears that the air in the organ pipe which we imagined to have produced part (*c*) must have had two types of vibration acting simultaneously, both simple but combining to form the curve of part (*c*). This is the same type of occurrence that has already been observed in strings.

Now consider part (*b*) of Figure 199. This was the curve ostensibly pro-

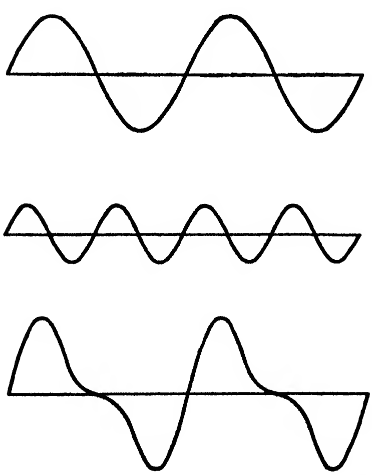


FIG. 200. PART (*c*) OF FIG. 199 AND ITS TWO SIMPLE HARMONIC COMPONENTS

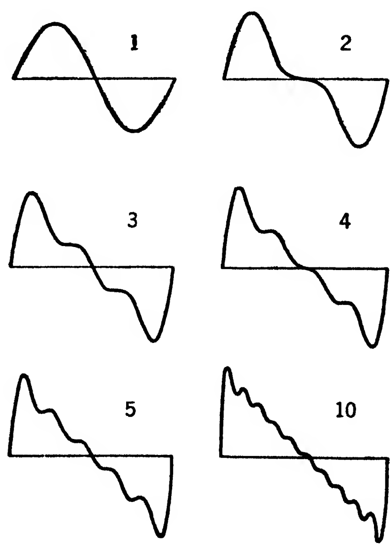


FIG. 201. THE EFFECT OF THE ADDITION OF SUCCESSIVE PARTIALS ON THE FORM OF THE COMPOSITE WAVE

duced by the violin string. The series of partial tones, all of which we are imagining to be present, will have the relative frequencies 1, 2, 3, 4, 5, and so forth. If the intensity of each upper partial is assumed to be less than that of the one below it — which is not always the case — it will be convenient to regard the relative amplitudes as 1, $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{5}$, and so forth, multiplied by that of the fundamental. Figure 201 shows the process of successively adding such a series of partial tones. The second curve is, of course, identical with that of part (*c*) of Figure 199. Those following show the effect of adding the successive overtones that have been specified. The inclined part of the resultant curve approaches more and more closely to a straight line as a larger and larger number of partials are added.

Compounding simple harmonic curves into their resultant is termed *synthesis*. The reverse process, breaking down composite curves into their simple harmonic components, is termed *analysis*. That any recurrent curve can be thus analyzed into a unique series of simple harmonic components was first announced by a Frenchman named Joseph Fourier in 1807. The statement is one form of what is termed *Fourier's theorem*. The consequences of this theorem bob up in most unexpected places in all portions of physics. Though they are especially prominent in the study of sound, the principle was discovered in the course of Fourier's study of the conduction of heat. Helmholtz made his acknowledgments to Fourier and restated his theorem in a form most readily applicable to the analysis of tone as follows (42:34):

> Any vibrational motion of the air in the entrance to the ear, corresponding to a musical tone, may be always . . . exhibited as the sum of a number of simple vibrational motions, corresponding to the partials of this musical tone.

Sound Spectra

The two components of the wave pictured in part (*c*) of Figure 199 were represented separately in Figure 200. There is, however, a simpler and more informative way of representing them. The two components differ from each other in two respects, wave-length and amplitude. These may be represented, as in Figure 202, by vertical lines, whose positions on a horizontal scale are governed by the frequencies (which are in inverse proportion to the wave-lengths) and whose lengths are determined by the amplitudes. This representation of the frequency and amplitude of the partials of a composite tone is called a *sound spec-*

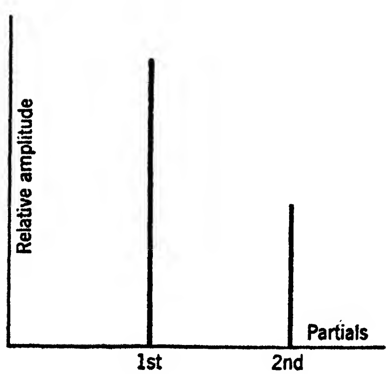
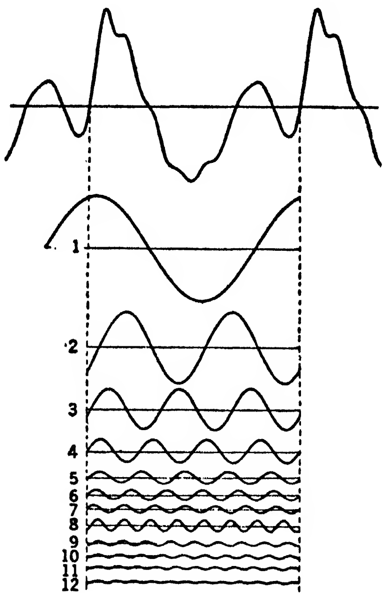


FIG. 202. SOUND SPECTRUM METHOD OF REPRESENTING THE PARTIALS INDICATED BY THE WAVES IN FIG. 200



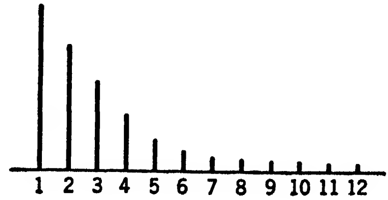


FIG. 203. CURVE REPRESENTING THE WAVE FORM OF THE TONE OF AN ORGAN PIPE

The component partials are represented here by simple harmonic waves and by a sound spectrum. (From *The Science of Musical Sounds*, by Dayton C. Miller. The Macmillan Company, publishers.)

trum. The upper left portion of Figure 203 shows the wave-form of the tone of an organ pipe. Below this is the array of simple harmonic curves into which the composite curve was analyzed, and at the right of the figure the corresponding sound spectrum. Each component simple curve is represented by a line indicating frequency by its position and amplitude by its length.

In the same way Figure 204 shows the actual wave-form of the tone of a

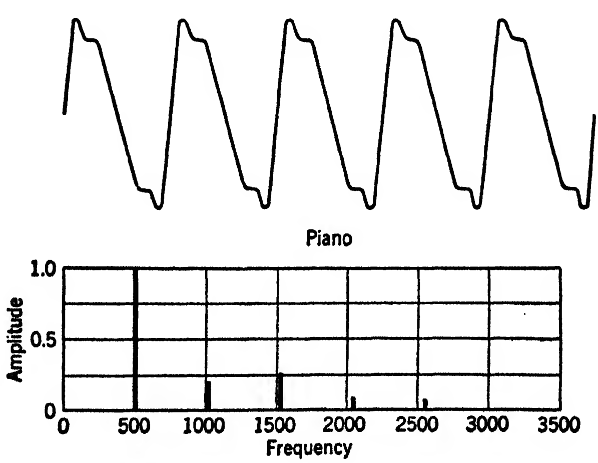


FIG. 204. WAVE FORM AND SOUND SPECTRUM OF THE TONE OF A PIANO
(From *Speech and Hearing*, by Harvey Fletcher. D. Van Nostrand and Company, publishers.)

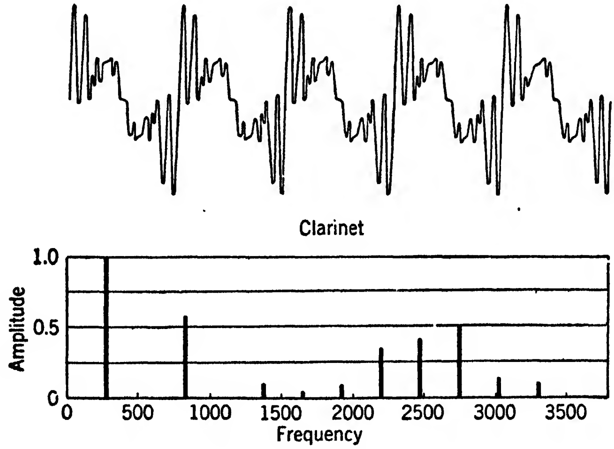


FIG. 205. WAVE FORM AND SOUND SPECTRUM OF THE TONE OF A CLARINET
(From *Speech and Hearing*, by Harvey Fletcher. D. Van Nostrand and Company, publishers.)

piano together with its sound spectrum. The frequency of the fundamental is 512. Figure 205 represents similarly the analysis of the tone of a clarinet with fundamental frequency 256.

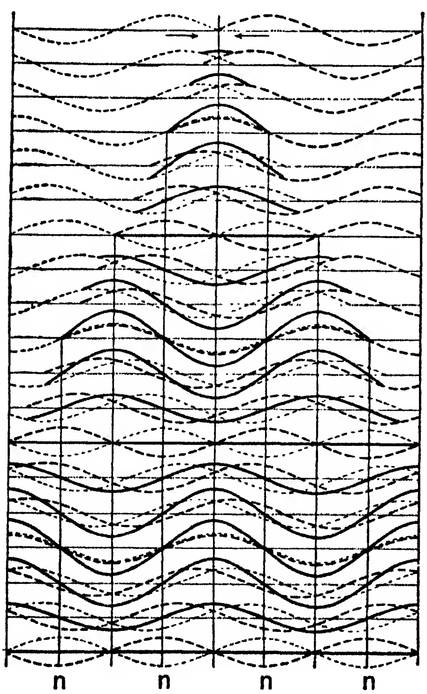


FIG. 206. THE COMBINATION OF TWO OPPOSITELY DIRECTED RUNNING WAVES INTO A STATIONARY WAVE

Stationary Waves and Running Waves

The vibration of a string in one or more segments constitutes an example of what are called *stationary waves.* The phenomenon is of first importance in the study of sound, since every musical instrument without exception depends for its action on the principle of stationary waves. The real nature of stationary waves can hardly be appreciated, however, without reference to the running waves of which they are composed.

If a long rope or wire is struck or sharply plucked near one end, a wave will be set up which can be seen to travel to the other end and return by reflection, possibly repeating its to-and-fro journey several times before it dies away. This is an example of a *running wave.* If instead of a single impulse, a steady train of identical waves had been set up at one end, the string would shortly have been bearing two such trains, moving in opposite direction, one the direct and the other the reflected train.

The effect of the combination of two oppositely directed trains of waves in a string is indicated in Figure 206. The two trains first encounter each other at the middle of the string. Beginning at the top, the successive stages in the resulting motion of the string where these two trains overlap are shown. To describe these successive stages would be awkward and unnecessary. They can best be envisioned by a careful scrutiny of the figure.

Some general observations may be made, however. At points on the string marked by the vertical lines *n*, *n*, *n*, *n*, no motion occurs. These points are commonly called *nodes*. Midway between these points the string moves up and down through a greater amplitude than anywhere else. These points are commonly called *loops*. Comparison of the resultant displacements with the component displacements due to the individual waves will show that at the nodes the displacements due to one of the waves is always equal and *opposite* to that due to the other. This is the reason for the continuously zero resultant displacement at the nodes. At the loops, on the other hand, the displacements due to the individual waves, while also always equal, are *in the same direction*. The two displacements, therefore, add at these points. Neither of these conditions obtain at intervening points.

Figure 207 is a photograph of a stationary wave pattern produced by two sets of ripples traveling in opposite directions. The ripples themselves

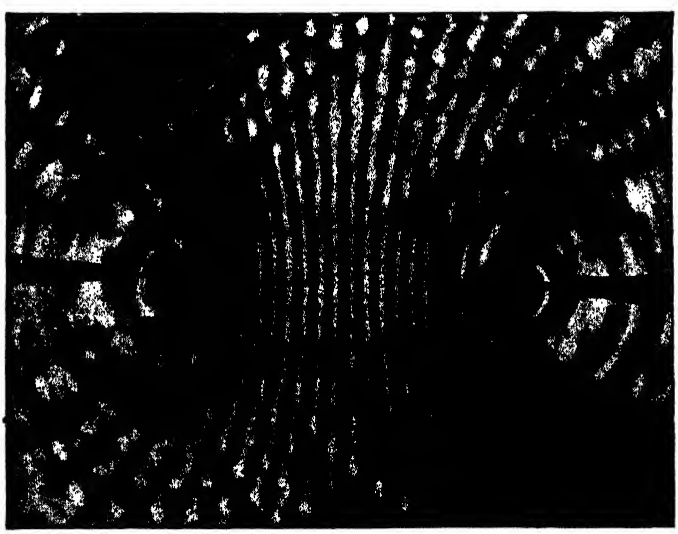


FIG. 207. STATIONARY WAVES ON WATER CREATED BY TWO OPPOSITELY DIRECTED TRAINS OF RIPPLES
(From Michelson's *Studies on Optics*, Fig. 5. University of Chicago Press.)

are produced at the two needle-points dipping into the liquid and are only faintly visible in the background as two sets of concentric circles expanding from their respective centers.

Thus stationary waves may be produced, and indeed always are produced, by the combined effects of two identical oppositely directed running waves in the same medium. This is a special case of interference already discussed in general terms (page 244). The segments into which vibrating strings have been seen to divide (Fig. 197) are thus identified as successive loops in a system of stationary waves. It will be evident that the distance between adjacent loops or adjacent nodes is one-half the length of the wave in that particular medium. Knowing the frequency, n, in addition to observing the wave-length, λ, in this way, the speed, v, of the wave in the string can be computed from the relation

$$v = n\lambda. \tag{1}$$

Boundary Conditions in Strings

The two ends of a string under tension are almost necessarily clamped. It is hard to imagine tension being applied under any other circumstances. In a sense, a streamer fluttering in a stiff, steady breeze is under tension with one end free; so is a heavy string suspended vertically from its upper end. These cases are somewhat trivial, however, and in general it is unavailing to try to consider what would be the attributes of the stationary waves set up in a string which does not have both of its ends clamped. It is well to keep such an imaginary possibility in mind, however, for it will shed some light on the behavior of stationary air waves in pipes, soon to be considered.

Because both ends of a string must be fixed and, therefore, constitute nodes, a string can vibrate in any integral number of segments (1, 2, 3, 4, ...). The familiar frequency sequence of the partial tones of strings (in ratios 1, 2, 3, 4, ...) is a consequence of this fact. The same sequence of partials would exist in the purely imaginary case of a vibrating string with both ends free. But for the case of a string with one end fixed and one end free the free end would be a loop instead of a node. The number of segments would clearly not be integral for this case, but would be an odd number of half segments ($\frac{1}{2}$, $\frac{3}{2}$, $\frac{5}{2}$, ...), and the frequency sequence would show the same ratios (1, 3, 5, ...). This is not a practicable possibility for strings, as has already been pointed out. It is mentioned here because it has a counterpart in the modes of vibration of stationary waves in pipes such as will now be considered. The parallelism between the two, even though merely theoretical in the case of one of them, is illuminating.

Stationary Waves in Pipes

Until one becomes accustomed to thinking in terms of longitudinal waves, it is hard to visualize a longitudinal stationary wave, such as characterizes sound waves in pipes. The process may be facilitated by a study of Figure 208. This represents the successive motions constituting a longitudinal stationary wave. The nodes are abreast of the dotted vertical lines *n, n, n, n*. The vibratory motion, instead of being transverse as in Figure 206, is longitudinal. The seven stages represent half a complete period.

The sinusoidal line indicates the magnitude of the displacements of the various layers of air from their normal position corresponding to uniform density. Thus the column of air will be seen to divide into segments characterized by nodes and loops just as truly as did a string. The displacements, though longitudinal, may be represented as transverse. To do so helps considerably in following the process. Thus the stationary wave pattern, the successive stages of which are pictured in Figure 208, is represented diagrammatically in Figure 209. The pattern (four nodes, five loops) is to a pipe what the last part of Figure 197 was to a string.

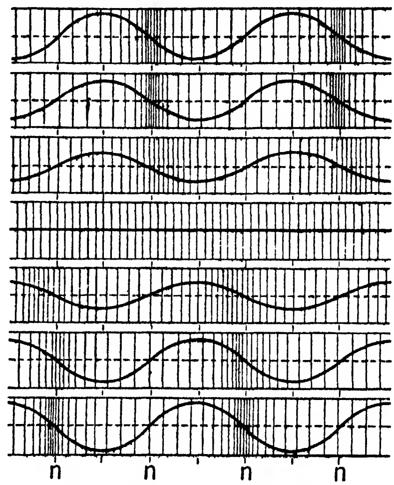


FIG. 208. MOTION INVOLVED IN A LONGITUDINAL STATIONARY WAVE

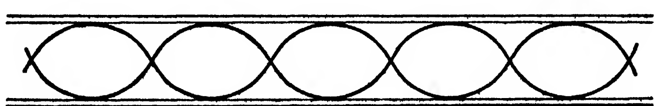


FIG. 209. DIAGRAMMATIC REPRESENTATION OF A STATIONARY LONGITUDINAL WAVE OF FIVE SEGMENTS

Boundary Conditions in Pipes

Unlike strings, pipes may produce either of two partial-tone sequences. One is the same as that of strings, the complete series of frequency ratios (1, 2, 3, 4, . . .), and the other is the odd-numbered series of frequency ratios (1, 3, 5, 7, . . .). The former is produced by *open pipes* (pipes having both ends open); the latter, by *closed pipes* (pipes having one end closed). The third possibility, pipes having both ends closed, is almost as much removed from practical possibility as is the corresponding case of a string having both ends free. Pipes which confined the sound within their interiors would scarcely make acceptable musical instruments. Besides, there must be access to the interior in order to blow the pipe. Whether the blown end of a pipe or wind instrument should act like an open or closed end might be in doubt if only its structural features were taken into consideration. Experiment shows, however, that the blown end usually produces an "open" effect.

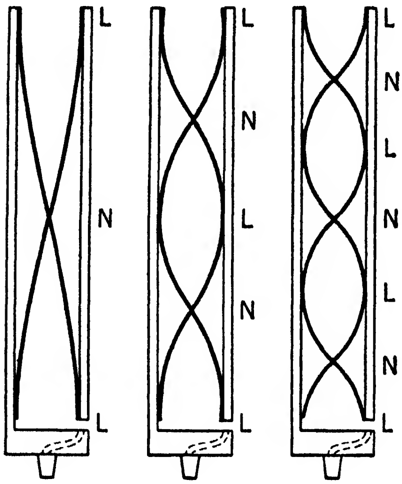


FIG. 210. THE FIRST THREE MODES OF VIBRATION IN AN OPEN PIPE

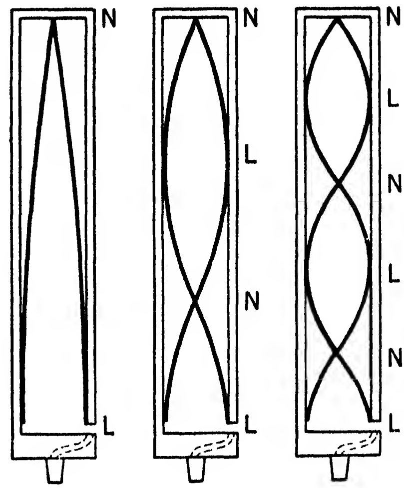


FIG. 211. THE FIRST THREE MODES OF VIBRATION IN A CLOSED PIPE

The reason for the difference in the action of open and closed pipes has already been intimated: the likeness or unlikeness of the boundary conditions. Figure 210 shows the first three modes of vibration in an open pipe. The ratios of the number of segments are 1, 2, 3. Figure 211 shows the first three modes of vibration in a closed pipe. The ratios of the numbers of segments are 1, 3, 5. Open ends always have loops; closed ends, nodes.

As in the case of a string the distance between adjacent nodes (or adjacent loops) marks half the wave-length of the sound. There is this difference, however. In strings such distances represent wave-lengths *in the string*, whereas in pipes such distances represent wave-lengths *in air*. Pipes are consequently more useful as indicators of the lengths of sound waves emitted to the surrounding air.

Manifestly then, an open pipe sounding its fundamental possesses a length approximately half that of the sound wave which it generates, and a closed pipe, approximately a quarter the length of the sound wave which it generates. In the second partial tone of the open pipe, the length of the pipe marks two halves that of the wave, the third three halves, and so forth. In the second partial tone of the closed pipe the length of the pipe marks three quarters of a wave-length, the third, five quarters, and so forth. The qualification involved in the word "approximately" arises from certain disturbances at the ends, affecting the exactness of the ratio. The discrepancy depends upon the ratio of the diameter of the pipe to its length if it is completely open, or upon the relation of the dimension of the aperture to the length if the pipe is not completely open.

The case of pipes which are not straight sided — brass instruments and certain others — is naturally somewhat less simple than that of the straight sided pipes discussed here, and will not be considered.

Reflection in Pipes

The closed end of a pipe acts on adjacent layers of air much as the fixed end of a string acts on its adjacent portions. The layer next to the closed end is clamped to the end almost as literally as the string is clamped to its support. It could move away only by virtue of the creation of a vacuum, which just doesn't happen in practice. It being thus "fixed," the longitudinal sound waves are reflected from it under the same condition which governs reflection of the transverse waves at the fixed end of a string. That is, there is a reversal of their displacements. This produces reflection without change of phase — which in this case means that condensations are reflected as condensations and rarefactions as rarefactions.

The opposite occurrence accompanies the reflection of a sound wave at the open end of a pipe. Condensations are reflected as rarefactions and rarefactions as condensations. This is not so easy to visualize, but when once seen, it explains why the open end of a pipe is always the scene of a loop of the stationary wave pattern within the pipe. The displacement of the particles constituting a condensation occurs in the direction of motion of the wave. That of the particles constituting a rarefaction occurs in the opposite direction. Hence, when a condensation traveling to the right is reflected as a rarefaction traveling to the left, the displacements of the two combining waves near the open end of the pipe are both to the right, and their magnitudes, therefore, add. The result is a maximum amplitude in that portion of the stationary wave pattern next to the open end of a pipe — thus defining a loop.

Bars, Diaphragms, and Plates

In nearly all the common musical instruments the vibrating element is either a string or a pipe. There are a few in which the vibrating element is a bar. This would include the xylophone (wooden bars) and the glockenspiel (metal bars) and might be extended to include reeds, which are really small and very flexible bars. A few instruments involve the vibrations of diaphragms or plates. The drum is an example of the former and cymbals of the latter. These instruments are somewhat non-standard, however, and need not be considered. Moreover, the vibrations of bars, diaphragms, and plates are extremely complicated. The upper partials are numerous, inharmonic, and unstable. That is to say, their frequencies are not integral multiples of the frequencies of the fundamental, as is the case with strings and pipes, and those which are to be elicited depend more on the mode of excitation of the instrument.

It should be realized also that other factors besides the direct contributions of upper partials contribute to the general effect produced by a musical instrument. An enormous contribution is made by the combinational tones between these partials, though very inadequate attention has been given to that point. Equally important is the effect produced by the circumstances of starting or stopping the tone and the mode of excitation in general. The effect of such instruments as the piano, for example, is in no small measure due to the original impact followed by the subsequent dying away of the intensity. This factor is not included in the "quality" of the tone as ordinarily defined by the physicist. As yet the physicist has been able to account for the properties of musical tone only to a limited extent. While the time will ultimately arrive when he can give a more complete account and even indicate major ways of improving the design of basic musical instruments, that day is not yet upon us. It is only in reproduction of music, such as by radio, phonograph, and sound pictures, that the science of music has taken the lead over the art of music.

Questions for Self-Examination

1. How may quality of a musical tone be associated with the accompanying wave-form? Who discovered it?
2. Who made the first association between pitches and lengths of vibrating strings? Between pitches and frequencies?
3. Show two ways of representing graphically the analysis of a sound wave. What are their respective advantages?
4. What are stationary waves and how are they produced?
5. Compare modes of vibration in "open" and "closed" pipes.
6. Compare "boundary conditions" in strings with those in pipes.

Problems on Chapter 25

1. What would be the outline of a vibrating string, photographed by time exposure, when its first and second partials (fundamental and first overtone) are especially prominent?
2. The bugle has no valves. How can different notes be sounded on it? To what notes is the bugler practically limited and why?
3. An open organ pipe has a length of 40 centimeters. What will be the frequency of the fundamental and of the first four overtones? Similarly for a closed pipe of the same length.
4. If the loudness of a musical sound is increased in such a way that all the partial tones are augmented in the same ratio, it is still possible that there might be a change in quality. Why?
5. When the first few partials (including the fundamental) of a tone are removed by an acoustic filter, the pitch remains unchanged but the quality is profoundly affected. Why (for both circumstances)?
6. A wire is L meters long and weighs M grams. What tension F in kilograms must be applied if the wire is to produce a tone whose frequency is n vibrations per second?

	L	M	n	F
6.	1.	2	200	33
	1.	4	150	37
	1.	6	100	24
	1.	8	75	18

	L	r	l
7.	120	$\frac{16}{15}$	7.5
	120	$\frac{10}{9}$	12.
	120	$\frac{9}{8}$	13.
	120	$\frac{6}{5}$	20.
	120	$\frac{5}{4}$	24.
	120	$\frac{4}{3}$	30.
	120	$\frac{3}{2}$	40.

7. The strings of a cello have a length L centimeters. By how many centimeters l must they be shortened by fingering to change the pitch by a frequency ratio r?

8. Find the length l_1 in meters of an open cylindrical organ pipe, d centimeters in diameter, which will give a note of frequency n vibrations per second, the end error being .6 of the radius; and find also the length l_2 of a closed pipe. Take the speed of sound as 340 m/sec.

	note	n	r	l_1	l_2
8.	C_4	16.35	40.	10.	5.1
	C_6	65.41	12.5	2.6	1.3
	A_8	440.	2.5	.38	.19
	C_{11}	2093.	1.	.078	.038

	Interval	Dissonant partials	Number beats
9.	*C–G*	none	
	C–F	none	
	C–E	10th, 8th	20
		9th, 7th	50
		5th, 4th	10
	E–G	7th, 6th	50
		6th, 5th	20
	F–G	10th, 9th	40
		9th, 8th	10
		8th, 7th	50
		1st, 1st	40
	E–F	2d, 2d	40
		1st, 1st	20

9. Find the dissonant partials between the respective intervals of a fifth, a fourth, a major and minor third, a tone, and a semitone. Take as representative of these respective intervals the combinations *CG*, *CF*, *CE*, *EG*, *FG*, and *EF*. Use partials up to the tenth inclusive. Take frequencies of *C*, *E*, *F*, and *G* respectively as 261.5, 329.6, 349.2, and 392. Take as the criterion of dissonance any number of beats between 5 and 50 per second.

LIGHT

CHAPTER 26

Elementary Properties of Light

Rectilinear Propagation of Light

It is commonly stated in textbooks that light travels in straight lines. The statement requires clarification, but is based upon a volume of common observation which extends back so far that it is impossible to identify its first author. Light is commonly contrasted with sound in this respect. It is pointed out, for example, that after passing through an aperture light continues directly on and casts a sharp shadow, whereas sound billows out in all directions and is audible, not merely in the direct line of the source and aperture, but at all points on the far side of the obstruction. The significance of this distinction will be developed in Chapter 33. There are, however, many apparent exceptions to the rectilinear propagation of light.

But, notwithstanding these apparent exceptions, the doctrine (if it may be so termed) of rectilinear propagation of light will probably continue indefinitely as a generalization on the main trend of commonly observed phenomena. It is not really a scientific "law," but within its limitations it is a useful concept and as such finds numerous applications. Too rigorous an insistence on its universal validity can, however, lead to difficulties, as Newton realized when he tried to incorporate this doctrine into his formulation of the essential properties of light.

Speculations on Speed of Light

A second ancient source of perplexity about light was the speed of its propagation. Empedocles (*ca.* 490–435 B.C.), the first Greek philosopher to give serious attention to light, did not credit it with instantaneous propagation as did many later philosophers, but believed that it moved with a finite speed (111:*1*:87). Pliny (A.D. 23–79) (111:*1*:249) was more explicit. He remarked that the speed of light was greater than that of sound — a deduction which any child can make from the common observation that one can see a distant hammer fall an appreciable interval before one hears the sound of its impact. A thousand years later Al-Biruni (973–1048), whom Sarton considers one of the greatest scientists of all time (111:*1*:707–08), declared that the speed of light was immense as compared with the speed of sound. And two hundred and fifty years after this, Roger Bacon expressed his faith that, though the speed of light had proved too great to measure, it was nevertheless finite (13:*12*:489).

This would appear to be all that could be expected, in the absence of experimental data. Unfortunately the lack of such data made it impossible to decide between this prophetic opinion and the opinion, which was also widely held, that the propagation of light occurred instantaneously. Both Kepler (64:9) and Descartes (quoted 105:125) held this opinion, though the latter, oddly enough, also promulgated the irreconcilable doctrine that the speed of light varies in different media.

Measurements of the Speed of Light

The first attempt actually to measure the speed of light was made by the Florentine Academy (27a:12) in 1667 following an earlier suggestion of Galileo (46:43). They undertook to determine the time required for two observers to flash lanterns back and forth at each other, when they were stationed nearly a mile apart. They expected thus to secure a measure of the speed of light. It is unnecessary to remark that the experiment was unsuccessful. It is important, nevertheless, not merely because it is significant historically, but also because some of the later successful determinations involved the same principle — sufficiently refined to cope with the prodigious speed which light possesses.

The first measured value of the speed of light was reported to the French Academy by Olaf Roemer (1644–1710) in 1676. Roemer had been studying the motions of Jupiter's moons and had observed that a greater time elapsed between successive passages of one of the moons behind Jupiter during the six months when the earth was receding from Jupiter than elapsed during the other six months when the earth was approaching. Roemer concluded that light required 22 minutes to travel across a diameter of the earth's orbit, thus giving a value of about 227,000 kilometers per second for the speed of light. In comparison with the later more accurate values (300,000) Roemer's value was much too low, though it was of the correct order of magnitude. It was, however, great enough to strain the credulity of many of his contemporaries beyond the breaking point. His work was at first unfavorably received, and it was not until eighteen years after his death that an unexpected but conclusive confirmation of his work put an end to the current doubts.

Of the work of Bradley, whose observations furnished the crucial verification and correction of Roemer's results, as well as of the great refinements introduced into the measurement of the speed of light successively by Fizeau, Foucault, and Michelson, space will allow but scant account. The method of Fizeau (77:340) is perhaps the most significant, partly because it furnished the model followed by the succeeding experimenters, but also because it embodied in an effective experimental method the principle which Galileo had suggested.

Fizeau replaced Galileo's first lantern shutter by a rapidly revolving toothed disk, the successive apertures and teeth of which allowed some nine thousand pulses of light to pass each second. He replaced Galileo's second lantern by a mirror placed about five miles away, which reflected

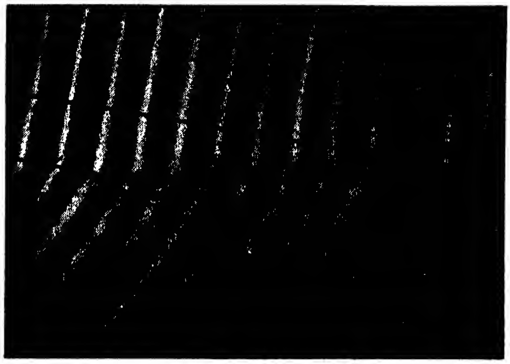


FIG. 212. REFRACTION OF RIPPLES
(From *General Physics for Colleges*, by Webster, Farwell and Drew, The Century Company, publishers.)

these light pulses back to the toothed disk. If the disk was stationary, the light could be made to return through the same aperture that it had traversed in the outward journey. But if the disk was turning fast enough, the path of the returning pulse of light would be blocked by the tooth adjacent to the original aperture, which in the meantime had moved into the path of the beam. If the speed of the disk was doubled, the light could again return, but now through the *adjacent* aperture, not through the original one. Calculation of the time which the light required to make the double journey was made possible by noting the speed of the disk and the number of teeth; from these the speed of light could be computed. The value thus obtained was about 5 per cent higher than that commonly accepted now.

Thirteen years later, in 1862, Leon Foucault avoided some of the inaccuracies which were inescapable in Fizeau's method by substituting a rotating mirror for the toothed wheel and thus secured greater precision of results. By further improvement of technique, Michelson, in 1880 and again in 1927, secured the highest accuracy thus far attained. The final value was:

$$299{,}796 \pm 4 \text{ km/sec.}$$

Whether the speed of light was greater in other media than it was in air, or less, was a moot question until as late as 1850 when Foucault, in his early experiments with the rotating mirror method demonstrated for the first time that "the velocity of light is less in water than in air."[1] This was twelve years before he made his precision measurement of the speed of light in air. During the seventeenth and eighteenth centuries the prevailing opinion had been that light traveled more rapidly in dense media than in air. Such was the opinion, already referred to, of Descartes. Newton also had cast the weight of his opinion on that side, under the impression that the refraction of light could best be accounted for in this way (91:226–28; 90:79). Possibly this opinion received some support by analogy with the known fact that sound traveled more rapidly in metals than in

[1] *Comptes Rendus*, *30*, 556 (1850).

air. In any case, the contention of Huygens in 1678 (61:32–38) that light must slow down upon entering dense media was almost entirely disregarded for a century.

Refraction of Light

There was as much opportunity for difference of opinion concerning the change in *direction* that characterized the passage of light into a dense medium, as there was concerning the change in *speed*. Aristotle, in his *Book of Problems*, correctly described the appearance of an oar dipped in water. Ptolemy (*ca.* A.D. 127–51) tabulated angles of refraction corresponding to angles of incidence for air into water, air into a piece of glass, and water into glass and sought for some law connecting these angles. He concluded that the ratio of the angle of incidence to that of refraction was constant for the same pair of media, a generalization which seems scarcely justified on the basis of his tables of refraction.

Ptolemy's tables were extended (though not entirely correctly) by Vitellio about 1270 and by Kircher (1601–80). Kepler, who, in addition to his astronomical studies, made important contributions to optics, was so impressed with the work of the former that he entitled his first book on optics (1604) *A Supplement to Vitellio.* He started his second book on optics, *Dioptrice*, written seven years later, by describing his experiments on refraction, in which he used the device illustrated herewith (Fig. 213). He, like Ptolemy, tried to discover some relation between the angles of incidence and refraction. The best he could do was to record that, for angles of incidence less than about 30° in air, the angle of refraction in glass was about two thirds the angle of incidence, a formulation not greatly in advance of that of Ptolemy, nearly fifteen hundred years before. It was natural for these early investigators to look for a relation between the angles, but in 1621 Willebrord Snel (1591–1626) discovered that the behavior of light upon refraction could be formulated, not by a ratio of the *angles* involved, but instead by a ratio of *trigonometric functions* of the angles. Snel's law, as it is now stated, is:

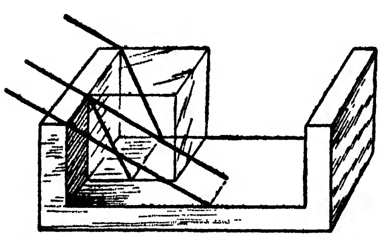


FIG. 213. KEPLER'S REFRACTOMETER
(From his *Dioptrice* of 1611.)

$$\frac{\sin i}{\sin r} = \mu \qquad (3)$$

where i represents the *angle of incidence*, r the *angle of refraction*, and μ the ratio of their sines, termed the *index of refraction* of the second medium.

Refraction and the Speed of Light

Though Snel's observation settled the ancient problem of the direction of refracted light, it did not determine whether light was retarded or accelerated in the process of refraction. That controversy continued until Foucault's measurements in 1850, already mentioned (page 289).

A simple illustration of refraction produced by the diminution of the speed of a wave is to be found in the change in direction of waves or ripples upon oblique approach to a beach. Figure 212 shows this effect on a train of ripples. The water, very shallow everywhere, is rendered still more shallow in the lower portion of the photograph by a sheet of glass laid on the bottom of the tray. The resulting diminution in the speed of the lower portion of the ripples produces the refraction shown. This is also true, we now know, with light waves entering such substances as glass or water.

Questions for Self-Examination

1. What can be said of the common impression that light travels in straight lines?
2. Tell the approximate value of the speed of light and outline the stages in the development of methods of measuring it.
3. Describe Roemer's method of determining the velocity of light by the eclipse of Jupiter's moons.
4. Define index of refraction and sketch the evolution of the idea.
5. Give an account of the old controversy over whether light traveled faster or slower in air than in glass.
6. Define index of refraction. How is it related to the speed of light?

Problems on Chapter 26

1. In a repetition of Fizeau's determination of the speed of light the distance between the wheel and mirror was 8500 meters. The wheel had 88 teeth. When the first occultation of the reflected beam occurred the wheel was rotating at 100 r.p.s. What value did this give for the speed of light? 3×10^8 meters/sec.
2. Roemer observed that the lapse of time between successive occultations of one of Jupiter's moons was seven minutes greater when the earth was receding from Jupiter than when it was approaching. The approximate period of the satellite being $42\frac{1}{2}$ hours, what was the speed of light given by these observations? Take the radius of the earth's orbit as 1.5×10^{11} meters. 2.2×10^8 meters/sec.
3. If, due to atmospheric refraction, the sun is visible for 15 minutes after it has actually set, what is its angular displacement? 4 degrees.
4. Assuming the rigidity of the ether to be approximately that of steel ($E = 8 \times 10^{10}$ newtons/square meter), what is the density of the ether in kilograms per square meter? 10^{-7}.
5. What percentages of error were involved in the second column of Ptolemy's table of refraction if the index of refraction of water is $\frac{4}{3}$? Errors up to 7 per cent.
6. Light incident on a block of glass at i degrees with the normal is refracted at r degrees. What is the refractive index μ? What is the velocity v per second of light in the glass?

i	r	μ	v
40	25	1.52	1.97×10^8
50	30	1.53	1.96×10^8
60	35	1.51	1.99×10^8
75	40	1.50	2.00×10^8

CHAPTER 27

Illumination

Modern Standards of Illumination

Until very recently artificial illumination was confined to providing the minimum amount of light that would enable man to continue the more essential activities normally carried on during the day. Artificial illumination competed merely with darkness, never with daylight. To imitate daylight, either in quality or quantity, was at first impossible and later prohibitively expensive on any scale large enough to be generally useful.

Within the last decade, however, new types of illuminant have provided a quality scarcely distinguishable from north-sky light, the best for general illumination. A campaign of encouragement has in addition brought many users to the point of providing a quantity of illumination rivaling and occasionally even exceeding that of daylight. This encouragement has not been entirely disinterested, to be sure, for the distributors of electric power have much to gain through general acceptance of the new standards of illumination. But the public also has much to gain, since the previous "starvation diet" of light has been taking its toll in inefficiency, accidents, and nervous ailments. There is scarcely a possibility that the public can be persuaded to provide more illumination than is good for it. In the meantime sound inducements for more illumination are being offered in this country in the form of electric rates which are much lower than those anywhere else in the world, as well as in improved lighting equipment at progressively lower cost. This form of propaganda by "the Utilities" need never cause any misgiving.

Luminous Intensity and Its Unit

Sources of "artificial" light vary greatly in their intensity. From night lamps, the luminous equivalents of a single candle, to beacon lights, whose luminous intensities may be expressed as the equivalent of millions of candles, the whole range of human requirements is covered. Moreover, a given source will invariably appear to be of different intensities when viewed from different angles. This effect is enhanced by reflectors designed to distribute light to the best advantage and becomes especially pronounced for such devices as headlights and searchlights. When the luminous intensity of a source is specified without qualification, it is either

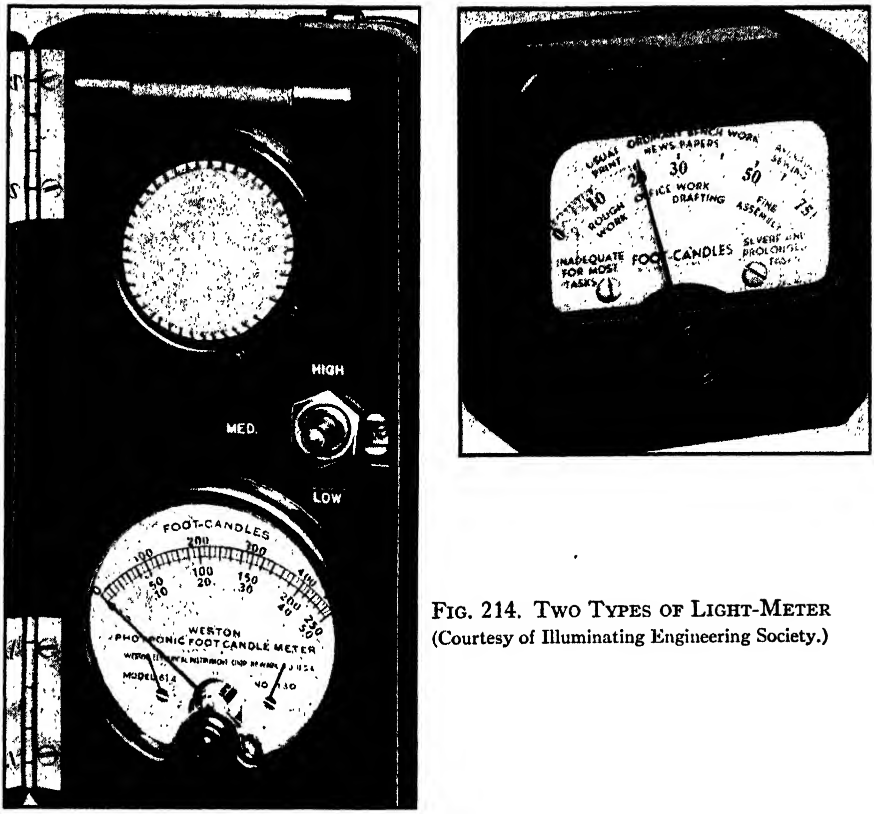


FIG. 214. TWO TYPES OF LIGHT-METER
(Courtesy of Illuminating Engineering Society.)

the average of all directions (*mean spherical candlepower*) or is limited by the context to a specific direction.

One of the few anachronisms of scientific terminology is the use of the term *candlepower* to describe the unit of luminous intensity. Only the name is anachronistic, however. The original unit was a candle, as the name indicates, of a particular kind, burning at a specified rate. But the most recent definition of the unit is a product of recent international agreement between England, France, and the United States, effective in 1941. This agreement was entered into before the onset of the European war, but there is no compelling reason why it should not be carried out, though for some reason Germany persistently refused to enter into it during the whole course of the negotiations.

The *new international candle* is defined as one sixtieth of the luminous intensity provided by one square centimeter of a "black body" at the temperature of melting platinum.[1] The term "black body" need cause no perplexity. It merely involves the requirement that the nature of the radiating surface shall be such that the surface appears perfectly black when

[1] *Revue de Métrologie Pratique et Légale* (2), *18*, 10 (1940).

cold. Perfect blackness in the sense of reflecting no visible light is, of course, an unrealizable ideal, but is one which can be approximated pretty closely by various artifices. It is a useful concept in the theory of radiation. In this instance perfect blackness must be invoked because the intensity of an incandescent surface at a given temperature depends somewhat on the degree of blackness of that surface when cold; the blacker it is when cold, the brighter it is when heated to incandescence.

Until a few years ago all illuminants were rated in candlepower. The more common sizes of electric light were procurable in 8, 16, or 32 candlepower sizes. Automobile lamp bulbs are still rated that way. All others, however, are rated, not in intensity, but in power consumption, that is, not in terms of luminous output but in terms of power input. While incandescent lamps dominated the lighting market this was a fairly convenient way of rating sources of light, for the luminous output was roughly proportional to power input. But with the advent of fluorescent lamps and other new types of illuminant, comparisons of intensity by power ratings have lost all significance. A twenty-watt fluorescent bulb gives several times as much light as a twenty-watt incandescent bulb.

Illumination and Its Units

Up to this point, the word "illumination" has been used in its usual loose, general sense. Science has commandeered it, however, as a technical term. As such it gives a numerical measure of the adequacy of the supply of light in the region where it is to be applied. It is, in a sense, of primary interest to the "consumer," whereas intensity is of primary interest to the "producer."

After light has left its source it is sometimes used directly, to attract attention, as with electric signs. But in the great majority of cases it is utilized indirectly, by its incidence on a surface, producing what is called technically *illumination*. Naturally, the illumination of a surface directly facing the source depends both on the luminous intensity of the source and on its distance from the source, being proportional to the former and inversely proportional to the square of the latter. The inverse square law is already familiar in gravitation and in sound and will appear again in connection with magnetic and electrostatic forces. It depends for its validity, as in the other cases, on the source's being concentrated at a point.

The English unit of illumination is the one most commonly used and is called the *foot-candle*. As the name indicates, it is the illumination at a surface one foot distant from a source of luminous intensity of one candlepower, the surface being normal to the source.

The metric unit is called the *lux*. It is similarly defined except that a distance of one meter is involved instead of one foot. The ratio of the two units is simply the square of the ratio between the meter and the foot or 10.76, the foot-candle representing the greater illumination. The lux will be the preferred unit in this chapter. If it becomes necessary to convert an illumination given in luxes to the more common English unit, foot-candles, the latter may be approximated by dividing the former by 10 (strictly 10.76).

As has already been intimated, standards of illumination have been rapidly rising. Until recently 50 luxes was considered sufficient illumination for sustained reading. Now 500 or even more are commonly provided, and the requirements are still rising. That this is not excessive is indicated by the fact that daylight illumination on the north side of a building on an average bright day is about 5000 luxes and in direct sunlight is about 100,000.

Measurements of illumination are commonly made by the "photovoltaic" type of light-meter, illustrated in Figure 214. They are more fully described on page 572 than is feasible here. They depend for their action on the fact that when certain "photovoltaic" materials are illuminated, they generate minute amounts of electric power. These tiny power impulses are made to energize electric meters, graduated to read in luxes or, more commonly, foot-candles.

Luminous Flux and Its Unit

The fact that the intensity of a source of light, measured in candlepower, may be quite different in different directions has already been mentioned. For some purposes the total amount of light emitted by an illuminant, the "flow" or *flux*, is of importance. Sometimes the intensity of a lamp is given in mean spherical candlepower, that is, the average of the intensities in all directions. In such a case the total flux could be said to be 4π times the mean spherical candlepower.

The factor 4π originates in the geometry of the sphere about the lamp as a center, and over which the illumination is supposed to be spread. The area of a sphere is 4π times the square of the radius. It is natural to take as the unit of luminous flux the flux from a source of unit mean spherical candlepower through one square meter of the surface of a sphere of one meter radius having the lamp at its center.[1] This is called a *lumen*. The area of the entire sphere will be 4π times the area through which the unit flux, as thus defined, passes. Consequently the total flux from a lamp of one mean spherical candlepower is 4π lumens.

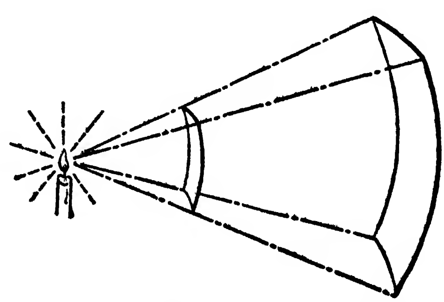


FIG. 215. LUMINOUS FLUX OVER UNIT SOLID ANGLE

[1] Though the lumen is defined here as a metric unit, there is no distinction between this and the English unit. The flux through one square foot of a sphere of one foot radius having the lamp as its center will obviously be the same as the flux described above. (See Fig. 215.)

The shortcomings of wattage as a measure of the luminous output of a lamp were noted above. It would be much more sensible to use luminous flux in this rôle. Indeed it would be eminently appropriate to do so. The present uninformative ratings of, say, two twenty-watt lamps, one incandescent and one fluorescent, would then be replaced by their flux ratings of 240 lumens and 800 lumens respectively. Unfortunately there is no prospect that such a change will occur in the near future.

There is, however, one respect in which the lumen is coming into common use. That is in rating the luminous efficiencies of lamps. Since the proper unit of luminous output is the lumen and of power input is the watt, the luminous efficiency of electric lamps is quite appropriately being specified in lumens per watt. There has been a steady increase in the efficiency of electric illumination since the time of Edison's first incandescent lamp. That lamp gave 1.4 lumens per watt. Tungsten filament lamps of comparable size give 12 lumens per watt and the new fluorescent lamps give 40. It is rather curious that the lumen, denied access to common use up to the present through lamp output ratings, seems to be entering by the back door of efficiency ratings.

Comparison of Luminous Intensities

Light flux is a form of power in the technical meaning of that term. Physicists would find their aesthetic sense gratified if only the practice had been established of measuring light in watts as mechanical and electrical power are measured and as rate of flow of sound energy is sometimes specified. But as with rates of energy flow in the form of heat, an arbitrary unit of light flux holds the field. For both heat and light the arbitrary basic units were established before their numerical relations with the absolute unit of power were known or, indeed, even imagined.

As with luminous flux, so with luminous intensity. Intensity, being simply flux per unit area, might well be specified in absolute units as watts per square meter instead of in candlepower. Since common practice does not sanction the absolute unit, intensities may only be compared with each other in terms of the purely arbitrary unit. The procedure of comparing intensities of sources is called *photometry* ("light measurement"), and the instruments commonly used for the purpose are *photometers*.

Until recently, almost all photometers involved the visual matching of illuminated areas. Though a human observer is very unreliable in any direct comparison of different intensities, he is fairly consistent in the ability to tell when two identical adjacent surfaces are of the same brightness when illuminated from the same kind of source. Visual photometers invariably utilize this ability, differing among themselves principally in the means taken to adjust to equality the illumination from the two sources.

The most common arrangement is to place the two sources at different distances until a match is secured. Another is to interpose a rotating sectored wheel in the path of the more intense illumination, adjusting the pro-

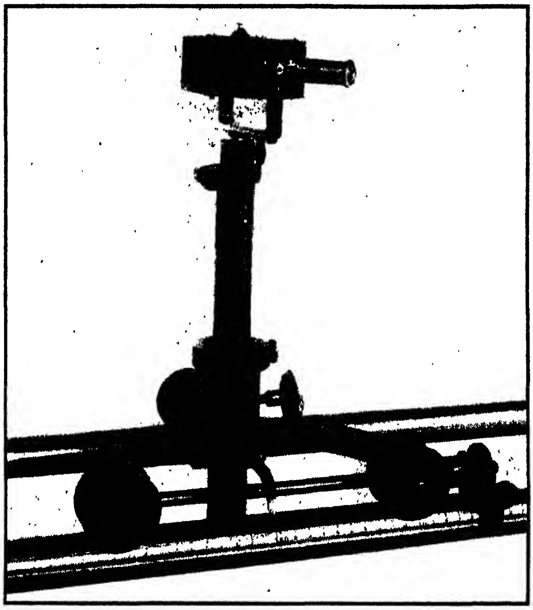


FIG. 216. A COMMON TYPE OF PHOTOMETER HEAD OF A VISUAL-MATCH TYPE

portion of the open to the closed sectors until a match is secured. Still another way is by the use of polarizers such as those described in the chapter on polarized light. The central feature in any visual-match type of photometer is the optical device for juxtaposing the illuminated surfaces so that both can be seen simultaneously. One arrangement for doing this, the so-called Lummer-Brodhun photometer head, is illustrated in Figure 216. The two lights being compared are off the scene, at measurable distances to right and left. A system of reflecting prisms in the central box enables the observer, upon applying his eye to the small oblique observing telescope, to see the opposite sides of the white screen receiving the light from the respective lamps, apparently side by side. The carriage is moved until a visual match is secured. Then application of the inverse square law — a questionable procedure unless the sources are small in comparison with the distances involved — enables the ratio of the intensities of the two sources in the given direction to be deduced.

The visual-match types of photometer have been rendered somewhat obsolete, however, by the photovoltaic type of light meter (page 572). Comparison of the pointer readings when the light meter is exposed to two illuminants successively, at the same distance, gives a direct measure of the relative intensities of the two sources. No adjustments to effect a match are required, as in visual photometers.

The replacement of the old visual-match types of photometer by the photovoltaic type brings its own problems, however. The chief problem is manufacturing photovoltaic surfaces which possess the same sensitivities as the human eye to the various colors making up the light being measured. The difficulty in solving that problem delayed the advent of the photovoltaic photometer for several years.

Comparison of Luminous Fluxes

The fact has already been pointed out that the usual illuminant distributes its flux quite non-uniformly. Its candlepower, therefore, is quite different when measured from different directions. It is desirable frequently to know the total flux emitted by a lamp, regardless of how that flux may be distributed. The preceding statements of luminous efficiencies of various types of lamp involve total flux rather than candlepower. The ordinary photometer, whether of the visual or the photovoltaic type, is not capable of yielding such information directly. In the purely theoretical case of a uniformly distributed flux, the candlepower in any direction multiplied by 4π (the area of the sphere of unit radius having the lamp at its center) yields the total flux. But since uniform distribution practically never obtains, no relation can be specified between the candlepower and the total flux from a given lamp. The value of the flux can be approximated in such a case by measuring the candlepower in a sufficient number of properly chosen directions, striking an average and then proceeding as in the case of uniform distribution of flux.

The indirection involved in the approximation just described may be avoided by the use of the so-called Ulbricht sphere. This is a hollow sphere, of large diameter in comparison with the longest dimension of the lamp being tested, coated on the inside with a special variety of highly diffusing white paint, and having set into it at one point a window of translucent glass. The lamp to be measured is placed inside of the sphere and between it and the window is placed a screen so that the direct rays from the lamp do not strike the window. The theory of the device shows that the brightness of the window is then directly proportional to the luminous flux of the lamp, provided that certain precautions are observed.

Whatever type of photometer is used, the intensity or the flux of a lamp under test becomes known only if that to which it is being compared is known. The basic standard, the "new international candle," has already been described. Its use is, however, necessarily confined to the principal standardizing laboratories. The common standard lamp is a specified variety of incandescent electric light, duly aged so that it has settled down to a constant performance, carefully guarded against mechanical and electrical misuse, and operated at a specified voltage. Such lamps, calibrated in terms of a basic standard, are securable from government standardizing agencies, in this country the United States Bureau of Standards.

The Automobile Headlight

Next to interior illumination, highway lighting is the most urgent requirement, at least in the United States. This is provided largely by flooding the roads ahead of the motor cars by lights on the cars themselves. Though this is far from an ideal solution of the problem of highway illumination, it is the only one feasible at present outside of populous areas. When adequate illumination is provided in this way, however, the glare is intolerable to approaching drivers. The central problem of automobile headlighting is how to get sufficient illumination without glare.

A solution to this problem which is in sight through the agency of polarized light is described in Chapter 35. It cannot be effected immediately, however, and in the meantime the less complete method now available is worthy of attention. Its principal defect is that it depends upon the initiative of drivers in momentarily pointing their headlight beams slightly downward out of consideration for the plight of other drivers approaching. Though only a simple flick of a conveniently located switch is involved, a disheartening proportion of those using the road are notably uncooperative in this respect.

The point of departure for automobile headlighting, as for other types of light projection, is the formation of a parallel beam. The standard device for converting the divergent light from a concentrated source into a substantially parallel beam in this connection is the so-called paraboloidal reflector. The principle is developed on page 392. For headlighting purposes the reflected beam from such a reflector is modified by its passage through an appropriately contoured cover glass. The beam is spread laterally by vertical flutes, and distributed vertically by small-angled prisms cast into the glass.

The depression of the beam for the convenience of approaching motorists is effected by a second filament in the headlight bulb, located slightly above the main filament. Figure 217 shows how this acts. It represents the vertical mid-section of a point source and paraboloidal reflector. The dotted lines show the normals to the surface at the points of incidence. The fact that the angle of reflection has the same value as the angle of incidence in each case will be evident.

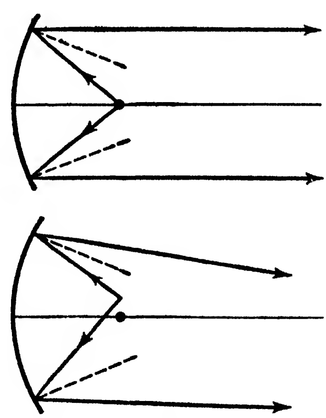


FIG. 217. THE EFFECT OF DISPLACING THE SOURCE ON A BEAM REFLECTED FROM A PARABOLOID

Some headlamps use the same artifice to deflect the beam to the right. To produce this, the second filament is mounted at the left of the main filament, as viewed from the driver's seat. Figure 217 may also be used to illustrate this effect by regarding it as a horizontal section instead of a vertical. Horizontal deflection, in headlamps which incorporate it, is invariably combined with depression. The second filament in these lamps is thus above and to the left of the main filament viewed from the rear.

Questions for Self-Examination

1. What does the term "candlepower" mean? Is the luminous intensity of a lamp increased by using a reflector?
2. Compare the *watt* with the *lumen* as units of luminous output of electric lamps.
3. How may the luminous output of lamps be compared with one another?
4. Describe the operation of the offset-filament type of automobile headlight.

Problems on Chapter 27

1. A 2-meter photometer of the intensity-matching variety shows equal illuminations of the screen when the lamps being compared are 75 centimeters and 125 centimeters distant respectively. If the 20-candlepower standard is the nearer of the two lamps, what is the candlepower of the other? Why would this type of photometer not give accurate results for frosted lamps? 56 cp.
2. On a line through two electric lights of candlepower 60 and 100 respectively, 10 feet apart, is a screen. Find the point between them where the two sides of the screen will have the same intensity of illumination. Find also the point, not between the two, where the illumination of one side of the screen is the same as that of each side before. 5.6 and 44 ft from the larger.
3. Three clerks have their desks side by side, 4 feet between centers. Eight feet above the center desk is a 500-candlepower lamp. What is the illumination at the center of each desk, taking obliquity into account? 7.8, 5.6 foot-candles.
4. If the desks of problem 3 are 4 feet $\times$ $2\frac{1}{2}$ feet, how many lumens does each desk receive, assuming the intensity at the center to be the average in each case? Does this assumption possess an equal degree of validity for each desk? Assume also that the flux from the lamp is uniformly distributed. 78 and 56.
5. A 100-watt incandescent lamp of the most efficient type delivers about 1200 lumens. The power equivalent of one lumen is about .0093 watt. What is the efficiency of the lamp? 11 per cent.

CHAPTER 28

Image Formation

The Pinhole as an Image-Forming Device

Common experience with photography and with picture projection, as in the cinema, has created an impression that for the production of images a complicated and more or less expensive equipment of lenses and optical accessories is required. While it is true that image-production is greatly facilitated by these devices, they are by no means indispensable. Faithful images may be formed, even photographs taken, with equipment no more pretentious than a diaphragm pierced with a small pinhole, backed by some sort of box or dark chamber (*camera obscura*) to block out stray light. Figure 218 is a photograph taken in this way. The great disadvantage in this type of photography is the faintness of the image, with consequent great length of exposure required. The accompanying photograph required twenty minutes' exposure in full sunlight.

Image formation by a pinhole depends on the circumstance that the angle subtended by such an aperture at an object whose image is being formed is invariably much smaller than the angle subtended by the object itself at the aperture. Though pinhole image formation has been on record ever since the time of Aristotle (7:*1*:333) the correct explanation for it was not forthcoming until 1575 (81:29), some two thousand years later.

FIG. 218. A PINHOLE PHOTOGRAPH

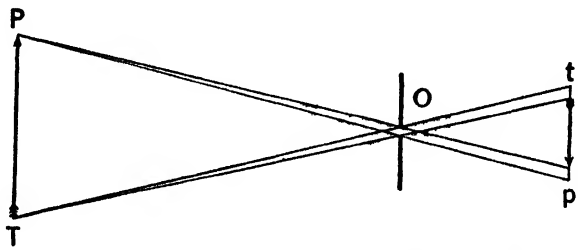


FIG. 219. THE PRINCIPLE OF PINHOLE IMAGE FORMATION

Figure 219 shows the principle. Light from the point *P* of the arrow constituting the object is confined to the area *p* on the screen and light from the tail *T* to *t*. Because the aperture subtends a small angle at the object, light from each extremity of the object (as well as for all points between) is spread out over a small area on the screen Consequently a pinhole image can never be sharp, but the sharpness can be increased up to a certain limit by diminishing the size of the aperture. This is, however, necessarily at the expense of brightness of the image. Tolerable sharpness requires tiny apertures and hence a very dim image results. This is the reason for the length of exposure required for the photograph of Figure 218.

Though the angle constituting the small cones of light radiating from each extremity of the object of Figure 219 has been greatly exaggerated in magnitude, it is still small in comparison with the angle *TOP* subtended by the object at the aperture. Since the latter angle is also equal to *top*, the angle subtended by the image at the aperture, it will be evident that the "spread" of the small cones of light must be very small indeed in comparison with the spread of the main cone spanning the entire image. That is, the angle subtended by the aperture at the object must be small in comparison to that subtended by the object at the aperture if an image is to be formed by a pinhole as was stated above.

The Birth of the Camera

The photographic camera is one of the simplest and most common types of image-forming instrument. The name comes from the old Latin term *camera obscura*, literally "dark chamber." These instruments were originally used as aids in sketching to form images which could be traced. Leonardo da Vinci presumably used such a device in the fifteenth century, for in one of his manuscripts is a drawing of a *camera obscura* (116:*2*:404) using a pinhole as the image-forming element. The first recorded use of a lens for this purpose was in 1599, in the writing of one Giambattista della Porta (99:363). Therein Porta invented the modern camera as far as the working principle of its optical system was concerned, though the chemistry of light-sensitive substances to record the images thus produced was not to come into existence for two and a half centuries.

How We Localize Objects

From every luminous point, emitted light spreads fanwise. We "see" these points because our eyes receive portions of the diverging cones of light. Very early in life we learn to interpret the complex sensations produced by the receipt of light and to infer the locations of the objects emitting it. Figure 220 represents a luminous point and the positions of two pairs of eyes that are regarding it. The cone of light entering the pupil of each eye is so narrow that it is pictured as a single ray. For the pair of eyes nearest the source, the degree of divergence of the two rays is greater than for the more distant pair. The nearer observer must "cross" his eyes more sharply to converge them upon the light. A different degree of muscular effort is involved in converging the eyes, corresponding to different distances. For normal persons possessing the use of both eyes, this muscular effort is translated into estimates of distance and is the chief source of visual information about distances of near-by objects. It enables the eyes to perform the function of rapid-fire surveying instruments, giving quick and accurate location of luminous points.

Common experience, however, deals, not with mere visible points, but with extended objects. Such objects are aggregates of infinite numbers of points and act accordingly in the process of vision. But to draw two or more diverging rays from many points of an extended object, such as the vertical arrow of Figure 221, would produce such a confusion of lines as to obscure the ideas that optical diagrams are intended to illustrate. Hence, it is usual to illustrate rays proceeding from only two points of an object, often the two extremities, as in this figure.

Moreover, it is not necessary that the objects under consideration shall be self-luminous, as might perhaps be assumed from the foregoing discussion. By far the greater proportion of the objects of our experience, on the contrary, become visible by light not originating in themselves, but reflected more or less directly from surrounding sources. From the standpoint of the optics involved, it is really a matter of indifference whether light originates in an object under observation or is only reflected from it.

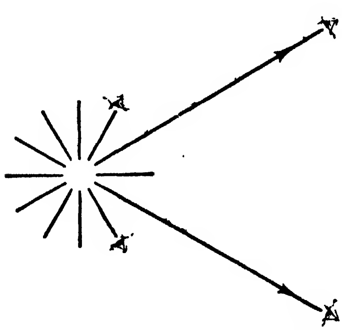


FIG. 220. HOW WE LOCALIZE A LUMINOUS POINT

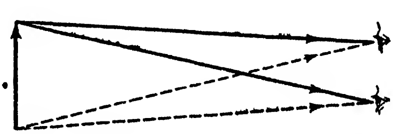


FIG. 221. VISUALIZING AN EXTENDED OBJECT

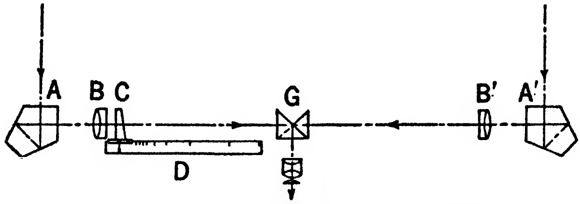


FIG. 222. OPTICAL SYSTEM OF COINCIDENCE RANGE FINDER

The Coincidence Range Finder

The precision with which objects can be located by binocular vision diminishes rapidly with increasing distance. The effort of convergence becomes so slight for distances of more than a few yards that, if it were not supplemented by other means of estimating distances, we should be continually reaching for distant objects just as babies do.

Convergence is greater the greater the distance between the two eyes of an observer. Accurate estimates of greater distances are facilitated by artificially increasing (in effect) this interpupillary distance. In *coincidence range finders* this "interpupillary distance" may be as great as seventy-five feet. The optical system of such a device is illustrated in Figure 222. It is used by artillerymen especially in naval practice. Smaller specimens are used as accessories to cameras to aid in focusing. Since convergence of the two lines of sight cannot be registered by muscular effort as in actual binocular vision, some other means must be used to correlate convergence with distance. A common method is to compensate the difference in direction of the two beams by interposing a prism in one of them. The optical system is so arranged that two portions of the field, displaced while the compensation is incomplete, are brought into coincidence (whence the name of the instrument) when the prism compensates the binocular angle. The distance of the target corresponding to that angle may then be read directly on a scale. Figure 223 shows the field of view of a coincidence range finder both before and after adjustment has been effected.

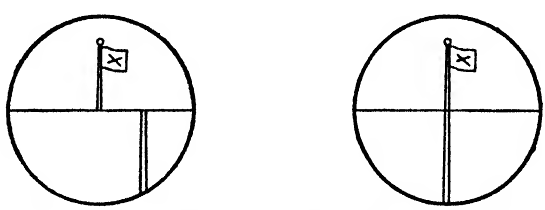


FIG. 223. FIELD OF VIEW OF COINCIDENCE RANGE FINDER

The Stereoscope

As one views a scene from different positions, the relative positions of foreground and background change. This is what produces the depth or *relief* characteristic of binocular vision. This relief can be provided in photographs, from which it is normally absent, by use of the *stereoscope.* This instrument requires two photographs side by side, one corresponding to the scene as it appears to each eye. Then, usually by prisms in the stereoscope, one eye of the observer is directed at one photograph and the other at the other, so that the binocular effect is secured just as if the two eyes were observing a single object directly from two different angles. The result is a rather striking illusion of depth. The stereoscope, in this form, is no longer commonly used, but it provided entertainment in the parlors of our grandparents. Amusement is now a mass-production enterprise, and the next step in stereoscopic vision will have to be taken in the field of moving-picture projection. This is now in a fair way to be realized. It is described on page 406.

How We See "Images"

Since perception of location is determined primarily by the directions from which light enters the eyes, the apparent location of an object is affected by modifying the course of the light as it proceeds from object to observer. Such modifications can occur in a number of ways, among the most common being through the agencies of mirrors and lenses. In a plane mirror, for example, the scene is completely transposed from where we know it to be. Of the light radiating from an object (Figure 224), two rays encounter a mirror at such an angle as to be reflected into the eyes of an observer. The angle of reflection for each of these rays is equal to the angle of incidence, a law which must have been known before men ever began to write. The divergence of these rays after reflection is such that they seem to have come along the dotted lines from a point as far behind the mirror as the object is in front of it. Light has not actually radiated from this point behind the mirror, it only appears to have done so because the eye has been trained to "see" an object at the point from which light appears to be diverging. This illusive reproduction of the object is what is termed an *image.* An extension of Figure 224 to cover the case of an extended object would show also that the image was the same size as the object upon reflection in a plane mirror.

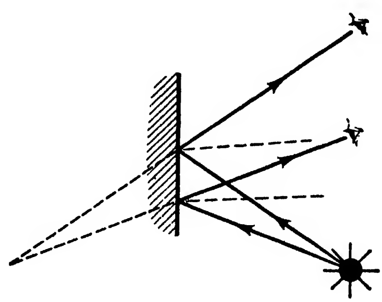


FIG. 224. DECEPTIVE LOCALIZATION BY REFLECTION

Image Formation by Spherical Mirrors

If the mirror is curved instead of plane, the same general principle applies, though with modifications. The divergence of the two rays entering the eyes is then affected by curvature of the mirror, so that the direction, distance, and size of the image is different than it would have been if seen in a plane mirror. Thus in the convex mirror of Figure 225, the increased divergence of the rays causes the image to seem closer than in a plane mirror. An extension of Figure 225 to cover the case of an extended object would show also that the image was smaller than the object, instead of being the same size, as in a plane mirror.

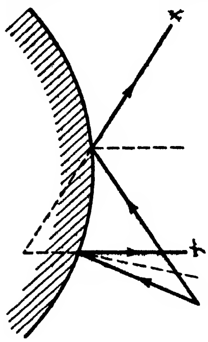


FIG. 225. IMAGE FORMATION BY CONVEX MIRROR

In a concave mirror, on the other hand, the decreased divergence of the reflected rays, as shown in Figure 226, causes the image to appear farther away than it would in a plane mirror, and the image to appear larger than the object.

Distinction Between Real and Virtual Images

Unlike the case of the convex and plane mirrors, however, the action of the concave mirror will be different for different distances of the object. Figure 226 applies to cases in which the object is nearer to the mirror than is the so-called *principal focus*. If the object is outside of the principal focus, a new phenomenon appears. In this case the divergence of the incident rays is not merely decreased by the reflection as before, but is actually reversed, the reflected rays becoming convergent (Figure 227). After passing through a focus the rays again diverge, and the observer then sees the image, not back of the mirror, as previously, but in front of it. In this case the light does not merely appear to have diverged from the image, as before; it *actually does* diverge. For this reason, this type of image is termed *real*, in contrast to the former type termed *virtual*. The images of real objects formed by plane and convex mirrors are always virtual, whereas the image formed by a concave mirror may be real or virtual, depending upon the distance of the object. It is sometimes of advantage to "catch" a real image on a screen, as is done in the cinema and by cameras. Virtual images cannot be so caught.

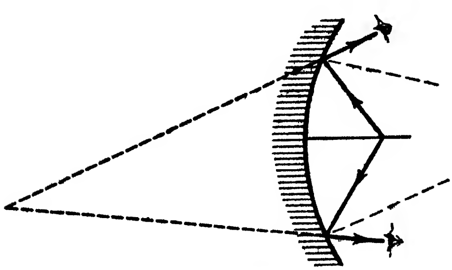


FIG. 226. VIRTUAL IMAGE BY CONCAVE MIRROR

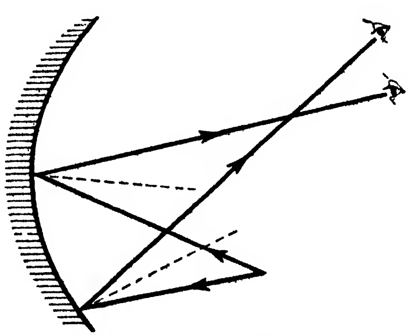


FIG. 227. REAL IMAGE BY CONCAVE MIRROR

An inspection of Figures 220 to 227 will show that the location of an image is determined by the intersection (real or virtual) of rays which originally diverged from the same point. In every case the rays have been shown ultimately entering the eyes of an observer. It should now be possible to realize that the presence or absence of an observer has nothing to do with the existence of the image. The condition for such existence is merely the intersection (whether real or virtual) of rays of light originally coming from the same point of the object. Henceforth the observer will be dispensed with in all diagrams, and optical problems will be considered solved when the appropriate intersections of rays have been effected.

The Principal Focus

One case of image formation is of particular importance — this is when the object is a great distance away. At any one point of such an object the mirror subtends a very small angle: so small that the rays reaching opposite edges of the mirror from that point are all but parallel. The plane in which parallel rays are brought to a focus is termed the ***principal focal plane*** of the mirror. The images formed by the mirror of all distant objects lie substantially in this plane. Thus the image of the sun formed by a concave mirror, as in the burning glasses of antiquity, lies in the principal focal plane. Objects which are so far away that their images lie substantially in the principal focal plane are said to be infinitely distant. The image of an infinitely distant point-object located on the axis will also be on the axis. This point is termed the *principal focus*.

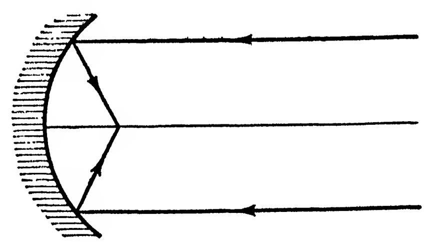


FIG. 228. IMAGE OF AN "INFINITELY DISTANT" OBJECT FORMED BY CONCAVE MIRROR

The advantage of bringing the images of distant and otherwise inaccessible objects within range in this way is, of course, that thereupon they may be viewed as minutely as desired. It is even possible to apply a magnifying glass to them, as one could do for an actual object close at hand. This combination of a real image at the principal focus of a mirror or lens with a magnifying glass applied to it constitutes the working principle of the telescope and of the compound microscope, both of which will be more fully discussed in Chapter 30.

Questions for Self-Examination

1. Describe the development of the idea of image formation by pinhole.
2. Trace the evolution of the optical features of the camera.
3. Describe how we localize objects by binocular vision.
4. Tell the principle of the coincidence range finder.
5. Show how to locate the image of an object formed by (*a*) a plane mirror, (*b*) a convex mirror, (*c*) a concave mirror; the last for a distant object and a near-by object.
6. Distinguish between real and virtual images and give an example of the formation of each with the aid of a diagram.

Problems on Chapter 28

1. If cameras had pinholes instead of lenses, would they have to be focused?
2. Where would be the image if a point-object were placed at the center of curvature of a spherical mirror? At the principal focus?
3. Would light, initially parallel, then reflected successively from a convex mirror and a concave mirror of the same radius, necessarily be parallel again? Explain.
4. Light converging immediately after reflection forms a real image; diverging, a virtual image. Light from a (real) object diverges as it strikes a mirror. What could be the meaning of the term "virtual object"?

CHAPTER 29

Spherical Reflecting and Refracting Surfaces

The Principal Focus of the Concave Mirror

A ray of light, parallel to the axis, strikes a concave spherical mirror as shown in Figure 229. The angle of incidence (that is, the angle with the normal) is α, and the angle of reflection, therefore, also has the value α. Since the normal makes an angle α with the incident ray, it must also make an angle α with the axis, to which the incident ray is parallel by hypothesis. Again, since the reflected ray makes an angle 2 α with the incident ray, it must also make an angle 2 α with the axis to which the incident ray is parallel. But the angle α between the normal and the axis has the value in radians of arc/radius, and the angle 2 α between the reflected ray and the axis has the approximate value arc[1]/focus. Therefore,

$$\frac{a}{r} = \alpha, \text{ and } \frac{a}{f} = 2\,\alpha, \therefore f = \frac{r}{2}. \tag{1}$$

Thus the principal focus of a spherical mirror lies on the axis midway between the mirror and its center of curvature, and the principal focal length is, therefore, half the radius of curvature. If the distant object lies off the axis, the image, while not now *at* the focus, will be found in or very near to the focal plane, that is, the plane through the focus perpendicular to the axis, so that equation (1) still applies (Fig. 230).

[1] Strictly, the arc a pertains only to the angle α, having the radius r. An arc having the radius f would be of slightly different length, but for mirrors ordinarily used in optical systems the angle α is very small and the difference is negligible.

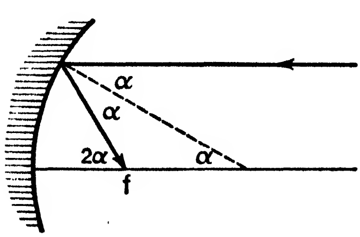


FIG. 229. THE POSITION OF THE PRINCIPAL FOCUS

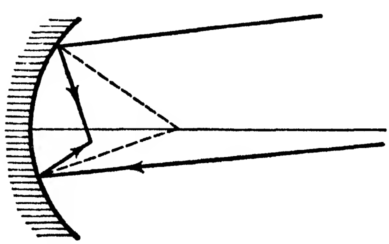


FIG. 230. THE POSITION OF THE OFF-AXIS IMAGE

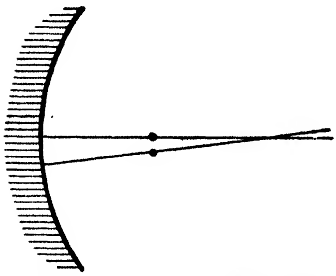


Fig. 231. The Concept of the Focal Plane

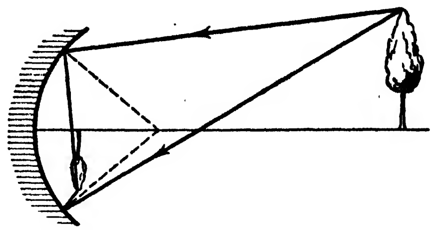


Fig. 232. The Image in the Focal Plane

This may be made evident by imagining the mirror rotated about its center of curvature until the axis is parallel to the rays from the point in question (Fig. 231). As long as the angle of such rotation is small, it will be evident that the image will be very close to the original focal plane.

This condition, or more accurately, the corresponding condition for image formation by a lens, is realized in focusing a camera on a distant large object, such, for example, as a tree. Let a tree (Fig. 232) be far enough away from a concave mirror so that the two extreme rays from any point of it, such as the tip, are nearly parallel and, hence, converge after reflection almost exactly in the principal focal plane. This does not mean that rays from another point, such as the base, will be parallel to those from the tip. If they were, no image would be formed. All rays would converge to the same point, a type of action of which no mirror or lens is capable, popular opinion to the contrary notwithstanding. Instead all the rays from the base are converged at a different point from those originating at the tip, though still in the focal plane. Thus rays from the tip and the base are separately converged, and the corresponding action of light originating at each point between results in the formation of an image of the tree.

The Object-Image Relation

If the object, while still at some distance from the mirror, is close enough so that the extreme rays from any point of it are no longer nearly parallel, the image will not be in the principal focal plane. To find the position of the image, assume a point object O (Fig. 233) on the axis of a mirror having the point C as its center of curvature. Let the distance of the object from the mirror be u, that of the image be v, and the radius of curvature of the mirror be r. Now note that

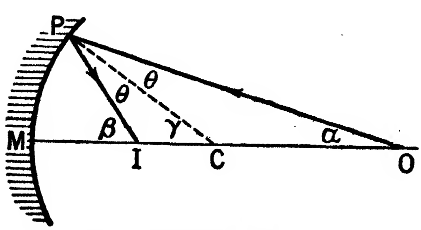


Fig. 233. The Geometry of the Object-Image Relation

$$\beta = \theta + \gamma \text{ and } \gamma = \theta + \alpha. \quad (2)$$

Eliminating θ between these two equations,

$$\alpha + \beta = 2\gamma. \tag{3}$$

But as in the consideration of Figure 229, for the small angles usually involved,

$$\alpha \doteq \frac{a}{u}, \beta \doteq \frac{a}{v}, \text{ and } \gamma = \frac{a}{r}. \tag{4}$$

Hence, substituting these values in equation (3),

$$\frac{1}{u} + \frac{1}{v} = \frac{2}{r} \text{ or, by virtue of equation (1), } \frac{1}{u} + \frac{1}{v} = \frac{1}{f}. \tag{5}$$

Equation (5) makes it possible to find the distance v of the image from the mirror, knowing that of the object u and the radius r of the mirror. It will be immediately evident that the first two of equations (4) are approximations and that consequently equation (5) is not rigorously correct. It is, however, so nearly correct that it is very useful in almost all practical cases. The closeness of approximation is greater as the angles assume smaller and smaller values.

As is readily seen from Figure 232, object and image are interchangeable. This means that a clear image of I could be seen at O, as well as a clear image of O at I. At each point the image of an object at the other would be "in focus." The two points are termed *conjugate foci* of the mirror. There are obviously an infinite number of pairs of such conjugate foci for a given mirror.

If, in the first form of equation (5), u is assigned a large value, corresponding to a very distant object, the value of $1/u$ becomes very small, approaching zero as u increases. The corresponding distance of the image is termed the principal focal length, f, of the mirror. Making these substitutions in equation (5),

$$0 + \frac{1}{f} = \frac{2}{r} \text{ or } f = \frac{r}{2}.$$

This is equation (1). It thus appears that the values ∞ and $r/2$ for u and v respectively (or the reverse) are simply one of the infinite number of pairs of conjugal focal lengths of a mirror.

Robert Smith's Graphical Method for Reflection

It is often convenient to be able to make a quick graphical determination of the location of an optical image. The simplest way of doing this was devised by Robert Smith, one of Newton's early successors at Cambridge University. It has already been observed that the image formed by a mirror is located at whatever point the rays intersect after reflection (page 305). It is always possible to trace two rays by making the angles of reflection equal to the angles of incidence and thus to establish the point

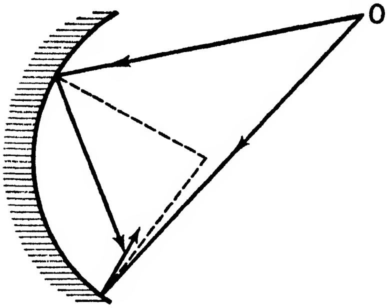


FIG. 234. LOCATING THE IMAGE BY ANGLES OF INCIDENCE AND REFLECTION

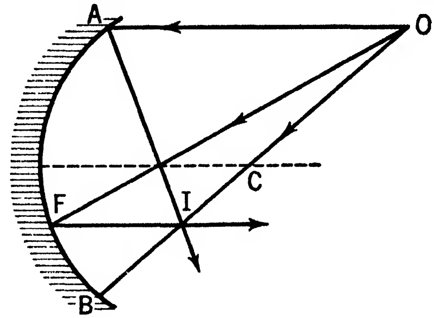


FIG. 235. LOCATING THE IMAGE BY ROBERT SMITH'S CONSTRUCTION

of intersection and, hence, to locate the image (Fig. 234). The method is laborious, however, and Smith simplified it immensely by taking advantage of some very elementary facts. Of the infinite number of rays radiated from the object, three of them are particularly easy to trace (Fig. 235):

(1) The ray *OAI* from the object which is parallel to the axis is reflected through the principal focus.

(2) The ray *OCI* which passes through the center of curvature is reflected back along itself.

(3) The ray *OFI* which passes through the principal focus before striking the mirror is reflected parallel to the axis.

Any two of these rays may be used to make a quick graphical determination of the location of an image. The laying-out of angles, the laborious part of ray-tracing, is entirely eliminated. Smith used only the first two of the foregoing rays. The third is a later addition, equally useful, both as a check and under certain circumstances as a substitute for one of the preceding.

Lateral Magnification

The relation of the size of the image to that of the object, termed the *lateral magnification*, is simply the relation of the corresponding distances from the mirror, as may be seen by reference to Figure 236. Suppose the image of the point *O* of an arrow is found to be located at *I*. One of the rays forming this image could be *OAI*, where the angles of incidence and reflection, α, are of course equal. Because of the equality of these angles, and the consequent similarity of the two triangles,

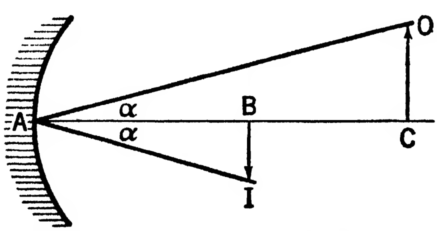


FIG. 236. RELATION OF THE IMAGE SIZE TO THE OBJECT SIZE

$$\frac{BI}{CO} = \frac{AB}{AC}.$$

But $\frac{BI}{CO}$ is the lateral magnification (to be termed m) and $\frac{AB}{AC}$ is the relation of the distance of the image from the mirror to that of the object. For the former the notation v has been adopted, and for the latter u. Designating also the inversion of the image by a negative sign (whence the lateral magnification for an erect image would be +),

$$m = -\frac{v}{u}. \tag{6}$$

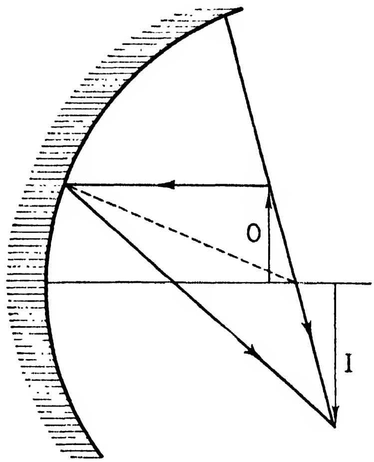


Fig. 237. Lateral Magnification of a Spherical Mirror

Suppose, for example, that an object is 10 inches distant from a concave mirror of focal length 6 inches. From equation (5)

$$\frac{1}{10} + \frac{1}{v} = \frac{1}{6}, \text{ or } v = 15.$$

From equation (6)

$$m = -\tfrac{15}{10} = -1\tfrac{1}{2}.$$

This is illustrated in Figure 237, O being the object and I the image. The positions of the object and image might have been interchanged, being conjugate foci of the mirror, whereupon, applying the same two equations,

$$\frac{1}{15} + \frac{1}{v} = \frac{1}{6}, \text{ or } v = 10;$$

$$m = -\tfrac{10}{15} \text{ or } -\tfrac{2}{3}.$$

For this case the image is closer to the mirror and smaller than the object, and still inverted. This is always the case whenever the object is outside of the center of curvature of a concave mirror, while the image is farther from the mirror and larger if the object is between the center of curvature and the principal focus.

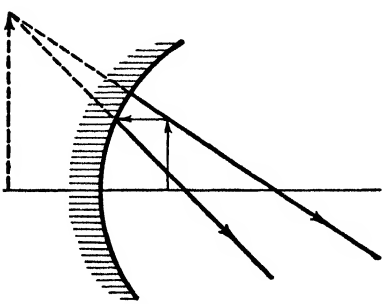


Fig. 238. A Virtual Image by a Concave Mirror

Virtual Images by Concave Mirror

The preceding treatment of the concave mirror has involved only cases in which *the object was "outside"* (farther from the mirror than) the principal focus. An important case is that in which the object is "inside" of the principal focus. It is illustrated in Figure 238. It will be noted that in this case the reflected rays do not converge to a focus as before. On the contrary, they diverge. Hence, as described in the preceding chapter (pages 305–06), the image is virtual instead of real. This case can be dealt with in the same way as that for a more distant object.

Consider the case of an object 4 inches away from a concave mirror of focal length 6 inches. Substituting in the second form of equation (5),

$$\frac{1}{u}+\frac{1}{v}=\frac{1}{f}; \frac{1}{4}+\frac{1}{v}=\frac{1}{6}, \text{ whence } v = -12.$$

A comparison of this result with Figure 238 indicates that the significance of the negative sign of v is that the image lies behind the mirror and, hence, is virtual. The image is also larger than the object, the lateral magnification being, from equation (6),

$$m = +\tfrac{12}{4} = 3.$$

The positive sign of m is associated with the fact that the image is erect.

Thus a concave mirror may produce either a real or a virtual image, real if the object lies outside of the principal focus, virtual if the object lies inside of the principal focus. The virtual image is always larger than the object for the concave mirror. Dentists, having to use mirrors of necessity, often use concave mirrors in just this way. Shaving mirrors are sometimes made concave, with a large radius of curvature, for the same reason.

The Convex Mirror

The action of a convex mirror lends itself to the same analysis as has just been made of that of the concave mirror. The formation of the image of a point source by such a mirror was described in the preceding chapter (page 306). Figure 239 shows graphically the case of an extended object. The image is virtual, erect, and smaller than the object. This is always true of the image of any real object formed by a convex mirror, a fact which renders the treatment of the convex mirror simpler than that of the concave.

Equation (5) may be applied to this case by regarding the radius of curvature and the principal focal length as negative. Thus for the numerical values $r = -12$ (and, hence, $f = -6$) and $u = 30$,

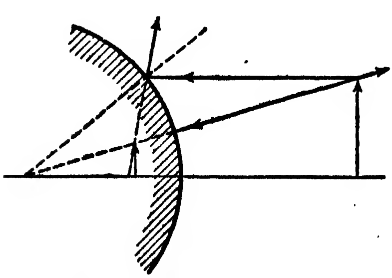


FIG. 239. THE IMAGE PRODUCED BY A CONVEX MIRROR

$$\frac{1}{30}+\frac{1}{v} = -\frac{1}{6}; \text{ whence } v = -5$$

and from equation (6),

$$m = -\frac{-5}{30} = +\frac{1}{6}.$$

Convex mirrors are not very widely used. Their principal utility is as rear-view mirrors for automobiles, but motorists seem to prefer plane mirrors for that purpose, notwithstanding the limited field of view which plane mirrors provide. Convex mirrors have, however, been recognized as image-forming devices for a long time. Euclid treated them at length in his *Catoptrics*, telling in proposition 21 that the images are smaller than the objects, and Ptolemy experimentally identified their "virtual foci" in the second century A.D. (100:69). Convex mirrors have been treated here principally for the sake of completeness of the outline and as a simple introduction to some optical principles.

Object-Image Relation for Refraction; Object in Air

To trace a ray from a point object O in the air to its image I in glass of index of refraction μ, the surface of the glass being the section of a sphere of radius r, note that in Figure 240, by definition of index of refraction, $\mu = \frac{\sin\theta}{\sin\phi}$. Limiting the case to that of angles small enough so that it can be said without appreciable error that $\frac{\sin\theta}{\sin\phi} \doteq \frac{\theta}{\phi}$, then $\mu = \frac{\theta}{\phi}$. (7)

But from the figure

$$\theta = \alpha + \gamma, \text{ and} \tag{8}$$

$$\phi + \beta = \gamma, \text{ or } \phi = \gamma - \beta. \tag{9}$$

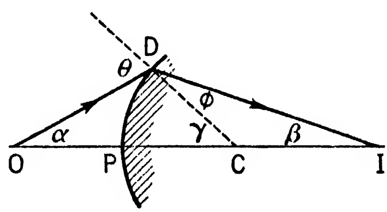


FIG. 240. THE IMAGE PRODUCED BY A SINGLE REFRACTING SURFACE (Object in air.)

Substitute (8) and (9) in (7).

$$\mu = \frac{\alpha + \gamma}{\gamma - \beta}, \text{ whence } \alpha + \mu\beta = (\mu - 1)\gamma. \tag{10}$$

Because of the smallness of the angles to which this treatment is limited, and of the corollary that the arc PD is not appreciably different from the perpendicular distance to D from the axis, substitutions may be made in equation (10) as follows, using u and v for object distance and image distance respectively as before:

$$\frac{\overline{PD}}{u} + \mu\frac{\overline{PD}}{v} = (\mu - 1)\frac{\overline{PD}}{r}$$

or

$$\frac{1}{u} + \frac{\mu}{v} = \frac{\mu - 1}{r}. \tag{11}$$

This is the object-image relation for a spherical refracting surface corresponding to equation (5) for a spherical reflecting surface. If the index of refraction μ and the radius r of the spherical surface are known, the image distance v can be computed for any given object distance u with the aid of equation (11). This relation, in a somewhat different form, was discovered by Huygens in 1703 (73:55).

A spherical refracting surface has two principal focal lengths instead of one, as does the mirror. Let u approach infinity; then, the corresponding value of v will be found to be

$$f_\mu = \frac{\mu r}{\mu - 1}, \tag{12}$$

where f_μ indicates the principal focal length in glass of index μ. This relation, in a somewhat different form, was discovered by Kepler in 1611. It is really a special case of Huygens' relation, though discovered a century earlier (130:*1*:204). Letting v approach infinity, the corresponding value of u will be found to be

$$f_1 = \frac{r}{\mu - 1}, \tag{13}$$

where f_1 indicates the principal focal length in air of index unity. Equations (12) and (13) correspond to equation (1) for spherical mirrors.

Graphical Method for Refraction

A graphical method of locating the image corresponding to any object distance is thus suggested. This method corresponds to that of Figure 235 for a mirror and, like that case, is due to Robert Smith.

(1) The ray *OAI* from the object which is parallel to the axis is refracted through the principal focus within the glass.
(2) The ray *OCI* which is directed toward the center of curvature passes through undeviated.
(3) The ray *OFI* which passes through the principal focus in air is refracted parallel to the axis.

Any two of these rays may be used to make a quick graphical determination of the location of an image. As in the case of the mirror, Smith used only the first two rays, *OAI* and *OCI*.

Lateral Magnification

The lateral magnification of a spherical refracting surface may be deduced by reference to Figure 242. With the image *IB* of the object *OA* located by the graphical method or with the help of equation (11), draw the ray *AP* to the ***vertex*** (as the intersection of the axis with the optical sur-

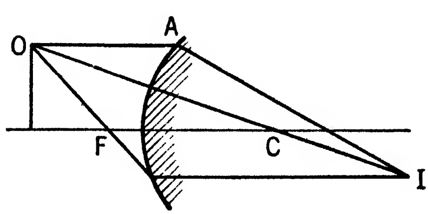


FIG. 241. THE IMAGE PRODUCED BY REFRACTION; ROBERT SMITH'S CONSTRUCTION

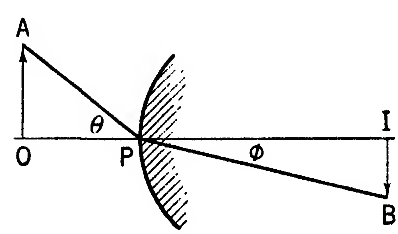


FIG. 242. THE LATERAL MAGNIFICATION OF A SPHERICAL REFRACTING SURFACE

face is termed) and then continue it to B, the tip of the image as already located. The angle ϕ will then be the angle of refraction corresponding to θ as the angle of incidence. The two angles are related by Snel's law (page 290), which in this case may be approximated as $\mu = \frac{\theta}{\phi}$, and, therefore, $\frac{\phi}{\theta} = \frac{1}{\mu}$, taking advantage once more of the smallness of the angles. Therefore, by the same type of approximation

$$\frac{IB}{v} = \phi \quad \text{and} \quad \frac{OA}{u} = \theta,$$

whence the magnification m is given by

$$m = -\frac{IB}{OA} = -\frac{\phi v}{\theta u} = -\frac{1}{\mu}\frac{v}{u}. \tag{14}$$

Equation (14) is the expression for the magnification of a spherical refracting surface corresponding to equation (6) for a spherical mirror. The minus sign is attached, as in equation (5), to indicate the inversion of the image.

To illustrate the use of equations (11) and (14), Tables 1 and 2 show object and image distances produced by a converging surface, the index of refraction being 1.5. The radius of curvature is +4 in Table 1 and −4 in Table 2. The student will compute the corresponding values of v and m. The two focal lengths will be seen by equations (12) and (13) to have the values 12 and 8 respectively. The significance of the negative signs will not be hard to interpret. To clarify them it is suggested that each case be represented by a simple diagram, the position and relative size of object and image being indicated. General rules on the sign convention will be found on pages 318–19.

u	v	m
∞	12	0
24	18	$-\frac{1}{2}$
16	24	-1
12	36	-2
10	60	-4
8	∞	∞
6	-36	$+4$
4	-12	$+2$
0	0	$+1$

TABLE 1

Object-Image Relation for Refraction; Object in Glass

Another case arises immediately out of the above, that is, its exact reversal — the object being within the glass. The object-image relation could be derived in the same way as equation (11) was, but it will be simpler merely to interchange u with v and μ with 1 in equation (11). From this comes the relation

$$\frac{\mu}{u} + \frac{1}{v} = \frac{1-\mu}{r}. \tag{15}$$

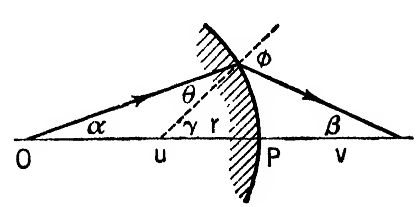


FIG. 243. THE IMAGE PRODUCED BY A SINGLE REFRACTING SURFACE
(Object in glass.)

Also, the equation being derived in the same way that equation (14) was,

$$m = -\mu \frac{v}{u}. \tag{16}$$

It will be evident that the expressions for the principal focal lengths for this case, corresponding to equations (12) and (13) for the case of the object in air, are

$$f_1 = -\frac{r}{\mu - 1} \tag{17}$$

and

$$f_\mu = -\frac{\mu r}{\mu - 1}. \tag{18}$$

For this case, as illustrated in Figure 243, the radius of curvature must be considered negative (see below).

u	v	m
∞	8	0
36	12	$-\frac{1}{2}$
24	16	-1
18	24	-2
15	40	-4
12	∞	∞
10	-40	$+6$
8	-16	$+3$
6	-8	$+2$
2	$-1\frac{3}{5}$	$+1\frac{1}{5}$
0	0	$+1$

TABLE 2

A table of image distances, v, and magnifications, m, for various object distances, u, in this case is appended. As in the previous case, the values should be checked and located on diagrams to get a clear picture of the action of such a refracting surface. At first thought the case of the object being embedded in an optically dense medium may seem rather bizarre. One has only to recall, however, that these conditions are fulfilled by anything embedded in glass or immersed in a globe of water or by a plano-convex lens lying, flat side down, on a printed page which is to be magnified.

Sign Conventions

Equations (11) and (15) apply only to the case of a convex glass surface. They may be made to apply equally to concave surfaces, by establishing some simple conventions as to sign. Unfortunately, complete uniformity on the matter of sign conventions in optics is far from realized, but the following, which are perhaps as common as any, possess some advantages of simplicity and consistency.

1. The radius of curvature r of a spherical refracting surface will be considered positive when the surface is convex to the incident light and negative when concave to the incident light. The four possible cases are illustrated in Figure 244. The first two will be recognized as having been discussed, the first being covered by equation (11) and the second by equation (15). The same two equations may be made to apply without change to the concave surfaces represented in the third and fourth cases respectively

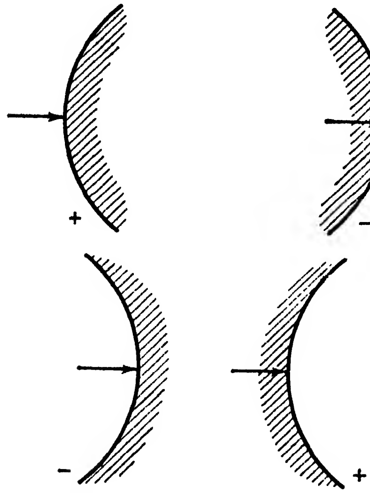


FIG. 244. SIGN CONVENTION FOR RADIUS OF CURVATURE
(Arrow indicates direction of light.)

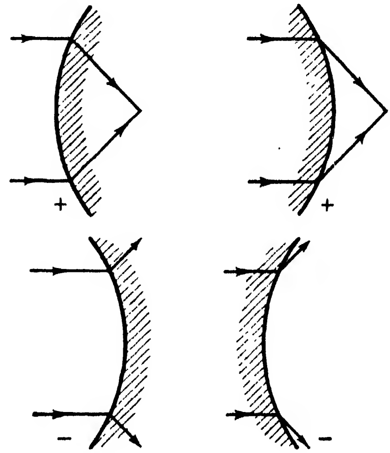


FIG. 245. SIGN CONVENTION FOR PRINCIPAL FOCAL LENGTHS

by giving the appropriate sign to the numerical value of r which is substituted in these equations.

2. The two focal lengths f of a spherical refracting surface will be considered positive when a parallel beam is rendered convergent in crossing the surface and negative when a parallel beam is rendered divergent (Fig. 245).

3. Image distances v will be measured from the vertex of the refracting surface and will be considered positive when in the direction that the light is traveling and negative when in the opposite direction (Fig. 246). This has the effect of placing real images at positive distances from their respective surfaces, and virtual images at negative distances. Both cases are in evidence in Table 1 and Table 2.

4. Object distances u will be measured from the vertex of the refracting surface and will be considered positive when in the direction opposite to that in which the light is traveling and negative when in the same direction as the light is traveling (see Tables 3 and 4). Negative values occur frequently when two or more successive refracting surfaces are involved. When u is negative the object is said to be virtual. The distinction between real and virtual objects appears in Figure 247. Some optical accessory, such as an auxiliary lens or refracting surface, which is not shown, is clearly required to produce the converging beam associated with a virtual object.

Image formation by concave refracting surfaces is illustrated in Tables 3 and 4. The first table is for an object in air, the second for an object embedded in glass (compare the third and fourth cases illustrated in Figure 245). For both cases, the index of refraction μ is 1.5 as for Tables 1 and 2.

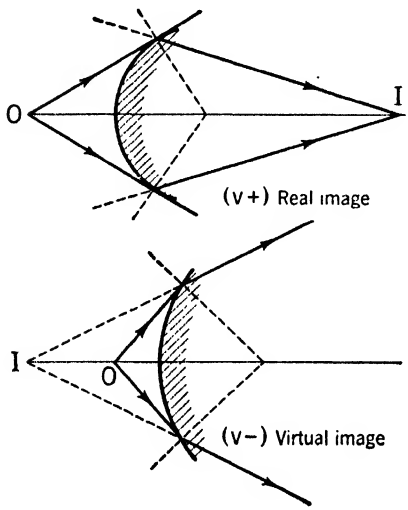


FIG. 246. SIGN CONVENTION FOR IMAGE DISTANCES

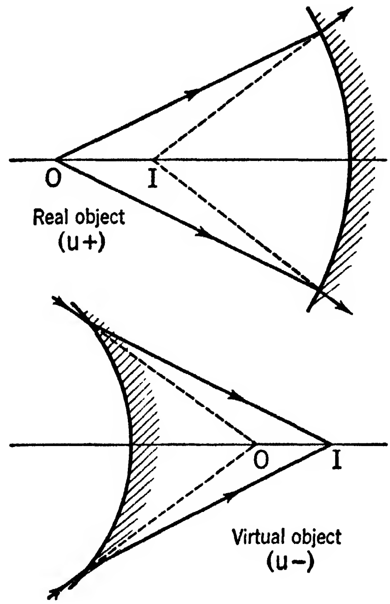


FIG. 247. SIGN CONVENTION FOR OBJECT DISTANCES

The radius of curvature is taken as − 4 and + 4 in the respective tables. Negative values of u are included. In both cases, a scale diagram should be made of the surface, the object, and the image. Table 3 is computed from equations (11) and (14), and Table 4 from equations (15) and (16).

u	v	m
∞	− 12	0
8	− 6	$+\frac{1}{2}$
4	− 4	$+\frac{2}{3}$
0	0	+ 1
− 2	+ 4	$+\frac{4}{3}$
− 4	+ 12	+ 2
− 8	∞	∞
− 16	− 24	− 1
− 24	− 18	$-\frac{1}{2}$

TABLE 3

u	v	m
∞	− 8	0
12	− 4	$+\frac{1}{2}$
0	0	+ 1
− 4	+ 4	$+\frac{3}{2}$
− 6	+ 8	+ 2
− 12	∞	∞
− 18	− 24	− 2
− 24	− 16	− 1
− 36	− 12	$-\frac{1}{2}$

TABLE 4

Questions for Self-Examination

1. What is a "principal focus" and what were the first recorded recognitions of it?
2. Clarify the common but incorrect statement that all rays from an "infinitely distant" object are parallel.
3. Describe Robert Smith's simplified method of locating images formed by a concave mirror and by a convex mirror.

4. An object is 15 inches away from a concave mirror of focal length 6 inches. Calculate the image distance. Also make a sketch showing how to locate the image graphically. Repeat when the object distance is 4.
5. What does a negative sign mean when applied to (*a*) focal length, (*b*) magnification, (*c*) image distance, (*d*) object distance, (*e*) radius of curvature?

Problems on Chapter 29

1. Light is incident on a water surface at 60° with the vertical. What is the angle of refraction? For water $\mu = 4/3$. 40°.
2. Through what angle is light deviated when incident upon water obliquely at 45°?
3. How deep is a tank of water which is estimated at 4 feet viewed directly from above? 5.3 ft.
4. A microscope is focused on a paper surface. A piece of plate glass is interposed. To bring the paper back into focus, the microscope is raised 4 millimeters. The upper surface of the plate is then brought into focus by raising the microscope another 6 millimeters. What is the index of refraction of the glass? 1.7.
5. If light has 50,000 wave-lengths to the inch in air, how many does it have in water? 67,000.
6. An object is placed u centimeters from a concave mirror of radius r centimeters. What is the distance v in centimeters of the image from the mirror? What is the magnification β? Locate the image on a diagram.

u	r	v	β
24	12	8	$-\frac{1}{3}$
12	12	12	-1
4	12	-12	3
0	12	0	1

7. An object is placed u centimeters from a convex mirror of radius r centimeters. What is the distance v in centimeters of the image from the mirror? What is the magnification β? Locate the image on a diagram.

u	r	v	β
24	-12	$-4\frac{4}{5}$	$\frac{1}{5}$
12	-12	-4	$\frac{1}{3}$
4	-12	$-2\frac{2}{5}$	$\frac{3}{5}$
0	-12	0	1

8. See page 317, Table 1.
9. See page 318, Table 2.
10. See page 320, Table 3.
11. See page 320, Table 4.

CHAPTER 30

Simple Lenses and Their Aberrations

The Development of Lenses

In southern Europe and in the Orient a peculiar type of bean known as the *lentil* is cultivated. Instead of being kidney-shaped, as are most of the beans we know, it is more like a flattened and sharp-edged doorknob. Its shape resembles a convex-sided lens enough so that it need not surprise us that the English word *lens* is simply the Latin word for lentil.

It is impossible to say who made the first lenses, which were presumably used as burning-glasses. In any case, the first comprehension of the way in which lenses act, the point of principal present importance, did not come until modern times. The first treatment of the subject was at the hands of Alhazen in the eleventh century, though it was clumsy and very incomplete. This treatment was considerably improved by Roger Bacon in the thirteenth century, especially as regards single refracting surfaces. Kepler in the early seventeenth century dealt with plano-convex and equibiconvex lenses almost in modern terms (130:*1*:203). The growth of comprehension of the action of lenses was rapid from that time on and reached substantially its present state within a century of Kepler's time.

Image Formation by Lenses

In the treatment of spherical reflecting and refracting surfaces in the preceding chapter, the principal objective was the development of relations between the object distances, u, the image distances, v, and the radii of curvature, r. These relations took the forms of equations (5), (11), and (15) of Chapter 29. There are similar relations for lenses. Perhaps the most useful of these is the *object-image relation,*

$$\frac{1}{u}+\frac{1}{v}=\frac{1}{f}, \tag{1}$$

an equation which has its exact counterpart in the second form of equation (5) of the preceding chapter, applying to spherical mirrors. The simplest way to derive this equation is to utilize once more Robert Smith's method of locating an image (Fig. 248). As with single reflecting or refracting surfaces, an incident ray initially parallel to the axis will, after leaving the lens, pass through a point on the axis which is termed the principal focus, the ray $OAFI$ of Figure 248.

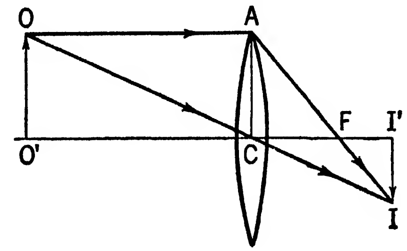


FIG. 248. THE GEOMETRY OF THE OBJECT-IMAGE RELATION FOR A CONVERGING LENS

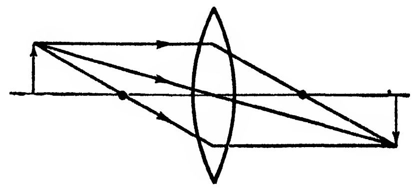


FIG. 249. ROBERT SMITH'S CONSTRUCTION FOR A CONVERGING LENS

Under certain conditions, one of which is that all angles involved are small, a ray through the center of the lens will not be deviated. Hence, *OCI* will qualifiedly be regarded as a straight line. By pairs of similar triangles,

$$\frac{I'I}{O'O} = \frac{CI'}{CO'}. \quad \text{Also} \quad \frac{I'I}{CA} = \frac{I'I}{O'O} = \frac{FI'}{CF},$$

whence

$$\frac{CI'}{CO'} = \frac{FI'}{CF}, \text{ or } \frac{v}{u} = \frac{v-f}{f},$$

from which comes by simple algebra

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f},$$

which is equation (1) above.

A lens has two focal lengths, one on each side, just as does a spherical refracting surface. But for the lens the two focal lengths are equal as long as the mediums on both sides of it are the same. For the spherical refracting surface, the ratio of its two focal lengths was simply the ratio of the refractive indices of the mediums on the two sides of the surface. In a sense this is also true of the lens as ordinarily used, the medium on each side being air.

In the first proportion of the foregoing derivation, namely $\frac{I'I}{O'O} = \frac{CI'}{CO'}$, the left-hand fraction represents the ratio of the size of the image to the size of the object and therefore represents the lateral magnification m. Substituting the values for the right-hand fraction, the equation becomes

$$m = -\frac{v}{u}. \tag{2}$$

The negative sign is interpolated to describe the inversion of the image when both v and u are positive.

Thus both the object-image relation given in equation (1) and the expres-

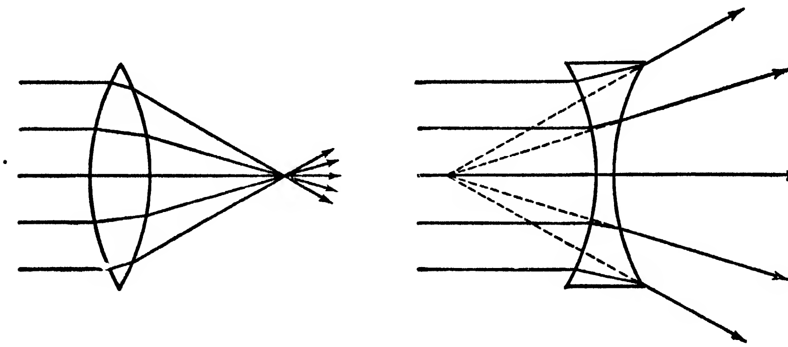


FIG. 250

ACTION OF CONVERGING LENS ACTION OF DIVERGING LENS

sion for the magnification given in equation (2) are the same for the converging lens as for the concave mirror. For the lens, as for the mirror — and, indeed, as for the spherical refracting surface as well — the same relations may be applied to other forms, due regard being given to the sign conventions stated on pages 318 ff. The two principal types of lens, corresponding to concave and convex mirrors, are the converging and the diverging types. The names arise from the effect on a beam of light consisting initially of parallel rays. The shape of converging lenses is characterized by greater thickness at the center than at the edges, and that of diverging lenses by greater thickness at the edges than at the center. A parallel beam is brought to a point by a converging lens and spread, as though it were diverging from a point, by a diverging lens, as shown in Figure 250.

In treating the concave mirror, two cases were emphasized, namely, that in which the object was outside of the principal focus and that in which it was inside. The same two cases exist in the action of the converging lens.

Object distance from first lens	Intermediate image distance from first lens	First magnification	Intermediate image distance from second lens	Final image distance from second lens	Second magnification	Over-all magnification
∞	+ 3	—	+ 3	− $\frac{9}{4}$	+ $\frac{3}{4}$	—
+ 12	+ 4	− $\frac{1}{3}$	+ 2	− $\frac{18}{11}$	+ $\frac{9}{11}$	− $\frac{3}{11}$
+ $7\frac{1}{2}$	+ 5	− $\frac{2}{3}$	+ 1	− $\frac{9}{10}$	+ $\frac{9}{10}$	− $\frac{3}{5}$
+ 6	+ 6	− 1	0	0	+ 1	− 1
+ 5	+ $7\frac{1}{2}$	− $1\frac{1}{2}$	− $1\frac{1}{2}$	+ $\frac{9}{5}$	+ $\frac{6}{5}$	− $\frac{9}{5}$
+ 4	+ 12	− 3	− 6	+ 18	+ 3	− 9
+ $3\frac{3}{4}$	+ 15	− 4	− 9	∞	—	—
+ $3\frac{1}{2}$	+ 21	− 6	− 15	− $22\frac{1}{2}$	− $1\frac{1}{2}$	+ 9
+ $3\frac{1}{4}$	+ 39	− 12	− 33	− $12\frac{3}{8}$	− $\frac{3}{8}$	+ $\frac{9}{2}$
+ 3	∞	—	∞	− 9	—	—
+ $2\frac{1}{2}$	− 15	+ 6	+ 21	− $6\frac{3}{10}$	+ $\frac{3}{10}$	+ $\frac{9}{5}$
+ $1\frac{1}{2}$	− 3	+ 2	+ 9	− $4\frac{1}{2}$	+ $\frac{1}{2}$	+ 1
0	0	+ 1	+ 6	− $3\frac{3}{5}$	+ $\frac{3}{5}$	+ $\frac{3}{5}$
− 3	+ $1\frac{1}{2}$	+ $\frac{1}{2}$	+ $4\frac{1}{2}$	− 3	+ $\frac{2}{3}$	+ $\frac{1}{3}$
− 12	+ $2\frac{2}{5}$	+ $\frac{1}{5}$	+ $3\frac{3}{5}$	− $2\frac{4}{7}$	+ $\frac{5}{7}$	+ $\frac{1}{7}$

It could, in fact, be illustrated by the identical numerical examples that were invoked in the case of the concave mirror. To avoid repetition, an example will be given which involves both a converging and a diverging lens acting together. The image formed by the converging lens is to be regarded as the object for the diverging lens. The example requires that the location of the final image formed by the second (diverging) lens be found for each position of an object in front of the first (converging) lens. The two lenses have respective focal lengths + *3* and − *9* and are separated by a distance *6*.

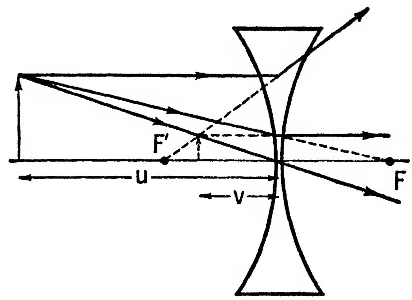


FIG. 251. ROBERT SMITH'S CONSTRUCTION FOR A DIVERGING LENS

The values in the accompanying table should all be checked by calculation, using only the object distances in the first column as given, together with the focal lengths and separation of the lenses. Interpretation of the results will be facilitated by locating the object and the two images on a separate diagram to scale for each row of calculated values. If any difficulty is encountered in interpreting the negative values of object distance found in the first and fourth columns, see page 319.

Shapes of Lenses

It will be evident that a given focal length of a lens can be secured in an indefinite number of ways. Not only can index of refraction be varied by proper choices of glass, but even for a given quality of glass, a lens can have a great variety of shapes, all producing the same focal length. Lenses are almost invariably bounded by spherical (or sometimes plane) surfaces, a plane being really a portion of a spherical surface of infinite radius. Thus, converging lenses may be formed by any of the combinations of spherical surfaces shown in Figure 252. The focal length might be the same in every case, notwithstanding the differences in shape. Similarly diverging lenses may be formed by any of the combinations of spherical surfaces shown in Figure 253.

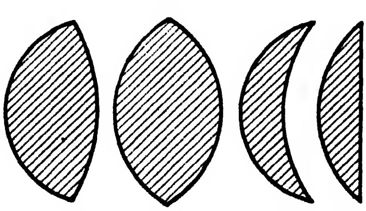


FIG. 252. SHAPES OF CONVERGING LENS

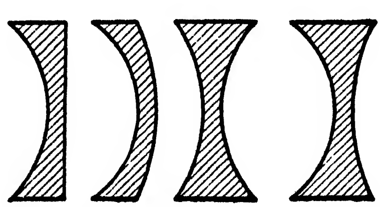


FIG. 253. SHAPES OF DIVERGING LENS

The Lens-Maker's Equation

Little has been said up to this point as to any relation between focal lengths of lenses, their shapes, and the qualities of glass of which they are made. For concave mirrors a very simple relation between focal length and radius of curvature was found (page 309). Similarly, for spherical refracting surfaces a less simple pair of equations was established, giving focal length in terms of radius of curvature and index of refraction (pages 316, 318). A similar relation will now be sought for lenses.

In 1693, Edmund Halley, Newton's friend and patron, derived a relation that has proven to be exceedingly useful.[1] With its aid the focal length f of a lens can be computed if the radii of curvature r_1 and r_2 and index of refraction μ of the glass are known. It is customarily stated in the form

$$\frac{1}{f} = (\mu - 1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right). \tag{3}$$

This is commonly called the *lens-maker's equation* for obvious reasons. It applies only to a "thin" lens, that is, a lens whose thickness is sufficiently small in comparison with its radii of curvature and object and image distances so that it may be disregarded. This is a justifiable approximation for most lenses. There is a corresponding equation for lenses that are not "thin," but it is less simple and will not be treated here.

With the aid of equation (3) an optician can determine the radii to which the surface of a lens may be ground to produce a required focal length. It applies to lenses in the same way that equation (1) of the preceding chapter did to spherical mirrors and that equations (12), (13), (17), and (18) of the same chapter did to spherical refracting surfaces. It has already been pointed out that a lens of any given focal length may possess any one of many different shapes. There is an indefinite number of combinations of r_1, r_2, and μ which will produce the same value of f. The great utility of the choice of shapes thus available to the manufacturing optician will be pointed out in another connection.

Two special cases of the lens-maker's equation had been developed by Kepler nearly a century before (65:123). He had observed that light was so bent when it passed through glass at angles less than 30° from perpendicularity to the surface, that angles of refraction were approximately two thirds of the corresponding angles of incidence. This foreshadowed the modern concept of index of refraction and in effect assigned to the refractive index of glass an approximate value $\frac{3}{2}$. Kepler then deduced geometrically that the focal length of an equibiconvex lens was equal to the common radius of curvature and that for a plano-convex lens it was equal to twice the radius of the spherical side. Both of these discoveries of Kepler will be seen to be embodied in the lens-maker's equation. Assigning to μ the value $\frac{3}{2}$ and setting $r_2 = -r_1$ for the equibiconvex lens, the equation

[1] *Philosophical Transactions* (abridged), *3*, 593 (1693).

becomes $f = r_1$. If $r_2 = \infty$ is similarly set for the second face of a plano-convex lens, the equation becomes $f = 2\,r_1$.

For equibiconcave and for plano-concave lenses the same approximate values of the focal lengths will be found, though of course with negative signs.

Derivation of the Lens-Maker's Equation

Halley derived his equation somewhat as follows. A lens is obviously a combination of two spherical refracting surfaces acting in succession. Therefore, the course of a ray acted on by a lens may be followed by using equations (11) and (15) of the preceding chapter successively. If the object is infinitely distant, the corresponding image distance will represent the principal focal length of the lens. In that case equation (12) replaces equation (11). This is the procedure that will be followed in deriving the lens-maker's equation.

The image of an infinitely distant object (see Fig. 254) as formed by the first surface would be at a distance

$$\frac{\mu r_1}{\mu - 1}$$

by equation (12) of Chapter 29, if the second surface did not intervene. This expression represents the distance (negative, of course) of the intermediate image from the second surface, except for the deduction to be made for the thickness of the lens. By limiting the derivation to the case of a "thin" lens, this distance may be taken as the object distance for the second surface.

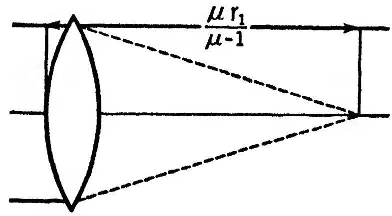


FIG. 254. THE LENS REGARDED AS TWO SUCCESSIVE REFRACTING SURFACES

Accordingly, the rays incident on the second surface of the lens of Figure 254 strike it as though they originated from a point at a distance.

$$-\frac{\mu r_1}{\mu - 1}$$

Substitute this for u in equation (15) of Chapter 29, and also substitute f for v, since the final image will be that of an infinitely distant object. Then

$$\frac{\mu}{-\dfrac{\mu r_1}{\mu - 1}} + \frac{1}{f} = \frac{1-\mu}{r_2},$$

r_2 being the radius of the second surface and f the focal length of the lens. Hence,

$$f = \frac{r_1 r_2}{(\mu - 1)\ (r_2 - r_1)},$$

or as more commonly stated,

$$\frac{1}{f} = (\mu - 1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right),$$

which is the lens-maker's equation.

The Limitation to Thin Lenses

The limitation of the lens-maker's equation to "thin" lenses has been commented upon. It would be necessary to modify the equation somewhat for lenses whose thickness was not negligible. While these modifications would not be at all difficult to introduce, it does not seem appropriate to do it here. (See 121:25.) Halley's comments in this connection may be of interest.[1] He says:

> But if you are so curious to consider the Thickness (t) (which is seldom worth accounting for), in the case of parallel Rays falling on a Plano-Convex of Glass, if the plane-side be toward the Object, t does occasion no Difference, but the Focal Distance $f = 2\,r$. But when the Convex Side is toward the Object, it is contracted to $2\,r - \frac{2}{3}t$; so that the Focus is nearer by $\frac{2}{3}t$. If the Lens be Double Convex, the Difference is less; if a Meniscus, greater. If the Convexity on both sides be equal, the Focal Length is about $\frac{1}{6}t$ shorter than when $t = 0$. In a Meniscus, the Concave side towards the Object increases the Focal Length, but the Convex towards the Object diminishes it.

Disregarding the thickness of the lens — as is commonly done to simplify the treatment of it — thus introduces a limitation to the accuracy of the resulting theory. Whether this theory fits the facts for a given lens used in a given way depends on whether the assumption of negligibility of thickness is satisfied for those particular conditions. If it is not, corrections must be introduced.

Aberrations

The element of approximation introduced into lens theory through confining attention to "thin" lenses is typical of a whole series of similar approximations, introduced either tacitly or explicitly at various stages of the foregoing treatment of mirrors, of spherical refracting surfaces, and of lenses alike. Besides the element of thickness (applying only to lenses) some other particularly important types of approximation have been made at various points in the foregoing treatment. On account of these, lenses always fall short of the "perfect" performance described by the equations constituting their theory. These discrepancies between the theory and the actual performance of uncorrected lenses are termed *aberrations*. For some reason, the corrections which are made necessary in consequence of disregarding the thickness of lenses are not customarily included in the list of aberrations. Aside from these, the most important are *spherical aberra-*

[1] *Philosophical Transactions* (abridged), *3*, 598 (1693).

tion, astigmatism, chromatic aberration, distortion, and ***curvature.*** The first two and the last two can be described now, but the third must await the development of the concepts of the following chapter.

On page 309 (footnote), 311 (eqns. 4), 315 (eqns. 7), 317, and 323, it was specified that consideration was being limited to small angles. This condition permeates the whole structure of geometrical optics. It consists, in fact, of two conditions instead of one. For one, if a spherical surface involved in any optical system subtends more than a few degrees at its center, imperfections are produced in the image which are said to be due to spherical aberration. For the other, if any of the rays from the object subtend more than a few degrees at a spherical surface, the resulting imperfections in the image are of a different type from those attributed to spherical aberration and are said to be due to astigmatism. The word astigmatism is derived from *a*, a Latin negative prefix, and *stigma*, a *spot* or *point.* Therefore, the word astigmatism implies that light originating at any one point of the object fails to converge to a point in the image. The term is not at all descriptive and fails to give much of any inkling as to the essential nature of astigmatism or as to how it differs from spherical aberration. One may say that spherical aberration originates in the angular magnitude of the spherical surfaces involved and that astigmatism originates in obliquity of the rays passing through those surfaces.[1] The nature of these two aberrations is now to be described.

Spherical Aberration

Spherical aberration will be considered principally with reference to the spherical reflector. It exists also in spherical refracting surfaces and in lenses, but in these its nature and the means of correcting it cannot be described quite as simply as with mirrors, though spherical aberration is the same in principle for all these cases. A brief consideration of the reflection from a concave mirror of large angular aperture will show that rays reflected from the edge do not pass anywhere near the "focus" toward which the central rays converge (Fig. 255). This effect was first commented upon in 1575 by Franciscus Maurolycus — both for reflection and refraction by spherical surfaces (81:39, 65). The first clear analysis was made by Huygens a century later (130:*1*:341).

The curve to which the reflected rays are tangent is called a *caustic.* Its cusp is the focus, to which the light from the central part of the mirror is reflected. Figure 255 treats only the case of parallel incident light, as Huygens did. It shows half of an axial cross-section of the caustic curve. The entire cross-section is frequently visible in the reflection of sunlight from the cylindrical interior of polished dishes, for example, as on the sur-

[1] The so-called *astigmatism* of the human eye, although it manifests itself in a somewhat similar way, is really an entirely different effect and should possess a different name. It exists even at points on the axis, unlike true astigmatism, and is produced by the existence of different radii of curvature in different radial planes of the refracting surface.

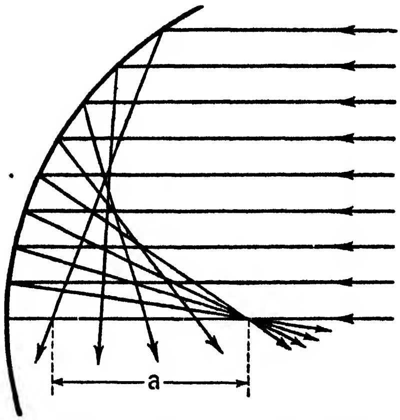


FIG. 255. A CAUSTIC BY REFLECTION

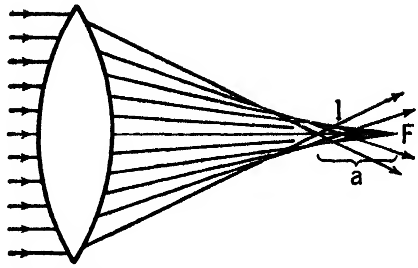


FIG. 256. A CAUSTIC BY REFRACTION

face of a tumbler of milk when the sunlight is incident almost horizontally. The complete "curve" is, of course, three-dimensional and has somewhat the form of the stem end of an apple with the stem removed. The distance a (Fig. 255) from the focus to the intersection of the marginal rays with the axis, divided by the focal length is a measure of the spherical aberration, or more properly, of the *axial spherical aberration* of the mirror. The same effect is common in lenses, being illustrated in Figure 256. It will be evident that the sharpest image will not be found at F, the point to which the central rays converge, for the image of a point would there be spread into a diffuse disk due to the diverging marginal rays. The radius of this disk of illumination would diminish as the lens is approached until the point l is passed, after which it would again increase. At l is said to exist the *circle of least confusion*, the radius of which, divided by the focal length of the lens, is a measure of the *lateral spherical aberration*. The axial spherical aberration a/f is evident in the figure.

The correction of spherical aberration, in either mirrors or lenses, is fairly simple. For mirrors, especially, the method of eliminating it has been known for many centuries. One point, however, should be emphasized. Spherical aberration can be corrected only for *one* object distance. A mirror or lens in which spherical aberration is eliminated for one particular distance of the object will still show spherical aberration for other object distances. To eliminate it for these other distances would require other mirrors or lenses, *a different one for each distance*. Fortunately this stricture is not as serious as might appear at first sight. All telescopes, for example, and most cameras are intended to focus on what are virtually infinitely distant objects. The elimination of spherical aberration for parallel rays will, therefore, meet all the requirements of these types of optical device.

The Parabolic Reflector

In the case of the mirror, reference to Figure 257 will show that by de-forming the margin of a spherical mirror outward through the proper angle, the marginal rays could be diverted to the focus. It has long been known that if a concave mirror possessed a form such that all incident parallel rays would pass through the same point after reflection, its cross-section would possess the form of the familiar curve known as a parabola. The first clear demonstration of this fact known in scientific literature is in a fragment written about the middle of the sixth century by one Anthemius of Tralles (25:10; 130:*1*:34). There are ill-founded traditions that both Archimedes and Claudius Ptolemy had done the same thing, ten centuries and four centuries earlier respectively. Either or both may have done so, for such of their writings as have survived indicate that they were easily capable of such an accomplishment. But the evidence that they actually did it is not strong enough to compel acceptance of the tradition at present (130: *1*:46).

Five centuries after Anthemius, that is, in the eleventh century, Alhazen wrote a whole treatise on parabolic mirrors.[1] Alhazen is known to have taken some of his other cues from Ptolemy, hence the fact that he treated parabolic mirrors extensively creates a mild presumption that Ptolemy may have written on the same subject. It is not known whether Alhazen was acquainted with the work of Anthemius. In 1278 this treatise of Alhazen was translated from Arabic into Latin by Wilhelm von Moerbeck, Archbishop of Corinth (25: 26), and from that time on knowledge of the parabolic mirror was more or less public property.

Obviously, a mirror which brings parallel light to a focus will convert the light from a point source at the focus into a parallel beam after reflection. The parabolic mirror is used very commonly in this way for searchlight reflectors and in automobile headlamps. But even a reflector which is perfect as far as spherical aberration is concerned cannot give a perfectly parallel beam because the source of light is not rigorously a point. The finite size of any practical source, such as a carbon arc or an incandescent filament, produces divergence in the reflected beam which is in strict proportion to the angle subtended at the reflector by the source of light.

Lenses may be shaped to aspherical surfaces to cause them to act in the same way as parabolic reflectors, but a more common way to make this correction is by combining lenses which have equal and opposite spherical aberrations. Lenses corrected for spherical aberration are termed ***aplanatic*** from Greek roots meaning "not spreading."

Astigmatism

The second principal aberration, astigmatism — except when it is due to defects in the image-forming system, as in the human eye — is produced by oblique incidence of light on the mirror or lens. It is like spherical aberration in that there is no one distance for which all rays from one point

[1] Not to be confused with Alhazen's larger work, the *Opticae Thesaurus*, which did not become generally known until much later than his work on parabolic mirrors.

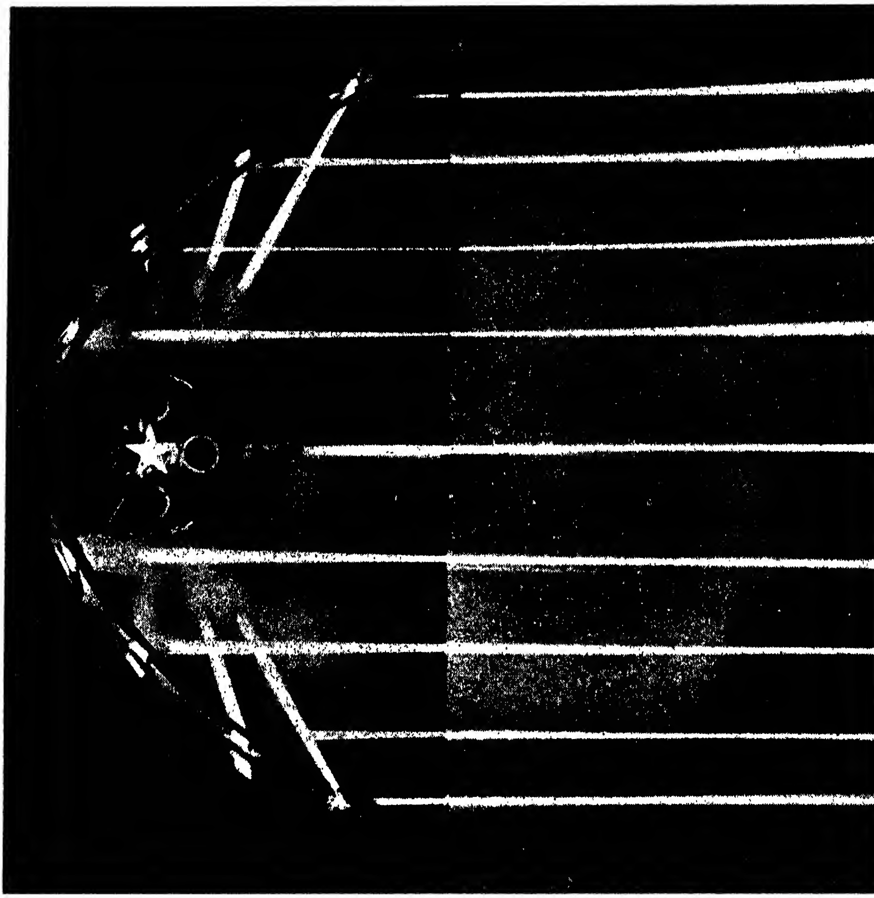


FIG. 257. REFLECTION BY SMALL PLANE MIRRORS ARRANGED ALONG A PARABOLA

of the object converge to a single point in the astigmatic image. Indeed, astigmatism is sometimes classified as one manifestation of a sort of generalized spherical aberration. But it is quite unlike spherical aberration (in the more usual restricted meaning of that term) in the pattern formed by the errant rays in their failure to converge to a point. This pattern may be visualized by different astigmatic patterns of a circle of luminous points (Fig. 258). In the first drawing the adjustment of focus has been made for what would correspond in spherical aberration to the circle of least confusion (page 330). Each point thus appears as a small disk. In the case of *spherical* aberration this disk becomes larger for any change of focus, whether an increase or a decrease. But the case is different for astigmatism, as the succeeding drawings show. Increasing the focal adjustment elongates the images radially and at the same time shrinks them in the other dimension to mere lines. On the other hand, decreasing the focal

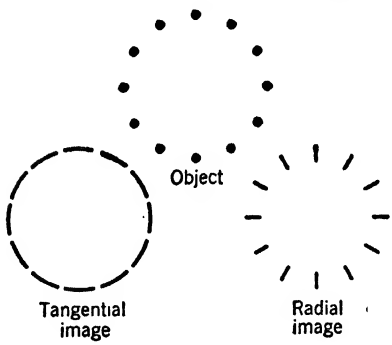


FIG. 258. THE TWO ASTIGMATIC IMAGES

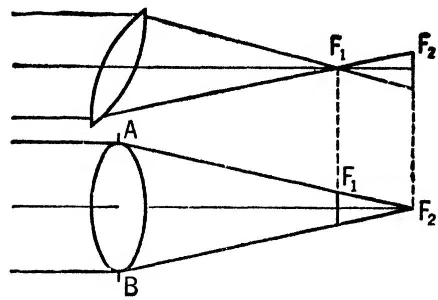


FIG. 259. ANOTHER MANIFESTATION OF ASTIGMATISM

adjustment elongates the images tangentially and at the same time shrinks them in the other dimension to mere lines at right angles to those of the preceding case. Thus astigmatic foci consist of two lines at right angles to each other and in different planes. One of the lines lies in the plane of incidence — as a radial plane containing the axis and an incident ray is called — and the other perpendicular to it.

Figure 259 illustrates astigmatism in a slightly different guise. The plane of incidence is the plane of the paper in the upper figure (which is the usual practice in optical diagrams). The astigmatic focal line, corresponding to the tangential lines of Figure 258, is represented end-on by the point F_1. The lower figure is the same arrangement as the upper, but viewed as the upper would appear if looked at from the top of the page. The lens is turned somewhat about the axis AB. The astigmatic focus corresponding to the radial lines of Figure 258 is represented end-on by the point F_2. Again it will be seen that the two focal lines are at right angles to each other and in different planes.

From the above it will be evident that if the object should consist of a right-angled mesh instead of a point, the vertical lines of the mesh would

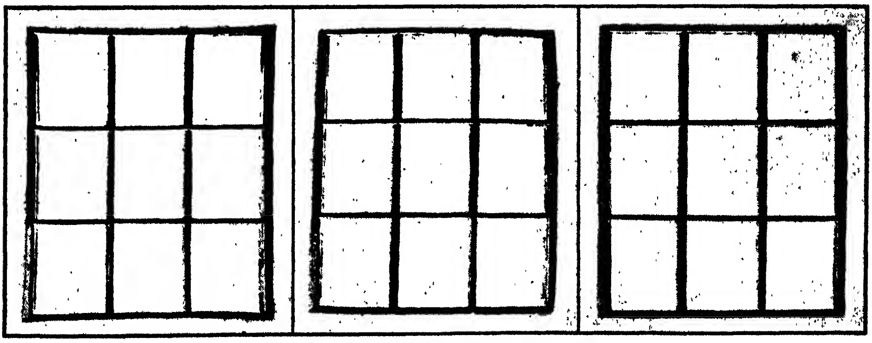


FIG. 260. DISTORTION
(Courtesy of the Eastman Kodak Research Laboratories.)

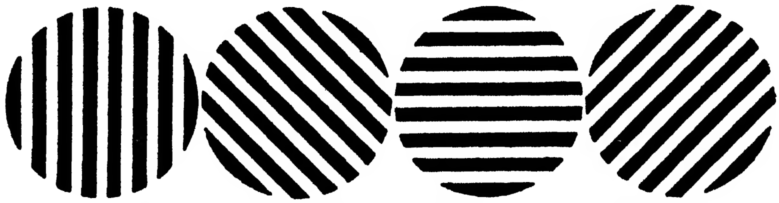


FIG. 261. A TEST OBJECT FOR VISUAL ASTIGMATISM
(The lines should all look equally black to a person of normal vision.)

be brought to a focus in a different plane than would the horizontal lines. Hence, a strongly astigmatic photograph of such a mesh could be in focus for one set of lines or for the other, but not for both at once. Similarly, in the astigmatic image of groups of parallel lines (Fig. 261), not all groups are in focus at once. This is the regular test for astigmatism of the eye, a test which anyone can make for himself. For this case, as has already been remarked, the astigmatism is produced by non-uniformity of the radii of curvature — usually in the cornea — in different radial planes instead of by obliquity of passage through a lens. The effect is much the same, however. The correction of visual astigmatism consists simply of cylindrical spectacle-lenses which will change the focal length for one of the astigmatic images to bring it to the same value as the other.

The reduction of astigmatism in lenses, unlike its correction in the human eye, is a matter of some difficulty. Only in recent years has the reduction become common enough so that other cameras besides the most expensive may be equipped with so-called anastigmat ("not astigmatic") lenses. It is not possible to develop, within the limits of this text, a statement of the principles upon which this correction is based.

The Minor Aberrations

In addition to the three major shortcomings of lenses, spherical aberration, astigmatism, and chromatic aberration — which has not yet been treated in detail — there are several lesser aberrations, the chief of which are distortion and curvature of plane. Unlike the other aberrations, distortion does not involve any failure of the light from one point of the object to converge to a single point in the image. As the name indicates, distortion consists, not of a "fuzzy" image, but of a misshapen image. Straight lines in the object become warped in the image, especially at the edges. A square object produces a bulging image or else an image the sides of which sag inwards. The former is ***barrel distortion*** and occurs in real images formed by uncorrected converging lenses. The latter is ***pincushion distortion*** and occurs in the virtual image formed by the same kind of lens. Examples are shown in Figure 260.

The edges of photographs showing distortion are usually somewhat out of focus. This indicates that the edges of the image would be in focus, not in the same plane as the center, but in front of or behind it. A saucershaped screen would be required to catch all portions of such an image in focus at once. This type of aberration is known as *curvature.*

Distortion and curvature are corrected by the use of twin lenses, usually somewhat separated, with a diaphragm between them. The distortion and curvature produced by one component are neutralized by equal and opposite aberrations produced by the other. Such a combination is termed a *rectilinear* lens.

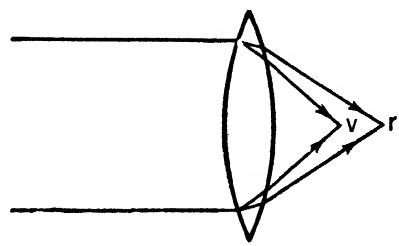


FIG. 262. CHROMATIC ABERRATION

Chromatic aberration, the only important deficiency of simple lenses that has not been discussed, is caused by the fact that the focal length of a lens is different for different colors. Hence, numerous images of the same object, different in color and size, are formed by a lens which is not corrected for chromatic (that is, *color*) aberration. The manifestation of chromatic aberration on a photograph is much the same as that of spherical aberration. The cause, however, is quite different. The detailed discussion of it must await the developments of the following chapter.

Questions for Self-Examination

1. Use Robert Smith's construction to locate the image of an object $7\frac{1}{2}$ centimeters away from a converging lens of focal length 3 centimeters; 4 centimeters away from the same lens; 6 centimeters away from a diverging lens of focal length − 9 centimeters.
2. Using the lens-maker's equation $\frac{1}{f} = (n-1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right)$, deduce the focal length (in terms of the radii of curvature) of (*a*) a plano-convex and a plano-concave lens, (*b*) an equibiconvex and equibiconcave lens, (*c*) a positive meniscus lens of which the radii bear the ratio 2:1. In all cases assume the index of refraction to have the value 1.5.
3. Describe spherical aberration. Tell how it may be corrected and what the limitations are on such correction.
4. Describe astigmatism both as it is produced in lenses and in the eye.
5. Describe the aberrations termed *distortion* and *curvature.*

Problems on Chapter 30

1. An object is l centimeters from a wall. A converging lens forms an image of it on the wall. When moved a distance d centimeters it also forms an image. Prove that the focal length of the lens f is
$$f = \frac{l^2 - d^2}{4l}.$$
2. Prove that the ratio of the sizes of the two images produced as in the preceding problem is
$$\left(\frac{l+d}{l-d}\right)^2.$$

3. A double convex lens, the ratio of whose radii is 6:1, produces parallel rays when a source of light is 2 inches away. What are the radii if the index of refraction is 1.5? 7 in. and $1\frac{1}{6}$ in.
4. A plano-convex lens is to be made of glass of index 1.6. It is to form a real image of an object placed 2 inches in front of it and to magnify it three times. What must be the radius of curvature? .9 in.
5. The lens-maker's equation, when the thickness t of the lens is taken into account, is

$$f = \frac{\mu}{\mu - 1} \frac{r_1 r_2 - \frac{\mu - 1}{\mu} r_2 t}{(\mu - 1)t - \mu(r_1 - r_2)}.$$

Calculate the focal lengths of the equibiconvex lens mentioned in the quotation from Halley on page 328, both with this relation and with its simplified form in equation (3) and verify Halley's statements.

6. An object is placed u centimeters from a thin converging lens of focal length f centimeters. What is the distance v in centimeters of the image from the lens? What is the magnification m? Locate the image on a diagram.

u	f	v	m
24	6	8	$-\frac{1}{3}$
12	6	12	-1
4	6	-12	3
0	6	0	1

7. An object is placed u centimeters from a thin diverging lens of focal length f centimeters. What is the distance v in centimeters of the image from the lens? What is the magnification m? Locate the image on a diagram.

u	f	v	m
24	-6	$-4\frac{4}{5}$	$\frac{1}{5}$
12	-6	-4	$\frac{1}{3}$
4	-6	$-2\frac{2}{5}$	$\frac{3}{5}$
0	-6	0	1

8. Two thin lenses, of focal length f_1 and f_2 centimeters respectively, are placed d centimeters apart. An object is placed a distance U centimeters in front of the first lens. How far V from the second will the image be formed and what will the magnification be? Make a diagram. (For data, see page 324.)
9. At what distance, u_1 (referred to the focal length f), must an object be placed from a converging lens in order that the image shall be magnified m times? From a diverging lens, distance being u_2? Interpret the signs in data and answers.

9.

m	$\frac{u_1}{f}$	$\frac{u_2}{f}$
2	$\frac{1}{2}$	$-\frac{1}{2}$
1	0	0
$\frac{1}{2}$	-1	1
$\frac{1}{3}$	-2	2
$-\frac{1}{3}$	4	-4
$-\frac{1}{2}$	3	-3
-1	2	-2
-2	$\frac{3}{2}$	$-\frac{3}{2}$

10.

r_1	r_2	μ	f
30	-30	1.5	30
30	∞	1.5	60
30	60	1.5	120
30	15	1.5	-60

10. A thin lens has radii of curvature r_1 and r_2 centimeters (positive if convex to incident light). The index of refraction of the glass is μ. What is the focal length f of the lens in centimeters?

11. Equibiconvex lenses made of various materials, each having radii of curvature of 10 centimeters, are immersed in water. Find the focal length, f_1, in air and that, f_2, in water for each lens.

11.

Material	μ	f_1	f_2
diamond	$2\frac{1}{4}$	4	$7\frac{3}{11}$
flint glass	$1\frac{3}{4}$	$6\frac{2}{3}$	16
crown glass	$1\frac{1}{2}$	10	40
water	$1\frac{1}{8}$	15	∞
air	1	∞	-20

12.

f_1	f_2	F
6	4	2.5
5	3	1.9
4	2	1 3
3	1	75

12. Two thin lenses in contact have focal lengths f_1 and f_2 centimeters respectively. What is the focal length F of the combination in centimeters?

13. A near-sighted person finds his distance of most distinct vision to be d centimeters. What should be the focal length f of his spectacles to make the distance of most distinct vision 40 centimeters with their aid? Assume the spectacle lens to be in contact with the simple lens which is to be regarded as the optical equivalent of the eye.

	d	f
13.	9	− 12
	15	− 24
	24	− 60
	33	− 190

	d	f
14.	50	200
	60	120
	90	72
	240	48

14. Solve problem 13 for a far-sighted person.

15. A normal person has 40 centimeters as his distance of most distinct vision. If, in order to examine a small picture, he uses a reading glass of focal length f centimeters, producing a lateral magnification m, how far u from the picture and l from his eyes will he place the lens? What will be the ratio α of the angles subtended at his eye by the image and the picture?

f	m	u	l	α
5	5	4	20	3
10	4	$7\frac{1}{2}$	10	$1\frac{3}{4}$
15	3	10	10	$1\frac{1}{2}$
20	2	10	20	$1\frac{1}{2}$
25	$1\frac{1}{2}$	$8\frac{1}{3}$	$27\frac{1}{2}$	$1\frac{11}{32}$

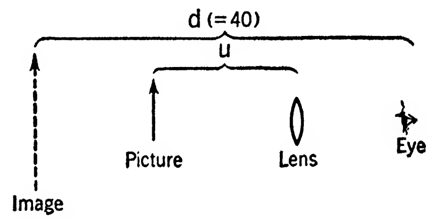


FIG. 263

16. Two thin converging lenses, each of focal length f centimeters, are placed d centimeters apart. How far V in centimeters from the second is the image of an infinitely distant object? (This gives the position of the principal focus of the *combination.*) Make a diagram.

f	d	V
2	6	4
2	3	− 2
2	1	$\frac{2}{3}$
2	0	1

17. A Ramsden eyepiece consists of two lenses of equal focal length f centimeters separated by a distance d centimeters. What are the distances l of the principal foci from the component lenses? Draw a diagram.

	f	d	l
17.	5.5	3.5	$\frac{2}{15}$
	4.5	3.	$\frac{9}{8}$
	3.5	2.5	$\frac{7}{9}$
	3.	2.	$\frac{3}{4}$

	f_1	f_2	d	l_1	l_2
18.	2	6	3.	$+\frac{6}{5}$	$-\frac{6}{5}$
	2	5	3.5	$+\frac{6}{7}$	$-\frac{15}{7}$
	1	4	2.5	$+\frac{3}{5}$	$-\frac{12}{5}$
	1	2	1.5	$+\frac{1}{3}$	$-\frac{2}{3}$

18. A Huygens eyepiece consists of two lenses of focal length f_1 and f_2 centimeters respectively, separated by a distance d centimeters, and so placed that light traverses first the lens of greater focal length. What are the distances l_1 and l_2 of the principal foci of the combination, each from its respective lens? Draw a diagram.

CHAPTER 31

Properties of Prisms

Early Attempts to Correct the Aberrations

It was in the middle of the seventeenth century that aberrations of lenses and mirrors became a live topic for study and experiment. Galileo had made one type of telescope famous early in the century, and Kepler had designed another. Astronomers had in the meantime become thoroughly awake to the value of what is now their principal instrument, and Huygens had become deeply involved in supplying them with telescopes which were longer than any ever made, before or since.[1] In the course of this work he devised a type of eyepiece which is in use to this day and still bears his name.

Since aberrations are magnified in just the proportion that images are, the use of these long telescopes was making the aberrations of their lenses most dishearteningly prominent. At this time chromatic aberration had not been identified, and it was generally supposed that lenses were subject to no other errors than those which arose from the spherical figure of their surfaces. The trend of the times was to try to correct these errors by the substitution of aspherical surfaces, attempts which were only partly successful for reasons which are common knowledge today.

Newton's Approach to the Problem

Newton, as an undergraduate at Cambridge University, had become interested in optical problems through reading Kepler's *Dioptrice*, published in 1611. Among others he had accordingly interested himself in telescopes, and as early as 1662, at the age of twenty, he was busily grinding specially figured lenses in attempts to improve their performance. Finding that chromatic aberration still remained after the spherical aberration had been reduced to a very tolerable minimum and concluding over-hastily from some rough experiments that it was inherently impossible to eliminate chromatic aberration from lenses, he turned to mirrors in place of lenses as telescope objectives. As Newton himself stated it (90:102):

> Seeing therefore the Improvement of Telescopes of given lengths by Refractions is desparate; I contrived heretofore a Perspective [2] by Reflection, using instead of an Object-glass a concave Metal.

[1] "Telescopes," *Encyclopaedia Britannica* (14th edition).

[2] For several centuries the English term for *telescope* was *perspective glass*, or more briefly *perspective*.

Encouraged by a considerable measure of success in his first reflecting telescope, Newton made a second, larger and optically better, and in December 1771 he donated this second telescope to the Royal Society of London, where it has remained to this day (Fig. 264).

Newton was not the first to think of the possibility of substituting a concave mirror for a lens as the objective of a telescope, but he was the first actually to make such an instrument. The first to design a reflecting telescope seems to have been Niccolo Zucchi in 1616 (130:*1*:307).

Important though the development of the reflecting telescope was, it was of far less significance than a discovery which grew out of Newton's attempts to circumvent the chromatic aberration which he believed to be inescapable in the refracting type of telescope. About a month after he had presented his reflecting telescope to the Royal Society, Newton wrote the following in a letter to the Secretary (19:*3*:5):

> I desire that in your next letter you would inform me for what time the society continue their weekly meetings; because if they continue them for any time, I am purposing them, to be considered of and examined, an account of a philosophical discovery which induced me to the making of the said telescope; and I doubt not but will prove much more grateful than the communication of that instrument, being, in my judgment, the oddest, if not the most considerable detection which hath hitherto been made in the operations of nature.

Newton's First Scientific Paper

The last clause in the foregoing letter is rather startling, coming from Newton. An equivalent statement from almost anyone else about one of his own discoveries would quite naturally encounter skepticism. Even Newton still had his reputation to make; for though he was now nearly thirty years of age and had already laid the foundation for his major scientific accomplishments, he was yet to publish his first scientific paper. In fact the paper which he read before the Royal Society a month after the foregoing letter, setting forth the "philosophical discovery" which was "the oddest, if not the most considerable detection which hath hitherto been made in the operations of nature," constituted the material for his first scientific paper, which was published in the *Philosophical Transactions* of the Royal Society for the year 1672.[1] The judgment of succeeding generations has fully vindicated Newton's high opinion of the value of this, his first contribution to scientific literature. Next to his discovery of the law of gravitation, which matured several years after this time, and which is commonly considered the greatest of all scientific discoveries, Newton's first scientific paper, entitled *A New Theory about Light and Colours*, is his best. Besides its scientific value, it is a model in clarity of exposition. An excellent short summary is to be found in his *Optical Lectures*, written about two years before in connection with his teaching at Cambridge University,

[1] (Abridged), *6*, 3075–87.

but not published until 1728, nearly fifty years after they were written. At one point he says (89:5):

> Concerning Light, I have discovered that its Rays, in respect to the Quantity of Refraction, differ from one another. Of those that have all the same Angle of Incidence, some will have their Angle of Refraction somewhat greater, others will have it somewhat less.... I moreover find that the Rays refracted the most produce purple Colours and those the least refracted produce red Colours... and so the Rays... do generate these Colours in order; red, yellow, green, blue and purple, together with all the intermediate ones that may be seen in the rainbow.... That you may not think we have declared to you Fables instead of Truth, we shall immediately produce the Reasons and Experiments on which these things are founded.

Newton's Discovery of the Spectrum

Newton was not the first to record an observation of the production of spectral colors by a prism, nor even the first to study it in some detail. Among earlier observers had been Marcus Marci in 1648 (130:*1*:314), Francisco Grimaldi in 1665 (49), and even Seneca in the first century (114:30). But Newton's discovery, though later than these, was made independently of them and went far beyond any of them toward being complete and exhaustive.

The scene as Newton undertook his experimentation with a prism pro-

FIG. 264. NEWTON'S REFLECTING TELESCOPE
(Photographed in the Royal Society of London.)

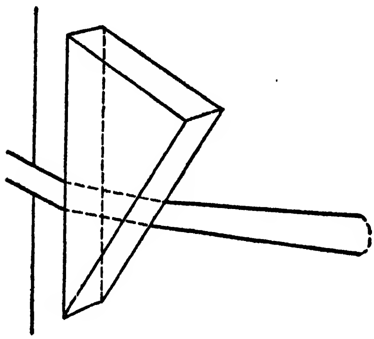


FIG. 267. NEWTON'S ILLUSTRATION[1] OF HIS PRISM EXPERIMENT

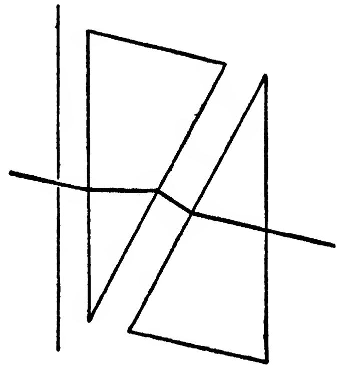


FIG. 268. RECOMBINATION BY A SECOND PRISM

cured for the purpose has been admirably reconstructed in the painting reproduced in Figure 265. What he saw is indicated in Figure 266.[2] The pinhole made in the window curtain formed a single white image of the sun and when the prism was interposed in the beam as shown in his own drawing (Fig. 267), this was divided into an infinity of spectrally colored images of the sun, serially displaced to form the pattern shown in Figure 266. Newton tried every conceivable arrangement of aperture, screen, and prisms in the process of identifying the origin and nature of the phenomenon under investigation. To two of these arrangements he gave particular emphasis. One was the recombination of the dispersed colors back into a single white-light image of the sun by the use of a second prism identical with the first, but so set as to produce contrary refraction (Fig. 268). This showed that spectral colors, when recombined, produced white light, an utterly new idea and one which was destined to produce violent controversy and criticism of his work. Another was what Newton termed his "experimentum crucis," his own illustration of which appears in Figure 269. Col-

[1] Figures 267 and 268 are redrawn from Newton's *Optical Lectures*, delivered at Cambridge University, England, in 1669 and filed in the University Archives. The lectures were not printed until 1728.

[2] Figures 265 and 266 are on the color plate, opposite pages 386–87.

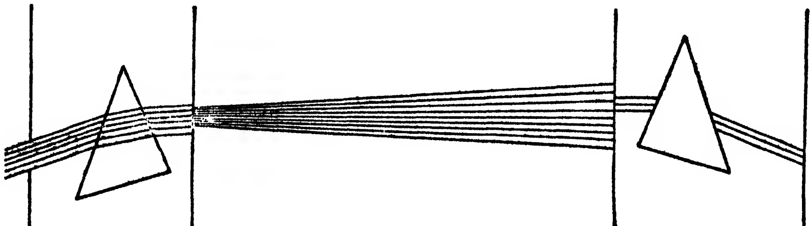


FIG. 269. NEWTON'S "EXPERIMENTUM CRUCIS"
(Redrawn from his *Opticks* of 1704.)

ors from the first prism were allowed, one by one, to pass through a second prism. Two significant observations resulted; first, there was no further dispersion. Green light passing through the second prism remained green, and similarly for other colors. Second, refraction by the second prism was greatest for violet light and least for red. Hence, in Newton's words:

> [*White*] *Light* is not similar or homogeneal, but consists of *difform Rays*, some of which *are more refrangible than others*; so that without any Difference in their Incidence on the same Medium, some shall be more *refracted* than others.

Thus went one of the most famous of all scientific papers (the italics are Newton's own).

Newton's Erroneous Opinion on Chromatic Aberration

It is somewhat ironic that such a definitive study as the foregoing should have led to one serious misapprehension, that misapprehension being on the very point which was the point of departure for Newton's famous study as outlined above. He concluded, erroneously, that it was forever impossible to design lenses which would be free from chromatic aberration. It is worth while to observe the basis for this error.

During the course of his experimentation Newton used five or six prisms, the refracting angles of which varied from 45 degrees to 64 degrees. The prisms of larger angle deflected their spectra further to one side than did the prisms of smaller angle, and also produced spectra which were longer than those produced by the narrower prisms. Newton surmised that the lengths of the spectra produced by the prisms were in strict proportion to the mean deviations of the beams of light produced by them. If, for example, his 64-degree prism deviated the light four times as far to one side as did the 45-degree prism, he expected to find that the two spectra had also a length-ratio of four to one; and within the limits of his apparently hasty measurements, he found this to be true. Thus far all was well. But at this point Newton made a curious blunder. It occurred to him to inquire whether the proportionality between mean deviation and length of spectrum applied also to other substances besides glass, and he accordingly made a glass-sided hollow prism wherewith to repeat the experiment using water. He should have found a decided departure from the expected proportionality: he should have observed, for example, that, for the same deviation, the water prism produced a spectrum shorter than that produced by the glass. Instead, he found that (90:31)

> in a Vessel made of polished Plates of Glass cemented together in the shape of a Prism and filled with Water, there is the like Success of the Experiment *according to the quantity of the Refraction.*[1]

[1] Italicized phrase indicates *proportional to the deviation* as the context here and at other points of Newton's treatment of the subject shows.

The cause of this blunder we shall probably never know. Its bearing on the problem of correcting chromatic aberration is not far to seek.

Newton's Erroneous Opinion on Chromatic Aberration

Achromatic lenses are today constructed of two components. A converging lens of crown glass is paired with a diverging lens of flint. The crown lens acts on the light in two ways: by bending it (refraction) and by spreading it into colors (dispersion). The flint lens acts in the same two ways, but with its refraction and dispersion both in opposite directions to those of the crown lens. It is so proportioned to the crown lens, moreover, that the opposite dispersions effected by the two are equal. Light dispersed into the colors of the rainbow by the crown lens is hence recombined into white light by passing through the flint lens.[1] But though the dispersions are equal, the refractions are not. The converging effect of the crown lens is only partly offset by the diverging effect of the flint. Newton's error was in concluding too hastily that lenses possessing equal and opposite dispersions must necessarily also possess equal and opposite refractions and that a combination of two lenses whose dispersions mutually neutralized could not possibly bring light to a focus. If this were true, then his conclusion that "the Improvement of Telescopes by Refractions is desparate" was inescapable.

The credit for correcting Newton's error and manufacturing telescopes possessing achromatic objectives is commonly given to John Dollond, who repeated Newton's experiments with water prisms and published the corrected results.[2] He had received very broad hints, however, from publications of a Swede named Klingenstierna and the French mathematician Clairaut, neither of whom were in a position to submit their theories to the test of the actual manufacture of achromatic lenses. Moreover, it subsequently developed that a certain Chester More Hall had anticipated Dollond by twenty-five years in making achromatic lenses, but he made only a few and did not publish his accomplishment. Hence, the invention was soon forgotten. Dollond, on the contrary, put the design and manufacture of achromatic lenses onto a sound foundation, in virtually its present form. Since he entered the scene at the proverbial psychological moment it is natural, if not strictly correct, to accord him the credit of having originated the current method of dealing with chromatic aberration in objective lenses.

Minimum Deviation

Newton's experiments with prisms consisted primarily in a study of what deviations from their original directions prisms produced in beams of light. The experiments concerned especially the differences in these deviations

[1] More correctly they neutralize only for two predetermined colors, leaving a slight residual aberration, which is, however, seldom of significance.

[2] *Philosophical Transactions*, *50*, 733 (1758).

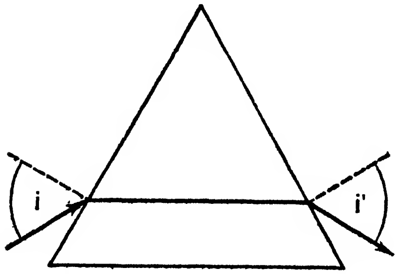


FIG. 270. THE POSITION OF MINIMUM DEVIATION

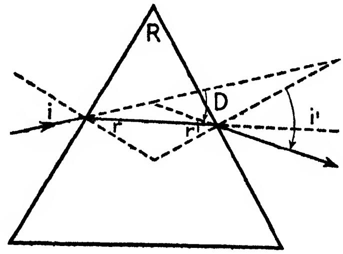


FIG. 271. REFRACTION THROUGH A PRISM

which the different colors composing white light experienced. Though he was not the first to observe either of these effects, he was the first to conduct an exhaustive study of them. But as an incidental element in this study, he made the first observation on record of the phenomenon known as *minimum deviation.* He observed that there was a certain angle of incidence for which the angle through which a beam of light was deviated by a prism possessed a minimum value. A prism so oriented with reference to the incident light is said to be in a position of minimum deviation, and the corresponding angle of incidence is said to be the angle of minimum deviation. Newton established the fact that for this position the prism was symmetrically located with reference to the rays incident on the prism and those leaving it. That is (Fig. 270), the angles i and i' are equal.

It is striking enough to see the beautiful colors of the spectrum projected on a screen from a beam of white sunlight. It is even more striking to see that spectrum, moving across the screen in response to a rotation of the prism, slow down and ultimately reverse its motion while the prism continues to rotate in the original direction. Newton states his observation of this effect as follows (89:19–20):

> If the Prism be held in the Sun's Light, and by a gentle Motion turned about its Axis, you will see the Colours, which it makes, to be by a continual Motion translated from Place to Place, in such a Manner, as sometimes they will appear to ascend, and then again descend. Observe therefore the Middle between these contrary Motions, when the Colours, now ascending and presently being about to descend, seem to be stationary.... Then the Inclination of the emerging Ray to the incident one will be least of all. Which, when it happens, the Refractions on both sides are equal, as shall be demonstrated hereafter.

The "demonstration," to be found in his *Optical Lectures* of 1669 (Sec. III, Prop. 25), is one of the first, if not the very first, applications of Newton's famous *Method of Fluxions.* This method later developed into the branch of mathematics now known as *Calculus.* Since a grasp of this very useful branch of mathematics cannot be presumed on the part of the reader, Newton's argument will be modified as follows.

Consider first the case where $i \neq i'$ (Fig. 271). The angle of deviation is always the angle between the incident and emergent ray. Imagine the direction of the ray shown herewith to be reversed. Though it traverses the prism in the opposite direction, it follows the same path as before, and, hence, its *deviation* is the same. That is, two different angles of incidence produce the same deviation. Therefore, if we gradually change the angle of incidence from i to i', the simultaneously changing angle of deviation starts and ends with the same value. Hence, disregarding the possibility of a constant, unchanging angle of deviation — which can readily be shown not to obtain — the deviation must in the meantime have first decreased and then increased (passed through a minimum) or *vice versa* (passed through a maximum). The former is what really occurs. This minimum value of the deviation occurs when the angle of incidence lies between i and i', no matter how small the difference between these two angles may be. So it must occur when $i = i'$.

The Role of the Prism in Refractive Index Measurements

It is convenient to set a prism at the position of minimum deviation in many of its uses in the laboratory. Not only is the spectrum at its sharpest and brightest for this position, but the measurement of index of refraction of the glass of which the prism is made — a frequent necessity — is greatly simplified by this adjustment. For this case, D being the angle of minimum deviation for a prism having a refracting angle R, the index of refraction is

$$\mu = \frac{\sin \frac{1}{2}(R + D)}{\sin \frac{1}{2} R}. \tag{1}$$

This commonly utilized relation was first established by Joseph Fraunhofer (1787–1826), a Bavarian about whom much will be said in the following chapter. It may be derived as follows. In Figure 271, the deviation is seen to take place in two steps: that at the first surface is $i - r$, and that at the second is $i' - r'$. The total deviation is the sum of these two, or

$$D = i - r + i' - r' = (i + i') - (r + r').$$

But from the figure $r + r' = R$. Also, in case of minimum deviation $i = i'$, and hence $r = r'$. Therefore

$$D = 2\,i - R \text{ or } i = \tfrac{1}{2}(R + D) \text{ and } r = \frac{R}{2}. \tag{2}$$

But applying Snel's law (page 290) to the ray entering the prism and substituting equations (2) in it,

$$\mu = \frac{\sin i}{\sin r} = \frac{\sin \frac{1}{2}(R + D)}{\sin \frac{1}{2} R},$$

which is equation (1). Determination of the index of refraction of a specimen of glass made into a prism thus requires only two angle measurements: the refracting angle R of the prism and the angle of minimum deviation D.

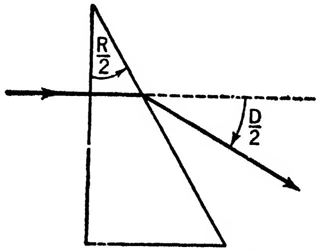


FIG. 272. DESCARTES' METHOD OF MEASURING INDEX OF REFRACTION

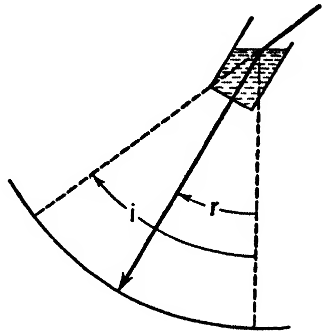


FIG. 273. NEWTON'S METHOD OF MEASURING THE INDICES OF LIQUIDS

It is worthy of note that, in the time of Galileo, Descartes utilized what was in effect this same method in measuring indices of refraction (quoted 103:135). His arrangement was really one half of the prism to which equation (1) applies (Fig. 272). The deflection was also half that denoted until now by D. Hence, equation (1) applies as it stands to Descartes' method.

Newton devised an almost identical way of measuring the index of refraction of liquids (103:135). A glass-bottomed container (Fig. 273) was mounted on a graduated scale. The device was tilted until the emerging beam was parallel to the carrier, the inclination to the vertical giving the angle of refraction. Then the liquid was removed and the angle of incidence similarly determined. A direct application of Snel's law then gave the value of the index of refraction of the liquid.

Total Reflection

Newton referred frequently — though in a purely incidental manner — to another phenomenon which like the production of spectra may be conveniently demonstrated with the aid of prisms. This phenomenon was *total reflection*. Without doubt he owed his recognition of it to Kepler, who, in his *Supplement to Vitellio* of 1604, was the first to mention it. The name *total reflection* arises from the fact that under certain conditions reflection of light may occur with only a very small loss of the incident light. This is in marked contrast to reflection from ordinary mirrors, which is usually characterized by the loss of about one quarter of the incident light. Prisms designed to provide total reflection are used wherever economy of light is important. One of the common examples of this is in the so-called prism binocular (Fig. 274). This is a double-barreled telescope the length of which has been shortened by reflecting the light back and forth to secure the necessary distance without the inconvenient length associated with the usual "spy-glass."

The nature of total reflection may readily be comprehended. Consider

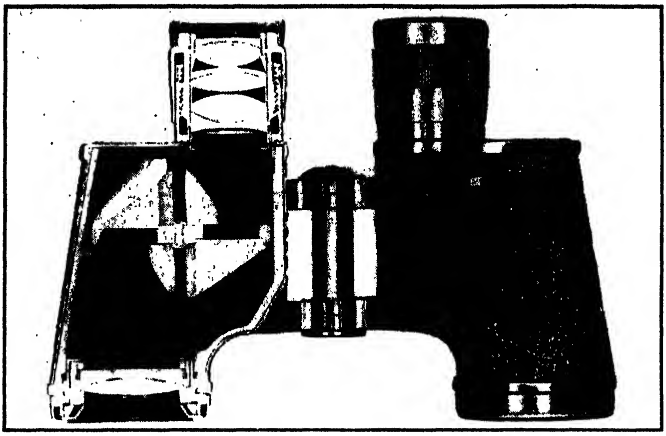


FIG. 274. THE OPTICAL SYSTEM OF A PRISM BINOCULAR
(Courtesy of Bausch and Lomb Optical Company.)

first the passage of light into a block of glass at a variety of angles (Fig. 275). It is evident that all rays incident within the right angle *jon* will after refraction be comprehended within the acute angle *j'on'*, in accordance with Snel's law, $\mu = \frac{\sin i}{\sin r}$. In particular, the grazing ray *jo*, making an angle of incidence of 90°, enters the glass after refraction at the much smaller angle *j'on'*. The magnitude of the angle *j'on'* may be deduced by substituting the value of μ and of i (90°) in Snel's law and solving for r. For crown glass the angle is about 40°; for water, nearly 50°. If the entire surface of the glass were covered except for a small hole at *o*, no illumination from any direction whatsoever could reach beyond the region *p'j'*. The rays *oj'* and *op'* are called *critical rays*. The angle which they make with the normal to a surface are characteristic of the medium. If their direction in glass is reversed, a converging cone from below only 80° in width spreads out to a full 180° upon emergence above the surface. Whatever the angle of incidence from beneath, as long as it is within the cone *p'oj'*, some of the light is reflected back into the lower medium. But the larger part of it penetrates the surface and emerges to be spread out through a full 180°.

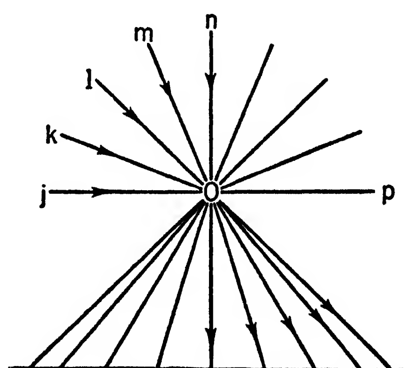


FIG. 275. THE SIGNIFICANCE OF THE CRITICAL ANGLE

If, now, a ray should originate in the lower medium (glass or water), outside of the cone limited by the critical angle, and proceed toward *o*, what would occur? Obviously it cannot emerge through *o*, for the critical ray, *j′o*, upon emergence took the grazing direction *oj*. There is, however, nothing to prevent its being reflected, and that is precisely what will happen. Since none of it can escape into the outer medium it is *all* reflected. This is the origin of the term *total reflection.*

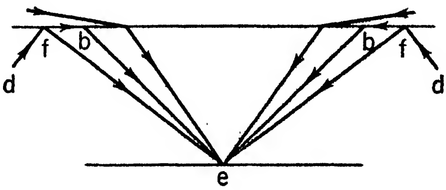


FIG. 276. THE GEOMETRY OF UNDERWATER VISION

A somewhat different manifestation of the same effect is what is sometimes termed the *fish's-eye view.* To an eye beneath the surface of water, the entire outside horizon is contained within the cone *beb* of Figure 276. The general impression must be somewhat like that which one would receive on looking up from a manhole through the circular open cover — except that within the illuminated circle all of the surrounding scenery is contained. Moreover, on looking toward what would be the upper portion of the side of the "manhole," one would see an inverted image of the bottom of the pool, reflected in the surface of the water (for example, the ray *dfe* in Figure 276). The state of affairs is well illustrated in Figure 277. The fisherman's legs are doubled, being seen both directly and inverted by reflection. The rest of him, seen past the edge of the "manhole," appears raised into the air.

The most common way of utilizing total reflection in optical instruments is by means of a prism whose angles are 45–45–90 degrees respectively.

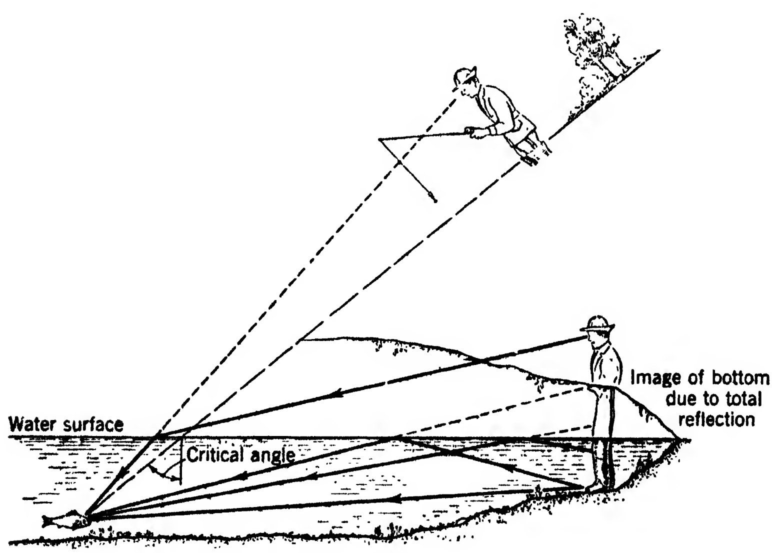


FIG. 277. THE FISH'S-EYE VIEW OF A FISHERMAN
(From *College Physics*, by A. L. Foley. P. Blakiston's Son & Co., Inc., 1933.)

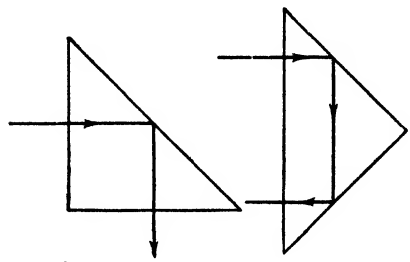


FIG. 278. TOTAL-REFLECTION PRISMS

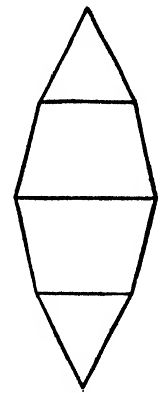


FIG. 279. THE LENS AS A GRADED ARRAY OF PRISMS

Since the critical angle for glass is about 40°, the angle of incidence on the hypotenuse of the prism, being 45°, is greater (Fig. 278). Total reflection, therefore, occurs at this face. The principal losses are by absorption within the glass and, more important, by the small amount of reflection at the two short faces of the prism as the light enters and leaves. It is in this way that the two prisms in each tube of a prism binocular act. The prisms are so disposed, moreover, that one of them reinverts the image which in the usual telescope is normally seen upside down, and the other rereverses the image similarly once reversed.

Perhaps the greatest scientific utility of the phenomenon of total reflection is in facilitating the measurement of indexes of refraction. Although a specimen in the form of a prism is required for the method of minimum deviation, a glass specimen in the form of a plate with a polished edge may be used for refractive index measurement by the method of total reflection. The principle involved is simply the determination, which is usually somewhat indirect, of the critical angle. From this the index of refraction may be calculated from Snel's law, since for refraction at the critical angle, the corresponding angle of incidence is 90°. Thus in $\mu = \frac{\sin i}{\sin r}$, substitute $i = 90°$, $r = \theta$, the critical angle, whence

$$\mu = \frac{1}{\sin \theta}. \tag{3}$$

Next to the lens, the prism is perhaps the most useful of optical devices. In a sense, the prism is more fundamental than the lens, in that the latter may be regarded as made up of a succession of prisms with a continuous gradation of angles (Fig. 279). Of the numerous uses which have been found for prisms — only a few of which have been mentioned here — the most significant in the development of physics is unquestionably the production of spectra. Notwithstanding this, the prism is not the only, nor even the most advantageous, instrument used nowadays in the study of spectra. While the prism still has its advantages in this field, other and more powerful devices have displaced it for certain purposes. The nature of some of these will become evident incident to a closer study of spectra themselves.

Questions for Self-Examination

1. Describe chromatic aberration and Newton's misconception about it.
2. Tell the story of Newton's discovery of the composition of white light.
3. What is the common method of correcting a lens for chromatic aberration, and what is the limitation on such correction?
4. What did Newton observe about the position of minimum deviation (make a sketch)?
5. Describe two ways of using a prism in the measurement of index of refraction.
6. What is meant by the terms "critical angle" and "total reflection"?
7. Describe and account for the "fish's-eye view."

Problems on Chapter 31

1. What is the lowest value of refractive index for which a 45–45–90 prism will be totally reflecting? 1.4.
2. If the fish of Figure 277 is 3 feet below the surface, what is the diameter of the apparent "aperture" at the water surface above him? 6.8 ft.
3. Achromatic lenses usually consist of two thin lenses in contact. Show that for such a case the reciprocal of the focal length of the combination is the sum of the reciprocals of the focal lengths of the individual lenses. (Apply equation (1) of Chapter 30 to find the position of the image — formed by the second lens — of the image of an infinitely distant object formed by the first lens.)
4. It is found that the two components of an achromatic lens require focal lengths in the ratio of −3:5, the crown component being positive and the shorter of the two. What will be the focal length of each of the components of the objective of a 20-foot achromatic telescope? 8 ft, $-13\frac{1}{3}$ ft.
5. Light incident on a block of glass at i degrees with the normal is refracted at r degrees. What is the refractive index μ? What is the speed v in meters per second of light in the glass?

i	r	μ	v
40	25	1.52	$1.97 \cdot 10^8$
50	30	1.53	1.96
60	35	1.51	1.99
75	40	1.5	2.

6. Light is incident from beneath on a smooth surface of water of refractive index 1.332, at an angle i with the normal. What angle r does the emergent ray make with the normal?

	i	r
6.	20°	27°
	30°	42°
	40°	59°
	48° 39′	90°

	D	μ
7.	45°	1.59
	50°	1.64
	55°	1.69
	60°	1.73

7. Light is caused to be incident on a 60-degree prism. The angle of minimum deviation is found to be D degrees. What is the refractive index μ?

8. What is the deviation d in degrees of a ray of monochromatic light upon passing through a 60-degree prism of refractive index μ, if the ray is incident upon one face at an angle i degrees with the normal?

	i	μ	d		*medium*	μ	θ		θ	μ
8.	60°	1.6	47°	9.	water	1.33	49°	10.	34° 34′	1.59
	50°	1.6	46°		crown glass	1.52	41°		38° 35′	1.64
	40°	1.6	51°		flint glass	1.7	36°		42° 34′	1.69
	34° 34′	1.6	64°		diamond	2.42	24°		46° 25′	1.73

9. The indexes of refraction of various media being μ, what is the critical angle, θ, for each medium?

10. The angle of emergence of the critical ray in the prism of problem 7 is found to be θ degrees. What value μ does this give for the refractive index?

11. The refractive index of a glass plate d centimeters thick is μ_r for red light (wave-length 6500 angstroms in air) and μ_v for violet light (wave-length 4000 angstroms in air). By how many n of its own wave-lengths is red light ahead of violet after both have traversed the plate, starting together? (1 angstrom = 10^{-10} meter.)

μ_r	μ_v	d	n
1.5127	1.5214	1	130
1.6038	1.62	1	250
1.6126	1.6213	1	130
1.7434	1.7723	1	440

CHAPTER 32

Color

Colored Lights vs. Colored Pigments

Newton's accomplishment in making the first correct and complete analysis of the structure of white light stands out, even today, as one of the more outstanding scientific contributions. But it was such a complete break with traditional views that the discovery was not only accepted slowly, it actually met with strong opposition for more than a century. This originated partly in professional jealousy as in his encounter with Robert Hooke (14:73), partly in contrary philosophical predilections (Schelling 113:270; Hegel 119:419), and partly in aesthetic considerations (Art. "Goethe," *Encyclopaedia Britannica*, 9th edition). There were, however, certain difficulties of a more substantial variety in Newton's color theory.

One of the obstacles which stood in the way of a ready acceptance even by men of science of Newton's discovery of the nature of color was its apparent disagreement with common experience in mixing colors. It was a matter of everyday knowledge then, as now, that mixing yellow and blue paints produces green. But it had been one of the implications of Newton's studies on color that yellow light superposed on blue would produce, not green, but a neutral gray, which lacked only the requisite intensity to be called white. Newton, however, does not seem to have recognized that this was involved in his own theory. The first actual observation of this effect to be recorded was made by Helmholtz in 1852, more than two hundred years later. The blue-yellow paradox requires explanation even today. It is not surprising that it proved a stumbling block to Newton's contemporaries.

Actually, there are two different phenomena involved. One is the mixing of colored substances (mixing "pigments" as the expression goes) and the other is the superposition or "mixing" of differently colored lights. A clear distinction between the two is the first step toward the understanding of either. But it is only the first of several steps that are necessary before the difficulty can be cleared up. The basis for what is observed when pigments are mixed was established by Newton. But, though he observed some of the facts involved in the mixing of lights, the explanation for those facts did not take even a preliminary form until more than a century after his time, and is even now not entirely conclusive.

The nature of the process of mixing pigments may best be seen by repeating Newton's original experiment with the prism. In this case, however, before allowing the light to strike the prism, cause it to be reflected from a surface painted with a solution of gamboge yellow, which is the pigment most commonly used to produce yellow paint. The color of the surface suggests that only yellow light is reflected from it, and one might perhaps expect, therefore, to see all the spectrum except the yellow blotted out. Instead, it will be seen that, not only the yellow, but also some red, some orange, and still more clearly some green are present in the spectrum. It is evident that the eye interprets this mixture of hues as yellow. Thus there are at least two ways in which the eye may be stimulated to "see" yellow. One is by light constituting a narrow band in the spectrum between orange and green, and the other, which may seem identical in effect, is by a certain mixture of all colors except blue and violet. The former is termed a *spectrally pure yellow*; the latter, while still stimulating the sensation of yellow, is spectrally very impure indeed. There is, in fact, an almost indefinite number of mixtures of colors which will stimulate the yellow sensation. It is not even necessary that yellow be present in the mixture. Light containing only spectrally pure red and green in a certain proportion — there being no spectral yellow whatever in it — will still appear yellow.

If, instead of gamboge yellow, a surface painted with Prussian blue should be substituted, a similar phenomenon would be observed. Besides the blue of the spectrum, violet and green would be evident. In this case, as before, it becomes evident that there is more than one way to excite the sensation of blue, and that it is possible, with the aid of a prism, to distinguish a spectrally pure blue from apparently identical hues not otherwise distinguishable.

The only color reflected by both pigments was green. If, now, the two pigments were mixed and a third surface painted with the mixture, all colors except the green would be absorbed by the mixture, some by one component and some by the other. The painted surface would appear green, and in this instance the prism would demonstrate that it was a spectrally pure green. All the other colors would be absent, absorbed by the mixture covering the surface from which the light was reflected. Hence, the mixture of yellow and blue pigments produced green because, by joint action of the two pigments, all other colors were absorbed. The production of green by a mixture of yellow and blue pigments is represented in Figure 281.

Refinements in Color Observation

Since the eye is not capable of discriminating between spectrally pure colors and mixtures of color stimuli producing the same effect, it is evident that the eye is not as effective in the analysis of light as the ear is in the analysis of sound. Two or more pitches sounded simultaneously will be detected by the normal ear as so many separate pitches, not as a single pitch constituting a sort of resultant of the several pitch stimuli. But with light, it is necessary to spread the illumination out into a spectrum, by means of a prism or equivalent device, before the observer can decide what the color components of any given stimulus really are. The unreliability of the unaided eye in color discrimination has created an immense amount of confusion in the formulation of theories of color vision. It is only within the last ten or fifteen years that it has become possible even to describe a color with sufficient precision and in numerical terms, so that it can be reproduced at will. In recent years, however, an adequate system of color terminology has come into existence under competent standardizing agencies, for example, the Colorimetry Committee of the Optical Society of America and the International Commission on Illumination; and instruments known as colorimeters, which analyze color stimuli in terms which facilitate the reproduction of a given color on demand, have come into common use.

Colorimeters work on the principle, which is not rigorously correct, that any color stimulus may be regarded as composed of a properly selected admixture of three spectral hues. The three spectral hues into which any color may thus be analyzed have often been termed "primary colors," and the history of theories of color vision is all cluttered up with arguments over which three colors were to be considered "primary," and whether, indeed, the number should be three. The originator in 1807 of the three-color theory was Sir Thomas Young (1773–1829). He is the versatile genius who discovered the nature of accommodation (focusing) and of astigmatism in the eye; who first established experimentally the wave-hypothesis of light; who first comprehended the nature of surface tension in liquids; who made contributions to elasticity that are acknowledged in the term "Young's modulus"; and who, by deciphering the Rosetta Stone, made it possible to read Egyptian hieroglyphics. Young considered that the "primary" hues were red, green, and blue, but it is now realized that there is a very wide range of choice as to which colors may be considered "primary." The desirability of standardization makes it more important that agreement should be reached on some set of three than that those three agree with any particular theory of color-vision. Hence, the selection of the three hues to be considered basic for colorimetric purposes has been, within certain limits, arbitrary. The choice, nevertheless, has fallen very close to the red, green, and blue which Young described as "primary."

Complementary Colors

The high point of Newton's experiments with the prism had been the discovery that white light was really the combination of all the spectral colors and that, even after being sorted out, the colors making up the solar spectrum could, by various means, be mixed together again, whereupon white light would once more appear. He said, referring to white:

> 'Tis ever compounded; and to its Composition are requisite all the aforesaid primary Colours, mix'd in a due Proportion.

But the observation that any spectral hue can be produced by mixing three "primary" colors implies that white, the combination of all spectral colors, can also be produced by a properly chosen combination of the three primaries. This may readily be tested experimentally by mixing lights which are of three hues which have been found to be "primary." An easy way to do this is to project three spots of light, one of each of the prescribed hues, and superpose them. The resultant will be found to be white. If the three are separated so that they overlap only in part, as in Figure 280 (opposite page 387), additional information may be adduced. One would expect to find that the overlapping of two hues produces an intermediate hue. This is found to be the case for the blue and green, the result being a bluish green, and for the red and blue, the result being violet. But a distinct surprise awaits in the overlapping region of red and green. There a pure yellow is seen. This verifies the statement of page 353 to the effect that yellow may be produced by mixing only red and green lights, neither of which contains any yellow component.

Further examination of the figure will bring out a final interesting point. The central white patch, being *red* + *green* + *blue*, will be seen also to be (*red* + *green*) + *blue*. But (*red* + *green*) has been seen to constitute *yellow*. Hence, the central white patch may be considered to be composed of *yellow* + *blue*. Thus, in mixing lights, the combination of yellow and blue produces white, as pointed out on page 352.

The yellow-blue paradox is now completely resolved. Light reflected from blue pigments has usually experienced absorption of only its red, orange, and yellow components (Fig. 281). Light reflected from yellow pigments has usually experienced absorption of its blues and violets. From light reflected by a mixture of the two all except the green has therefore been absorbed and the mixture consequently appears green. So much for pigments.

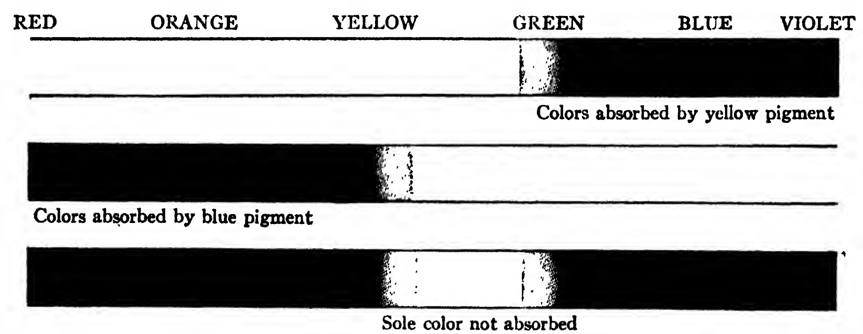


FIG. 281. SPECTRAL ABSORPTION OF YELLOW AND BLUE PIGMENTS SEPARATELY AND MIXED

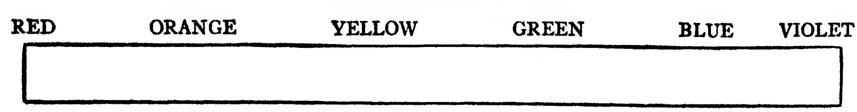


FIG. 282. THE SPECTRUM ASPECT OF MIXING BLUE AND YELLOW LIGHTS

When corresponding *lights* are mixed, however, blue contributes green, blue and violet, yellow contributes red, orange, yellow and green, and the two together contribute every color of the spectrum. The result is white (Fig. 282). The basis for the difference in mixing blue and yellow pigments and mixing yellow and blue lights should now be evident. The former is a *subtractive* process; the mixture of pigments absorbs the total of all the hues absorbed by the two pigments separately and reflects only the hue that neither pigment absorbed, namely, green. The latter, however, is clearly an *additive* process; in it colored lights are superposed, and the resultant, being the true sum of the separate hues, produces white. This may be traced back ultimately to Newton's observation that white is the sum of all colors.

But yellow and blue is not the only pair of hues which combines into white. The figure shows that the combination of green and violet has the same effect, as does the combination of red and blue-green. Any two hues which, when added together, produce white are termed *complementary colors*. These are the only three pairs which this particular trio of primaries can produce which are complementary. But there is an almost indefinite number of primary trios and, hence, a corresponding number of complementary pairs. In his *Opticks* Newton concocted a scheme for the representation of complementary pairs by arranging the spectral hues on a circle in such a way that any two diametrically opposite hues would be complementary (Fig. 283). It is to be understood that the colors progress gradually through all the intermediate spectral stages as one goes around the circle; they are not in sharply divided blocks, as would be suggested by the division into sectors. The center of this circle, being the sum of all colors, would be white. Proceeding outward, say toward the red, the color becomes at first a faint pink, then a deeper and deeper red until at the circumference the hue is "saturated." Other colors are arranged similarly.

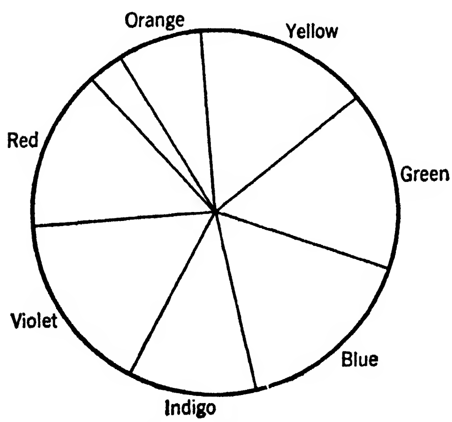


FIG. 283. NEWTON'S COLOR CIRCLE
(From his *Opticks* of 1704.)

Current Theories of Color Vision

The three-primary-color theory of Thomas Young was amplified by Hermann von Helmholtz (1821–94) fifty years later and is today commonly called the Young-Helmholtz theory. There have been other theories, notably that of Ewald Hering (1834–1918) and of Christine Ladd-Franklin

(1847–1930). No one of the theories accounts for all the known phenomena of color-vision.

The more difficult of these phenomena to deal with are the various manifestations of color-blindness. Color-blindness, as is well known, most commonly affects the ability to distinguish between red and green. It seems to be a sex-linked characteristic, being confined almost entirely to men, like baldness and hair on the face. The first scientific study of color-blindness was made by John Dalton (1766–1844), the founder of the atomic theory of chemistry. In 1794 he read before the Manchester Literary and Philosophical Society a paper entitled *Extraordinary Facts relating to the Vision of Colours*. Dalton's interest in the subject arose from his own color-blindness. He first became aware of it when as a boy, being present at a review of troops and hearing those around him exclaiming on the gorgeous effect of the red uniforms, he asked in what respect the color of a soldier's coat differed from that of the grass on which he trod.[1]

It is probably no mere coincidence that Young's theory of color-vision, the first of the series, was formulated shortly after Dalton's identification of color-blindness. In ability to correlate the various forms of normal and abnormal color-vision, the honors are about equally divided between the three theories. The great difficulty with them all is that, though each postulates a separate set of "receptors" in the retina for each primary color, there is no anatomical evidence for the existence of any of the assumed variety of receptors. A modified version of the Young-Helmholtz theory is current in physics and technology merely because it lends itself the most readily to the practices of colorimetry.

Color Photography

The increasing vogue of color photography justifies a brief description of the basic process. Figure 284 is a color camera disassembled to show the arrangement of the interior. By an arrangement of reflectors, three pictures are taken at once. The picture at R is produced by the direct light from the lens, substantially as in an ordinary camera, except that there is a red filter in front of the film. But before reaching this film, the light has passed through two lightly silvered forty-five-degree mirrors, each of which diverts about a third of the light at right angles. That portion reflected from the first mirror passes through a green filter and then exposes the film at G. That which penetrates the first mirror and is reflected from the second passes through a blue filter and exposes the film at B.[2] Thus light from a pure red object will expose only the film R, that from a pure green object only the film G, and that from a pure blue object only the film B.

The three films are then developed in the usual way as black-and-white negatives, and positives are made from them. Wherever the negative was

[1] "Dalton," *Encyclopaedia Britannica* (9th edition).

[2] This blue filter is often dispensed with. The same effect is attained by using the ordinary old-style photographic emulsion for this film. This emulsion is sensitive only to blue.

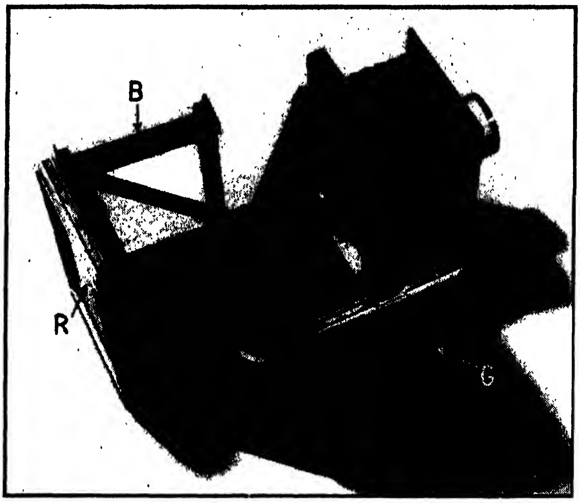


FIG. 284. INTERIOR ARRANGEMENT OF A COLOR CAMERA

transparent the positive is opaque and vice versa. These positives are then dyed, each a separate color. Film *R* is dyed a light blue. This blue affects only the "opaque" portions of the positive leaving the transparent portions uncolored. Similarly, film *G* is dyed red, in the same way, and *B* is dyed yellow. Thus in each case, where a color was bright in the object, the corresponding positive is transparent, and where a color was absent, the corresponding positive is dyed its characteristic color.

The three colored positives are next superposed, and white light is sent through them. Where the object was red, the *R* positive is transparent, the *B* positive is yellow, and the *G* positive is red. Only the red portion of the white light can get through both the red and yellow spot; hence, that portion of the image will appear red. A similar process takes place for the other two primary colors. Where the object was white, all three positives are transparent, and the white light will therefore penetrate them unaffected. Where the object was black, all three positives are colored, and none of the white light can penetrate all three colored films. Such portions of the image consequently appear black. The color reproduction is not one hundred per cent faithful to the original, of course, but it is surprisingly good.

The "Technicolor" process, common in colored motion picture practice, differs from this only in detail up to making the positives. Then, the opaque portions of the positives are processed so as to swell, producing raised surfaces. The three positives, instead of being dyed, are then run over separate colored inked rollers. The reproduction is the same as for the preceding case.

The "Kodachrome" process, though similar in basic principle, is utterly different in detail. The three photographic emulsions and two color

FIG. 285. THE MOST FAMOUS OF RAINBOW PHOTOGRAPHS
(From *Clouds*, by G. A. Clarke. London, Constable and Company, 1920.)

filters are all made up into a single five-ply film only .02 millimeter thick. This film is then developed, reversed, and dyed intact. This is an amazing process when one recalls that the three emulsions must be dyed each a different color. At present it can be done only at the Eastman factory, to which exposed Kodachrome films must be sent. The Kodachrome process produces only one copy of a film, though it is possible to re-photograph the finished film. This is unlike the two previous processes which permit an unlimited number of transparencies to be made from a single set of negatives.

The Rainbow

Perhaps the most persistent of all quests in the field of color has been the perennial inquiry about the nature of the rainbow. The main facts about the appearance of the bow have been well known for centuries. The accompanying illustration is perhaps the best rainbow photograph ever taken. The radius of the main or *primary* bow subtends an angle of between 40° and 42° at the eye of the observer — 40° at the inner or violet edge and 42° at the outer red edge, the two-degree interval spanning the intermediate colors of the spectrum. The red edge, moreover, is a purer color, more nearly saturated than is the violet edge. The space within the concavity of this primary bow may be seen to be filled with diffusely scattered white light, notably brighter than the space outside the convexity.

If conditions are favorable, an outer or *secondary* bow may be seen. The

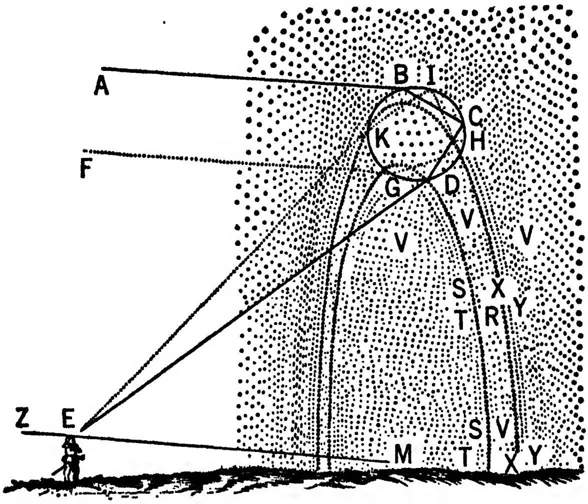


FIG. 286. DESCARTES' DIAGRAM OF THE RAINBOW
(From his *Les Météores* of 1637.)

radius of this subtends from 50° to 53°, with the red inside at 50° and the violet outside at 53°. The whole is dimmer than the primary bow. The red edge, as in the primary, is purer than the other colors, and the region of scattered white light is *outside* the bow instead of inside as with the primary. Sometimes some fine supernumerary bows appear, like ripples, just inside of the primary and outside of the secondary bow.

The most prominent of these appearances is, of course, the primary bow. Attempted explanations of this date clear back to Aristotle (9:*3*:371 *b*). He was followed in this attempt by a distinguished series of writers,[1] culminating two thousand years later in Descartes in 1637. Descartes' was the first really correct explanation of the rainbow, as far as it went. It may be taken from his own illustration (Fig. 286). The dots in that illustration represent falling drops. While they are in the region where the observer sees the primary bow, say *G*, the course of sunlight through them is as shown in the ray *ABCDE* through a magnified drop. Each drop acts like a prism; hence, when the observer at *E* looks in slightly different directions, he sees the colors of the rainbow. Suppose *DE* to be the direction of the red portion of the bow. Since blue is deviated more than red in the process of refraction, the drops at slightly lower altitude will send blue to the observer. Hence the red side of the primary bow is the outer portion and the blue the inner. The primary bow will be seen to be formed by light that has experienced two refractions and one reflection in traversing a raindrop.

The secondary bow is formed by light that has experienced two reflections

[1] Seneca (114:29), first century; Vitellio (130:*1*:83), 1269; Qutb-al-din (111:*2*:23, 762, 1018), thirteenth century; Theodorich of Saxony (58:23; 130:*1*, 58, 173), about 1310. See also Gilbert's *Annalen der Physik*, *52*, 406; Maurolycus (81:79), 1611; de Dominis (130:*1*:172), 1611.

within the drops (ray *FGHIKE*). The direction characterized by the lesser refraction will this time be lower than that characterized by the greater; hence the red will lie at the inner edge of the secondary bow instead of at the outer as in the primary. All this Descartes deduced by observing the passage of light through spherical flasks of water and correlating its behavior with the observed characteristics of the rainbow.

Descartes then sought the reason why inside of the primary bow there is a region of diffuse uncolored illumination, as is also the case outside of the secondary bow, there apparently being no reflection to the eye of the observer from raindrops between the two bows. This proved to be a laborious undertaking. Using Snel's law he is said to have traced some ten thousand rays through a drop at various angles of incidence and discovered the existence, in a spherical drop, of an angle of maximum deviation (Fig. 287), corresponding to the angle of minimum deviation later discovered by Newton in the prism (pages 343 ff.). As in the case of the prism, the intensity of the light of any particular color coming to the eye of the observer increases to a sharp maximum as the line of sight approaches the primary bow from beneath. Then it diminishes suddenly to zero. What the observer sees as a semi-circular streak of violet light, therefore, is really a whole half-disk of violet light, increasing sharply in intensity toward the edge (Fig. 288, opposite pages 386–87). He also sees superposed on this a red disk, larger by two degrees; hence, its brilliant edge projects beyond the edge of the violet disk. The intervening colors have similar effects. Inside of the "bows" where all colors overlap, the effect is, of course, that of white light. This diminishes in intensity as the line of sight leaves the bow and approaches the common center of its colored arcs. Descartes found that similar considerations applied to the secondary bow except that the luminous region was outside of the bow instead of inside. Thus the reason for the regions of diffuse illumination was identified. At the same time the reason was supplied for the greater purity of the red edges of the bows, these edges not being diluted by the presence of other overlying colors.

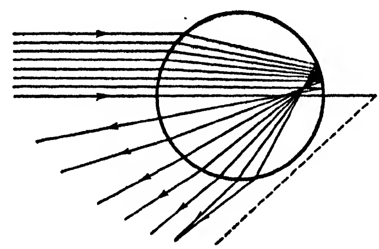


FIG. 287. THE ANGLE OF MAXIMUM DEVIATION IN REFRACTION THROUGH A SPHERE

Descartes' explanation of rainbow phenomena was a masterpiece as far as it went. In two respects, however, it was incomplete — though through no fault of Descartes. The time was not yet ripe for the final word to be said on the subject. The first omission was Descartes' ignorance of why refraction through raindrops and through prisms produced color. He had simply observed that it did and had to leave it at that. Newton, as has been seen, furnished the answer to that problem a generation later. The second omission was treatment of the supernumerary bows. This was a diffraction phenomenon, and was identified as such in 1804 by Sir Thomas Young, the originator of the Young-Helmholtz theory of color-vision. Since diffraction of light has not yet been discussed, a consideration of the supernumerary bows will have to be postponed.

Questions for Self-Examination

1. Describe and account for what appears in the overlapping regions when circles of red, green and blue lights are projected on the same screen.
2. Tell the results of adding blue and yellow lights; of mixing blue and yellow paints. Account for the difference.
3. How do color cameras produce their effects?
4. Describe and account for the appearance of the primary rainbow. How is the phenomenon of minimum deviation involved?

CHAPTER 33

Lengths of Light Waves

Light: A Corpuscle or a Wave

The foregoing treatment of the effect of reflection and refraction on light deals primarily with what light *does*, especially with ways in which it has been utilized for human convenience. Utilitarian aspects form an absorbing and important part of the story of the unfolding of any branch of science. But experience has shown that even the utilitarian possibilities become restricted unless the quest is carried beyond what an agency can *do* into the realm of what it *is*. Of course, the principal answer to any question of what it is comes through a sufficiently broad inquiry into what it does. But seldom do purely utilitarian manipulations furnish the required information.

Perplexity over the nature of light has been rife ever since the time of the Greeks. Indeed, today's perplexity on this question is strikingly similar to that of twenty-five centuries ago. In this respect light is unique among all the subtopics of physics. In the others there has either been an entire absence of tenable theories, as in gravitation, or a progressive modification of concepts until theory and observed phenomena coalesced into a more or less complete agreement. It is only in light, which, paradoxically, more than any other field of science, has lent itself to extensive and accurate study in the laboratory, that we are still struggling over almost the same apparently incompatible concepts that perplexed the Greek philosophers.

The two antagonistic views of the nature of light which have thus been pitted against each other in varying forms are that of a stream of particles or corpuscles [1] versus that of a sequence of waves. Until the time of Newton the issue between these theories lay in the field of mere controversy like most issues of the pre-scientific era — and like some even today. Newton was perhaps the first to seek experimental evidence, but even with some very good evidence at hand, he appears to have had considerable difficulty in making up his mind. From his extensive experimentation with light he concluded that certain of the implications of both theories were inescap-

[1] The term *corpuscle* was first used in scientific literature to describe the supposed atomic particles of matter by Robert Boyle in his *Defense of the Doctrine Touching the Spring and Weight of Air* (1662).

able, and tried to combine them. Perhaps the most representative short formulation of his views is as follows:[1]

> Assuming the rays of light to be small bodies emitted every way from shining substances, those, when they impinge on any reflecting or refracting superficies, must as necessarily excite vibrations in the aether as stones do in water when thrown into it.

Notwithstanding Newton's indecision, his successors attributed to him an espousal of a corpuscular theory. Perhaps there was some warrant for this, for in his efforts to "straddle the fence" he had fallen more often on that side than on the other. Bulwarked by Newton's enormous prestige, the corpuscular theory entered on almost a century of triumphal prevalence. The opinion of Hooke in 1665 comparing a spreading pulse of light to "the rings of waves on the surface of the water" (57:57) and that of Huygens in 1690 that light "spreads by spherical waves like the movement of sound" (61:20) were completely disregarded. The same was true for even Newton's own statement that rays of light, even though regarded as small bodies, "must as necessarily excite vibrations in the aether as stones do in water when thrown into it."

Newton's Rings

Newton had made a very thoroughgoing study of the color patterns seen when light traversed thin films. His first films were of air, formed by laying the convex side of an extremely long-focus lens on a plane piece of glass. Figure 289 shows what Newton saw when he looked through the film of air formed in this way. The central portion of the pattern was bright. The reason for this was not entirely understood and could not be without the wave concept. It will be developed presently. But at a certain distance from the point of contact was seen the smallest of several concentric dark rings. We now know that this marked a region where the separation of the plates was one quarter of the wave-length of the light being used. That portion of the light which was twice reflected in the air film consequently traveled a half wave-length further than that transmitted directly. Hence when the two rays combined, they were in opposite phases of vibration and would destructively interfere with each other. The same condition obtained wherever the glass surfaces were separated by three quarters wave-length, five quarters, seven quarters, etc. Between these dark rings were bright rings, produced by path differences of a full wave-length, two wave-lengths, three wave-lengths, etc., the corresponding distances between the plates being one half, two halves, three halves wave-length, etc.

The Interpretation of Color in Newton's Rings

Naturally, the size of a given ring depended on the wave-length of the

[1] *Philosophical Transactions*, 7, 5087 (1672).

light used, the ring being larger for red light than for blue light. Newton observed the dependence of size upon color. He said (90:208):

> I found the circles which the red light made to be manifestly bigger than those which were made by the blue and violet. And it was very pleasant to see them gradually swell or contract accordingly as the Colour of the Light was changed.

Thus Newton made an association which today is interpreted as inescapable evidence that color is associated with wave-lengths. But he came even closer to the wave hypothesis than this. Trying to explain the division of light into reflected and transmitted (or absorbed) portions whenever it strikes a surface of any kind, he imagined every beam of light to be divided into sections of uniform length, one section being easily transmitted, the next easily reflected, and so on in alternation. In his own words (90:278–85),

> Every Ray of Light in its passage through any refracting Surface is put into a certain transient Constitution or State, which in the progress of the Ray returns at equal Intervals, and disposes the Ray at every return to be easily transmitted through the next refracting Surface, and between the returns to be easily reflected by it....
>
> The returns of the disposition of any Ray to be reflected I will call its *Fits of easy Reflection*, and those of its disposition to be transmitted its *Fits of easy Transmission*, and the space it passes between every return and the next return, the *Interval* of its Fits....
>
> If the Rays which paint the Colour in the Confine of yellow and orange pass perpendicularly out of any Medium into Air, the Intervals of their Fits of easy Reflection are the $\frac{1}{89000}$th part of an Inch. And of the same length are the Intervals of their Fits of easy Transmission.

The context shows that Newton took the thickness of the film at the first bright ring as the measure of the "Interval of the Fits," which we should now say marks half the wave-length of light. The yellow-orange part of the spectrum is now known to possess a half wave-length of $\frac{1}{85000}$th of an inch. Thus Newton not only postulated a periodic nature for light, he measured the periodicity and deduced a figure within 5 per cent of the modern value of the wave-length.

Change of Phase in Newton's Rings

Nor was this all. Newton observed the rings in the reflected portion of the light as well as in the transmitted portion (Fig. 291). Their visibility was, indeed, much greater by reflected light. He observed that the center of this system of rings was dark, instead of light as before. Though he did not recognize it, this was almost as clear a manifestation of a property of waves as was the periodicity which he had previously deduced. It provided an excellent example of change of phase at reflection, one of the basic properties of wave motion.

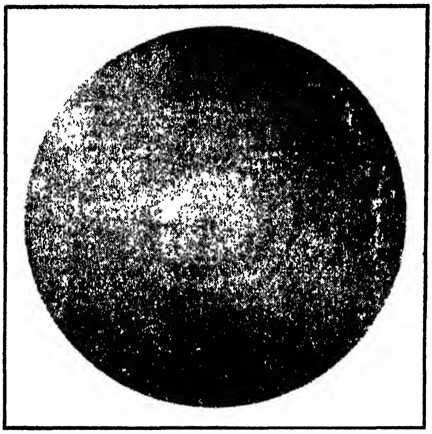


FIG. 289. NEWTON'S RINGS BY TRANSMITTED LIGHT

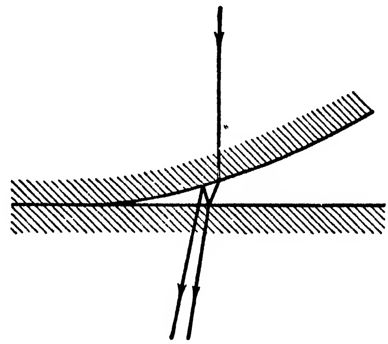


FIG. 290. THE PRODUCTION OF NEWTON'S RINGS BY TRANSMITTED LIGHT

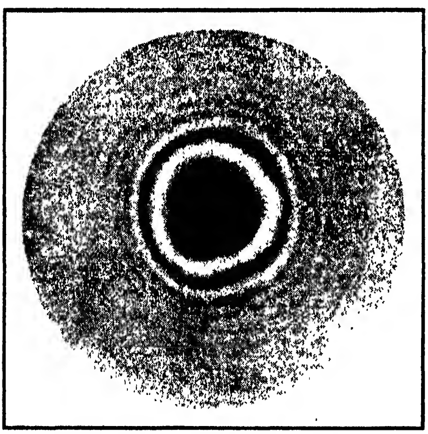


FIG. 291. NEWTON'S RINGS BY REFLECTED LIGHT

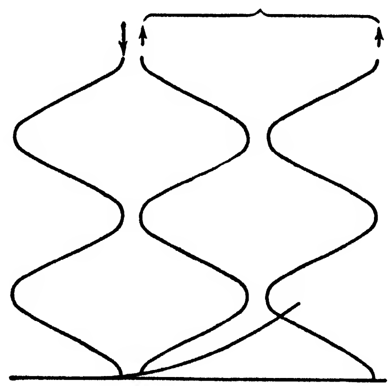


FIG. 292. DESTRUCTIVE INTERFERENCE DUE TO CHANGE OF PHASE AT REFLECTION

The phenomenon of change of phase at the reflection of a mechanical wave and of a sound wave has already been encountered. Here it occurs in light. The parallel with the case of sound is especially close and need not be traced in detail. Reflection from the first of two surfaces in contact involves a 180-degree change, that from the second involves no change of phase (Fig. 292). If the distance between the surfaces producing the reflection is small in comparison with a wave-length of light, the combination of the two reflected waves will produce destructive interference. Hence comes the black center (of the Newton's-ring pattern), which if it were not for the reversal of phase by reflection at the first surface would be bright.

Young's Identification of Interference

Newton's heroic attempt at interpretation of his rings is what lends significance to the phenomenon. But this significance is immensely increased by an extension, both to the phenomenon and its interpretation, made by Thomas Young (1773–1829) more than a century and a half later.

At the beginning of the nineteenth century a London physician named Thomas Young reported to the Royal Society some experiments which, when their full import was realized, turned the tide of scientific opinion from the corpuscular to the wave theory of light. Young had done extensive experimentation with sound, and had apparently been deeply impressed with the phenomenon of beats. The fact that two sounds could so combine as to produce silence was most easily explainable on the basis of their wave properties and had by common consent established the wave theory of sound. In connection with his medical studies Young had already discovered the mechanism of accommodation (focusing) of the eye, and through that channel had become interested in the study of light itself. Observing some indications of a phenomenon in light comparable to beats in sound, he was led to question the prevailing corpuscular theory of light and ultimately to espouse the wave theory which was, in consequence, to prevail for more than a century. The work for which he is best known was the experimental demonstration that light exhibited attributes of wave motion to an extent that justified abandoning the prevailing corpuscular theory attributed to Newton.

The experimental basis for the wave hypothesis of light as Young formulated it was *interference*. The fact has already been observed (page 245) that two trains of water waves may be so superposed that in certain regions the troughs of one train will lie continuously on the crests of another, thereby producing zero disturbance. Intervening regions, instead of being characterized by quiescence, exhibit increased disturbance on account of

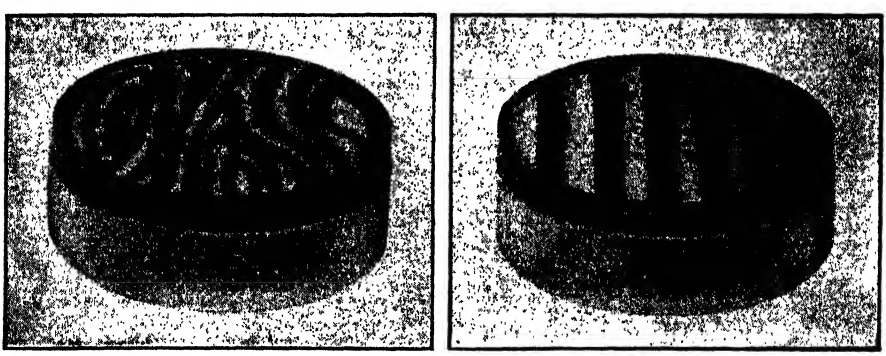


FIG. 293. INTERFERENCE IN A WEDGE-SHAPED FILM OF AIR BETWEEN TWO GLASS PLATES

At left, window glass; at right, plates with "optically plane" surfaces. (Courtesy of Bausch and Lomb Optical Company.)

combinations of waves occurring by the continuous superposition of the crests of one train on the crests of the other, and similarly for the troughs. *Destructive interference* is said to occur between the two trains of waves in the former case and *constructive* interference in the latter. Similarly, two sound waves may be so combined as to produce alternate regions of silence and enhanced sound. The phenomenon of interference, of which the foregoing are familiar examples, is easily comprehensible in the case of combining waves, but would be utterly incomprehensible in the case of combining streams of particles. So when Young demonstrated in 1803 that two beams of light could, under properly controlled conditions, be made to combine in such a way as to produce alternate regions of darkness and light, he was rightly considered to have identified in light a characteristic property of waves.

The First Determination of Wave-Length of Light

Newton had made his measurements yield some rudimentary information on the interval between the "fits of easy reflection" for yellow-orange. Young carried the same idea much further, deducing the wave-length ranges of the conventional colors of the entire spectrum. His values, reproduced below, agree very well with those assigned today. For convenience in comparison, his values (stated in inches) are converted into centimeters, the unit upon which modern wave-length tables are based.

WAVE-LENGTH RANGES OF SPECTRAL COLORS

From Young's measurements on Newton's rings.

Color	Inches (in millionths)	Centimeters (in millionths)
red	24.6–26.6	62.5–67.6
orange	23.5–26.6	59.7–62.5
yellow	21.9–23.5	55.7–59.7
green	20.3–21.9	51.6–55.7
blue	18.9–20.3	48.1–51.6
violet	16.7–18.9	42.4–48.1

Young took his data in large measure from Newton's own measurements of the diameters of interference rings, checking Newton's statements by observations of his own. The thickness of film which Newton had found at the first yellow-orange ring was $\frac{1}{89000}$ inch, which, when multiplied by 2, gives a wave-length of 22.5 millionths (page 365). This may be compared with the corresponding values in Young's table above.

The "Black Spot" in Liquid Films

Young's next observation, like those on Newton's rings, was simply an improvement and extension of what had been done before. Newton had made rather careful observations on colors of soap bubbles. Though bubbles did not lend themselves readily to measurements of thickness as did air films between convex glass surfaces, the colors were more brilliant. Newton describes his observations in these words (90:214):

> The Colours emerged in a very regular order, like so many concentrick Rings encompassing the top of the Bubble. And as the Bubble grew thinner by the continual subsiding of the Water, these Rings dilated slowly and overspread the whole Bubble, descending in order to the bottom of it, where they vanished successively. In the mean while after all the Colours were emerged at the top, there grew in the center of the Rings a small round black Spot... which continually dilated itself till it became sometimes more than $\frac{1}{2}$ or $\frac{3}{4}$ of an Inch in breadth before the Bubble broke.

This observation of the "black spot" in a liquid film was of considerable importance. Young himself observed the same thing later, under circumstances which made more extended examination possible.

He described his observation in these words (134:368):

> When a film of soapy water is stretched over a wine glass, and placed in a vertical position, its upper edge becomes extremely thin, and appears nearly black, while the parts below are divided by horizontal lines into a series of coloured bands.

Figure 294 is a photograph of Young's own drawing of the appearance of a soap film under these circumstances. Though the appearance of a soap film is somewhat difficult to demonstrate to large groups, Young's demonstration has become somewhat of a classic. The "black spot" is of particular interest. The name is rather misleading. The top of the film is not black in the sense of being opaquely so. On the contrary, the transparency is even more marked there than elsewhere. It is so called because, when the film is "highlighted" by strong reflection, that portion fails to share in the general brilliance and hence appears black by contrast. This absence of reflection is the same effect that appeared at the center of Newton's rings when seen in reflected light, and is due to the same cause except that now it is the back surface of the film which produces the 180-degree change of phase instead of the front.

Reduction of Reflection by Thin Films

An interesting application has recently been made of this absence of reflection from a film which is sufficiently thin. Undesirable reflections are often encountered, such as the reflection of a window or of an electric light on a near-by glass-covered picture which makes the picture nearly or quite invisible until the angle of view is changed. If the reflecting glass surface could be covered with a thin film, most of the reflection would be eliminated. The difficulty has lain in producing films which were sufficiently thin and yet rugged enough to survive. This was first successfully done in 1939. Figure 295 shows the increased visibility produced by this treatment of glass. The right third is uncoated, the left third has a coating of the optimum thickness, and the center portion has a coating of inter-

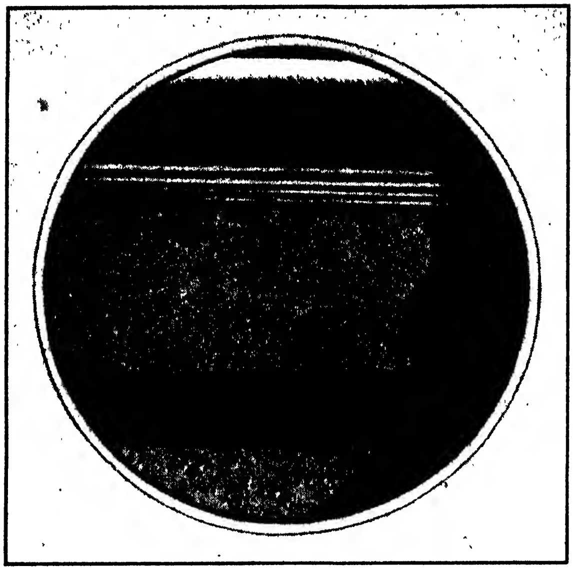


FIG. 294. PHOTOGRAPH OF YOUNG'S DRAWING (THE ORIGINAL IN COLOR) OF A VERTICAL SOAP FILM ON THE END OF A TUBE
(From his *Lectures on Natural Philosophy* of 1807.)

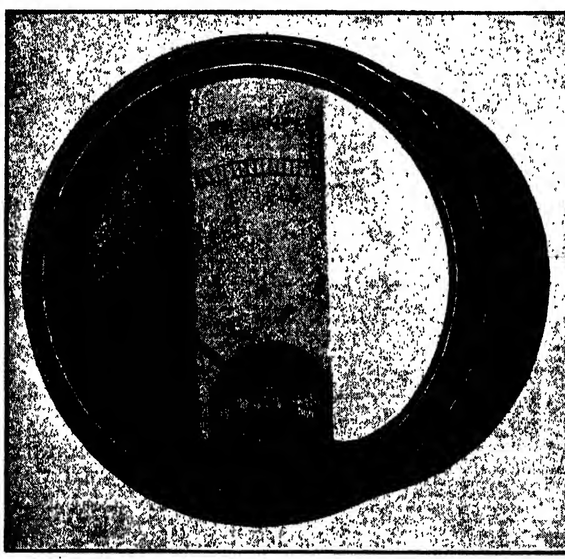


FIG. 295. VISIBILITY THROUGH THE COATED, PARTIALLY COATED, AND UNCOATED PORTIONS OF THE GLASS COVER OF AN ELECTRICAL INSTRUMENT
(Courtesy of Dr. Katharine Blodgett, of the General Electric Company.)

mediate thickness. This technique will solve the troublesome problem of loss of light by reflection at surfaces of the lenses used in photography and in picture projection.

Grimaldi's Observation of Diffraction

The work of Newton and Young meshed at still another and very crucial point. In 1665 was published the work of a friar named Francesco Grimaldi (1618–63), who had noticed some peculiarities in the behavior of light which had passed successively through two tiny apertures. Grimaldi had given the name *diffraction* to the new phenomenon. It was produced, among other ways, by sending a beam of sunlight successively through two tiny apertures, as in Figure 296, and observing what happened to it. He described his observation as follows (77:297):

> [When the light is incident on] a smooth white surface it will show an illuminated base *IK* notably greater than the rays would make which are transmitted in straight lines through the two holes. . . . This is proved as often as the experiment is tried by observing how great the base *IK* is in fact and deducing by calculation how great the base *NO* ought to be which is formed by the direct rays. . . .
>
> Further it should not be omitted that the illuminated base *IK* appears in the middle suffused with pure light, and at either extremity its light is colored.

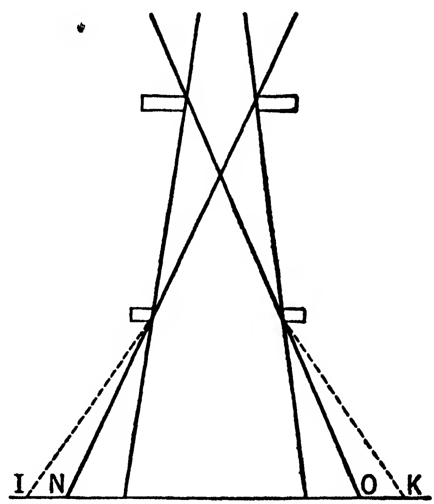


FIG. 296. GRIMALDI'S OBSERVATION OF DIFFRACTION
(From his *Physico-mathesis de Lumine* of 1665.)

Newton not only referred to Grimaldi's experiments (90:317 ff.), but he repeated and improved upon them. Apparently prompted by Grimaldi's observation of color effects, Newton compared the widths of the luminous patches produced by the lights of different portions of the spectrum, and remarked (90:336):

> Comparing the Fringes made in the several colour'd Lights, I found that those made in the red Light were the largest, those made in the violet were the least, and those made in the green were of a middle bigness.

Here, too, as we see it, Newton was well within the domain of the wave hypothesis, for as will be shown shortly, the scale on which diffraction effects occur should, on that hypothesis, be proportional to the wavelengths, as Newton found it to be. But even with a man of Newton's

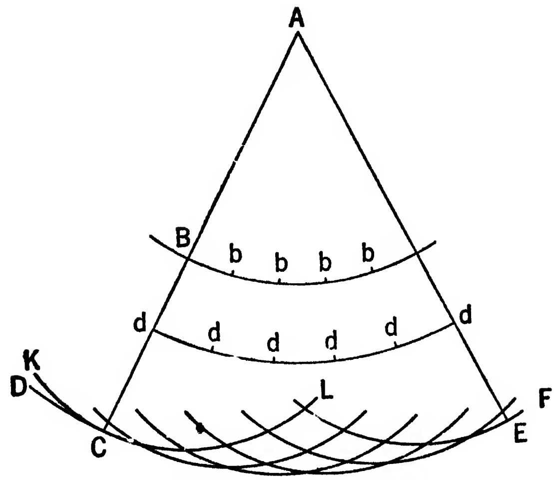


FIG. 297. HUYGENS' PRINCIPLE
(Redrawn from his *Traité de la Lumière* of 1690.)

caliber, the preoccupation with one theory blinded him to the significance of the evidence pointing toward another; and it was left for Thomas Young to find the most convincing evidence on the wave hypothesis.

Huygens' Principle

Young utilized in this connection a principle which Huygens had formulated more than a hundred years earlier, but which had remained largely unnoticed. Huygens had stated his principle in the following words (61:19):

> Each particle of matter in which a wave spreads ought not to communicate its motion only to the next particle which is in the straight line drawn from the luminous point, but it also imparts some of it necessarily to all the others which touch it and which oppose themselves to its movement. So it arises that around each particle there is made a wave of which that particle is the centre. Thus if *DCF* [Fig. 297] is a wave emanating from the luminous point *A*, which is its centre, the particle *B*, one of those comprised within the sphere *DCF*, will have made its particular or partial wave *KCL*, which will touch the wave *DCF* at *C* at the same moment that the principal wave emanating from the point *A* has arrived at *DCF*; and it is clear that it will be only the region of *C* of the wave *KCL* which will touch the wave *DCF*, to wit, that which is in the straight line drawn through *AB*. Similarly the other particles of the sphere *DCF*, such as *bb*, *dd*, etc., will each make its own wave. But each of these waves can be infinitely feeble only as compared with the wave *DCF*, to the composition of which all the others contribute by the part of their surface which is most distant from the centre *A*.

The principle which Huygens here sets forth is now commonly stated somewhat more succinctly as follows: *Every point on an advancing wave front acts as a source from which secondary waves continually spread.* The masking of the sideways-moving portions of the secondary waves by the more energetic forward-moving portion, upon which Huygens commented, may be reduced by causing a wave to pass through a small aperture. The sideward components will then become visible. The passage of a ripple through an aperture (Fig. 299) illustrates this effect. It was a similar spreading of beams of light, which under Huygens' principle seems so natural, that gave rise to the phenomena which Grimaldi had classified under the name *diffraction* in his publication of 1665.

Young Confirms Huygens' Principle for Light

Young took advantage of this spreading of light by diffraction in another experiment, the one for which he is probably best known. If ripples originate in two closely adjacent sources, their overlapping regions will exhibit alternate areas of constructive and destructive interference (Fig. 300). Young obtained a similar effect with sunlight, which he reported to the Royal Society in 1803. A cone of sunlight, spreading from a tiny aperture in the screen S (Fig. 301), became incident on two small apertures, a and b, very closely adjacent to each other in a second screen. The two cones spreading from these apertures overlapped. The interference pattern in this overlapping region was made evident by interpolating a third screen, on which appeared parallel light and dark bands which exhibited colors. Young described his observations as fol lows (134:365):

> When the two beams are received on a surface placed so as to intercept them, their light is divided by dark stripes into portions nearly equal. The middle of the portions is always light, and the bright stripes on each side are at such distances, that the light, coming to them from one of the apertures, must have passed through a longer space than that which comes from the other, by an interval which is equal to the breadth of one, two, three or more of the supposed undulations [Fig. 298] while the intervening dark spaces correspond to a difference of half a supposed undulation, of one and a half, of two and a half, or more.

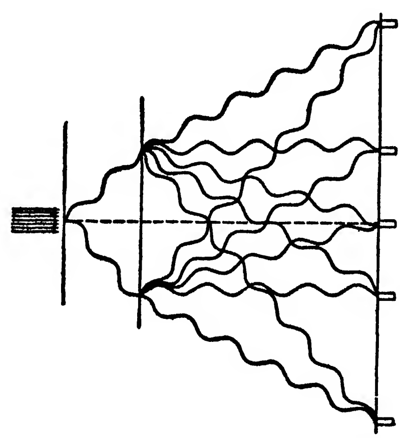


FIG. 298. CONDITION FOR REINFORCEMENT OF DIFFRACTED LIGHT WAVES

Wave-Length from Double-Slit Diffraction

The observation of interference bands produced in this way, added to the thin-film observations, furnished enormous support for the wave hypothesis. But one of the most significant parts of the undertaking was causing the observations to yield the values of the wave-lengths of the

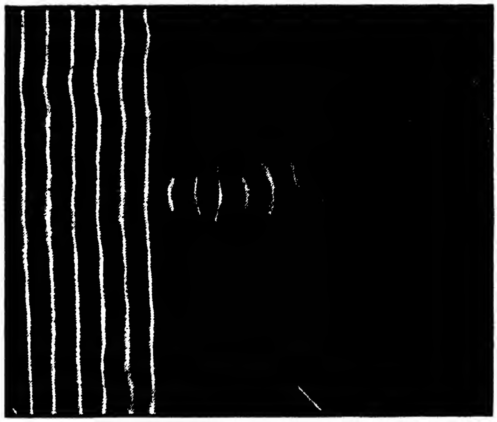


FIG. 299. DIFFRACTION OF A RIPPLE UPON PASSING THROUGH AN APERTURE
(From *General Physics for Colleges*, by Webster, Farwell, and Drew. The Century Company, publishers.)

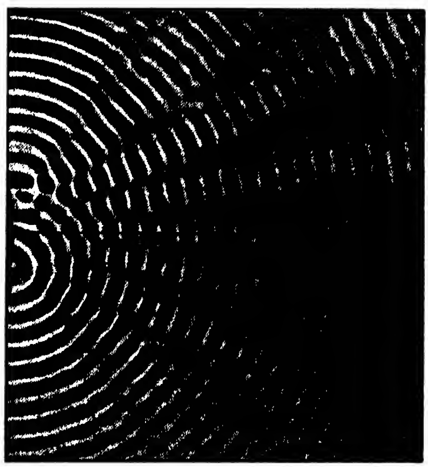


FIG. 300. INTERFERENCE PATTERN BETWEEN TWO TRAINS OF RIPPLES
(From *General Physics for Colleges*, by Webster, Farwell, and Drew. The Century Company, publishers.)

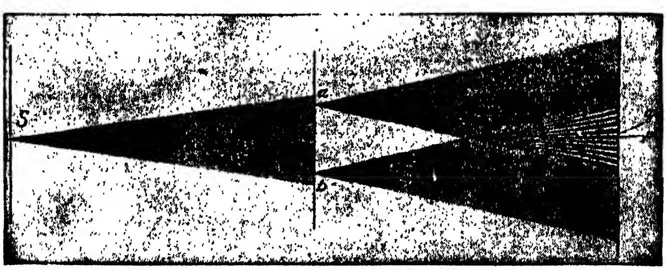


FIG. 301. INTERFERENCE IN THE OVERLAPPING REGION OF TWO DIFFRACTED BEAMS OF LIGHT
(From *Studies on Optics*, by A. A. Michelson. The University of Chicago Press, publishers.)

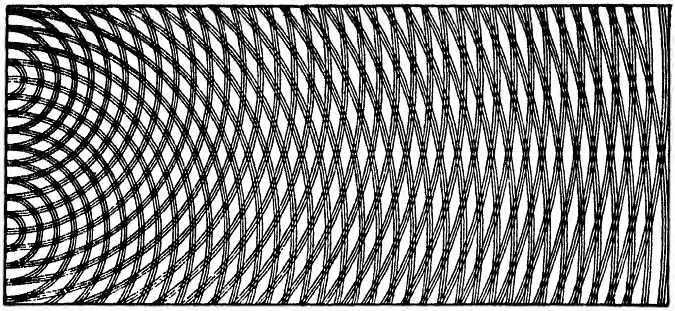


FIG. 302. YOUNG'S REPRESENTATION OF OVERLAPPING WAVES
(From his *Lectures on Natural Philosophy* of 1807.)

FIG. 303. APPEARANCE OF INTERFERENCE PATTERN FROM TWO APERTURES
(From *Studies on Optics*, by A. A. Michelson. The University of Chicago Press, publishers.)

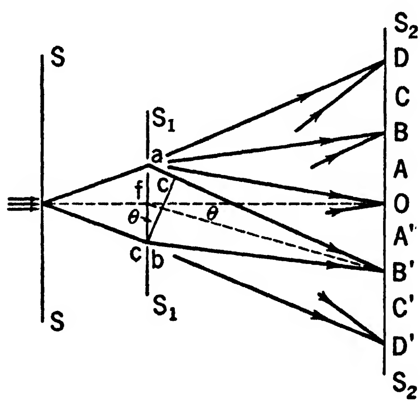


FIG. 304. THE GEOMETRY OF INTERFERENCE BETWEEN DIFFRACTED BEAMS

FIG. 305. APPEARANCE OF DIFFRACTION PATTERN PRODUCED BY PASSAGE OF LIGHT THROUGH A SLIT
(From *Studies on Optics*, by A. A. Michelson. The University of Chicago Press, publishers.)

light used in producing the interference patterns. Young had extracted this information from measurements on the interference patterns in thin films. His ability to do so in this case also and the fact that the two sets of results checked each other excellently could scarcely fail to lend strength to the hypothesis which Young was invoking to account for his results. It is no wonder that Young said, in introducing his work to the Royal Society:[1]

> The proposition, on which I mean to insist at present, is simply this, that fringes of colours are produced by the interference of two portions of light; and I think it will not be denied by the most prejudiced, that the assertion is proved by the experiments I am about to relate.

Referring to the determinations of wave-length in this way he said (134:365):

> It appears that the breadth of the undulations constituting the extreme red light must be supposed to be, in air, about one 36 thousandth of an inch, and those of the extreme violet about one 60 thousandth.

Reference to Figure 304[2] will show how these wave-lengths could be deduced. The point O is a region of constructive interference for all wave-lengths. Light from the two apertures travels an equal distance in arriving at O and hence will be "in step" upon arrival. If at B' (and B) are found the first regions of constructive interference for a particular wave-length, say red, then ac must represent a difference of path of one wave-length between the two beams reaching B' which originate from the two apertures a and b respectively (compare Fig. 298). If the distance ac can be deduced, the wave-length of red light will become known.

This can readily be done with the aid of elementary geometry. The angle OfB' (Fig. 304) is very small and nearly equal to abc. (For an interference band to be large enough to be seen, OB' would usually be less than $\frac{1}{1000}$ of Of.) Hence with tolerable accuracy it can be said that

$$\frac{ac}{ab} = \frac{OB'}{Of}. \tag{1}$$

But when B' represents the first bright region off of the axis, the distance ac represents one wave-length of light, which will be termed λ. Also ab represents the distance d between apertures, Of the distance L from the apertures to the screen and OB' the width w of one interference band. Hence,

$$\frac{\lambda}{d} = \frac{w}{L} \quad \text{or} \quad \lambda = \frac{wd}{L}. \tag{2}$$

[1] *Philosophical Transactions*, *94*, 1 (1804).

[2] In this and succeeding diagrams dealing with diffraction, the angles are enormously exaggerated. Their actual values would be, almost without exception, small fractions of a degree. It is the smallness of these angles that permits what otherwise appear to be unjustifiable substitutions of chords for arcs, angles for sines or tangents, and similar approximations used in the derivations.

Since w, d, and L may all be directly measured, it becomes possible to determine the value of λ, the wave-length of the light. This was Young's deduction. It gave him values of wave-lengths wherewith to check those secured by interference in thin films. The agreement was such as ultimately to vindicate his stand for the wave hypothesis.

Fresnel and the Single-Slit Pattern

Young had observed that when light passed through a single small aperture, and indeed when it passed only one sharp edge of an obstacle, alternate light and dark bands were produced as in the case of two apertures, though the bands were different in many details. He erroneously attributed these bands to interference between light reflected from the edges of apertures, or between direct beams and the light so reflected. At this point he was corrected by a young French military engineer, named Augustin Fresnel (1788–1823). Fresnel had established a reputation in the field of light, had independently discovered the phenomena which Young had been studying, and had, in fact, carried the subject much further. When Young's work came to his attention it stimulated an extended friendly correspondence between the two men. Fresnel pointed out Young's error and gave a correct account of the cause of these diffraction phenomena. His treatment of the case of the single aperture is a classic which may well be quoted verbatim. But first the diffraction pattern produced by a single slit-formed aperture should be observed (Figure 305). It should be compared with the pattern due to two slits (Fig. 303). The widths of the bands in this case are not uniform, and the intensity diminishes rapidly on each side of the center. Fresnel accounted for these appearances as follows (31:114):

> Let AG (Fig. 306) be the aperture through which the light passes. I shall at first suppose that it is sufficiently narrow for the dark bands of the first order to fall inside the geometrical shadow of the screen. Let P be the darkest point in one of these two bands; it is then easily seen that this must correspond to a difference of one whole wave-length between the two extreme rays AP and GP. (To show this) let us imagine another ray, PI, drawn in such a way that its length shall be a mean between the other two.... Corresponding elements of the aperture on each side of I send to the point P vibrations in exactly opposite phases, so that these must annul each other.
>
> By the same reasoning it is easily seen that the darkest points in the other dark

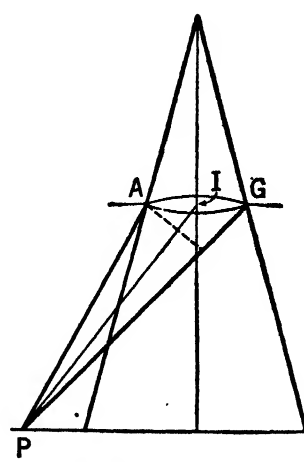


FIG. 306. FRESNEL'S TREATMENT OF DIFFRACTION DUE TO A SINGLE APERTURE

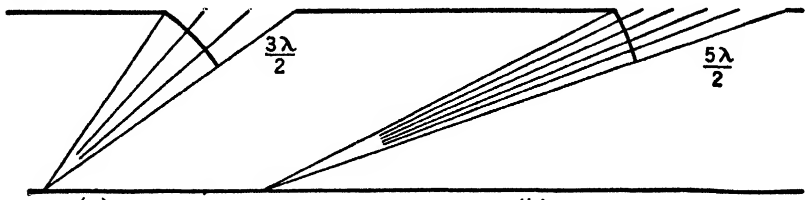


FIG. 307. CONDITIONS OF FIRST TWO SUBSIDIARY MAXIMA IN DIFFRACTION

> bands also correspond to differences of an even number of half wave-lengths between the two rays which come from the edges of the aperture; and in like manner, the brightest points of the bright bands correspond to differences of an uneven number of half wave-lengths.... This is true with the exception of the point at the middle which must be bright.

The last sentence accounts for the fact that the central band of the single-slit diffraction pattern is twice as wide as the bands on either side. To account for the diminution in intensity of the bands upon receding from the center, imagine the aperture divided into an odd number of subdivisions. At a point on the screen, for example, such that the diffracted light travels $1\frac{1}{2}$ wave-lengths further from one edge of the aperture than from the other (Fig. 307 (*a*)), there will be, by Fresnel's preceding observations, a region of maximum brightness. But this maximum will be much less bright than the adjacent central bright band. For if the aperture be in imagination divided into three subdivisions, the path difference between corresponding points in adjacent subdivisions will be one half wave-length. Between the outside edges of either adjacent pair of subdivisions, the path difference will be one wave-length; and hence light from these two areas will nullify each other just as it did at point P (Fig. 306) in Fresnel's observations. Hence, only the remaining third of the area is effective in producing light. Similarly, at a point on the screen such that the over-all difference of path is $2\frac{1}{2}$ wave-lengths (Fig. 307 (*b*)), only one fifth of the area of the aperture is effective, etc. Figure 305 indicates the appearance of the diffraction pattern produced by a single slit. This should be compared again with Figure 303. The comparison will show that the interference pattern due to two slits is simply superposed upon the diffraction pattern due to each.

Diffraction and Resolving Power

If the diffracting aperture is a circular opening, the pattern of a point source will consist of concentric circles, fading out rapidly with radial distance from the center. Such patterns are always formed whenever light passes through circular apertures, such as the lenses of microscopes and telescopes. The lenses themselves are powerless to correct the spreading produced by diffraction, as astronomers and biologists are well aware. For

FIG. 308. ARTIFICIAL DOUBLE STARS AT DIFFERENT SEPARATIONS
(Courtesy of Dr. J. A. Anderson, of the Mount Wilson Observatory.)

this reason, the spread-out images of two points which are very closely adjacent often overlap sufficiently to make it difficult or impossible to tell whether two points or one are under observation. Close double stars (Fig. 308) and the fine structure of cells are often indistinguishable from this cause. In such cases the adjacent objects are said not to be "resolved" by the telescope or microscope. The "limit of resolution" of the instrument has been passed. To be able to separate such a pair of objects, either the aperture producing the diffraction (and containing the objective lens) must be made larger or the wave-length of the light must be made shorter. Hence, blue light, or even ultra-violet light, is sometimes used in microscopy when maximum resolving power is required.

Multiple-Slit Diffraction; the Grating

If parallel light from a point source is acted upon by a converging lens, it will be brought to a focus. If a series of closely spaced parallel apertures (a *grating*) be interposed in front of the lens, the original image will not be affected, except to be reduced in intensity. But there will appear in addition a number of subsidiary images on each side of the central image, the successive images diminishing rapidly in intensity in each direction from the middle. The reason for this multiplication of images is not far to seek.

In addition to the original wave-front (Fig. 309 (*a*)) the grating now produces additional wave-fronts such as that of Figure 309 (*b*) and 309 (*c*)

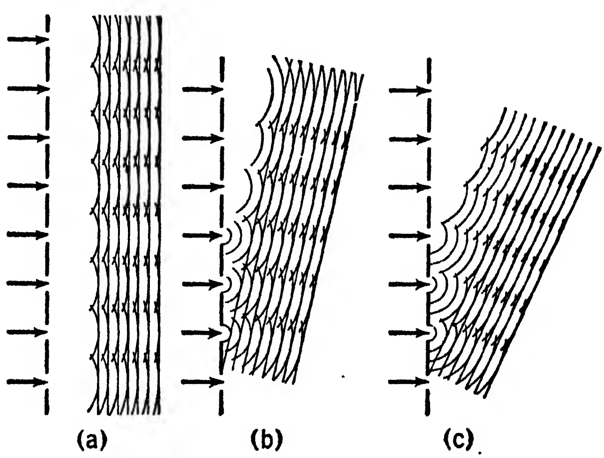


FIG. 309. THE ACTION OF A GRATING

Each of these is brought to a focus by the lens and forms its own image. But the direction of each of these subsidiary wave-fronts will depend upon the wave-length of the incident light. White light will thus produce a series of continuous spectra. In this the grating is comparable to the prism in its action on light. Gratings are indeed more extensively used than prisms to produce spectra.

But to be effective, a grating must consist of many thousands of lines ruled extremely close together. Rulings spaced fifteen thousand to the inch are fairly common. Moreover, the accuracy of such rulings must be extremely high. The construction and operation of ruling-engines for making gratings is one of the most exacting of the useful arts. The difficulty of ruling gratings on glass has led to the practice of ruling them on polished metal; the combination of reflection and diffraction from the metal does what combined transmission and diffraction accomplish through the glass grating. The first man to make extensive use of gratings in the production of spectra was Joseph Fraunhofer (1787–1826), whose work will be more extensively described in the following chapter. He used them first to measure the wave-length of light from a sodium lamp.[1] His gratings were very primitive, consisting of fine wires stretched across the space between two parallel screws, the spacing being determined by the pitch of the screws. Even with this primitive equipment, however, he secured good results, four of his most concordant observations being in millimeters, .0005891, .0005894, .0005891, and .0005897. The mean .0005893 is almost exactly the value now assigned.

The method of using a grating to determine the wave-length of light is precisely the same as Thomas Young's use of the double slit. In fact a double-slit is a rudimentary grating consisting of two rulings. The grating has until recent years been by far the most accurate instrument for the measurement of wave-lengths. Thus Young, though he never dreamed of a real grating, devised the principle which later made possible the precise measurement of wave-lengths.

The finest of Fraunhofer's gratings had a spacing of .0528 millimeter between the centers of adjacent wires. Thus in Figure 304 the sine of the angle θ had the value $\lambda/.0528$. Fraunhofer's mean value for θ was such as to give the wave-length of sodium light correct within .005 per cent. This astonishing degree of accuracy was prophetic of what the grating was to accomplish in the following hundred years. It developed, indeed, into the most powerful tool available for the study of spectra.

Characteristics of Grating Spectra

A grating produces many spectra, instead of just one spectrum as does a prism, the *order* being identified as first, second, etc., each way from the center. Hence, each of the spectra produced by a grating is less bright than the single spectrum produced by a prism. But gratings possess ad-

[1] *Philosophical Magazine* (5), *26*, 245 (1888).

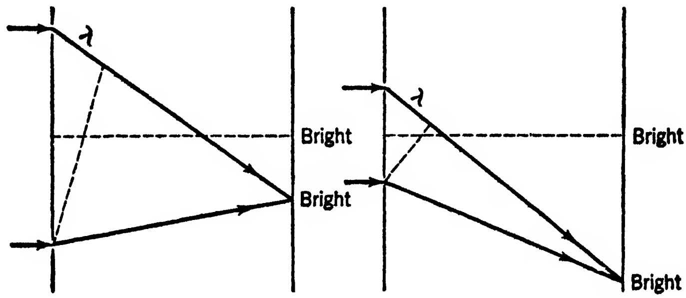


FIG. 310. DEPENDENCE OF WIDTH OF INTERFERENCE BANDS ON DISTANCE BETWEEN SLITS

vantages which more than offset this handicap. They do not produce the absorption which is characteristic of the passage of light through considerable thicknesses of glass or quartz. A very small grating may be capable of spreading out a spectrum through as wide an angle as would require a bulky array of prisms. Most useful of all, perhaps, is a certain uniformity of grating spectra which is not characteristic of prism spectra. Within certain limits a spectrum formed by one grating is exactly similar to that formed by any other, the two being exact copies of each other except for a difference in scale. Moreover, any distance measured along a grating spectrum is approximately proportional to the corresponding wave-length range. Spectra produced by two prisms, on the other hand, cannot safely be compared with each other even when reduced to the same length. The relative widths of the blue and the green sections, for example, may be quite different or, in extreme and somewhat unusual cases, the two sections might actually be reversed in relative position. Also, it is characteristic of a spectrum produced by a prism that the long-wave-length end (the red) is compressed and the short-wave-length end (the violet) is greatly expanded, in comparison with the corresponding spectrum produced by a grating. The grating spectrum is termed a *normal* spectrum, in consequence of the very useful approximate proportionality between distances along the spectrum and corresponding wave-length ranges.

The "coarseness" or "fineness" of an interference pattern produced by two slits depends upon the distance between the slits, a large distance producing a fine concentrated pattern, a short distance producing a larger-scaled pattern (Fig. 310). This is also true of the interference pattern produced by a grating. Thus the angular spread per unit of wave-length range, known as the *dispersion*, will depend solely on the grating space, in an inverse proportion. Thus the dispersion of a grating depends entirely on the number of rulings per unit width and not at all upon the total number of lines.

The so-called *resolving power* of a grating depends, on the other hand, solely on the total number of rulings and not upon the spacing. This term, resolving power, as used here, is closely analogous in meaning to the same term as used on page 379. A high resolving power implies in this case high ability of a grating to produce separate images of closely adjacent wave-lengths in a spectrum. Fraunhofer's gratings did not possess sufficient resolving power to produce separate images of the sodium line, now known to be double. Hence, his wave-length was .0005893 mm, the mean of .0005890 mm and .0005896 mm, the wave-lengths of the two components. The numerical measure of the resolving power of a grating or a prism is the ratio of the mean wave-length to the difference between the wave-lengths which are barely resolvable. Thus it requires a resolving power of $\frac{5893}{6}$, or approximately 1000, to produce separate images of the two components of the sodium line.

The Ruling of Gratings

The "ruling" of gratings is a delicate, as well as a useful art. A prodigious amount of labor has gone into the construction of "ruling engines." The heart of the machine is the screw which advances the ruling diamond the required interval between successive strokes. Making this screw requires months of skilled labor. The completed ruling engine must be so mounted as to be free from jars, and stringent measures are taken to maintain a constant temperature while a ruling is in progress. To rule gratings as much as six inches wide, with fifteen thousand to twenty-five thousand lines to the inch, all to the required high accuracy and uniformity, is a major undertaking. Such gratings are literally priceless, for so large is the element of luck in the manufacture of gratings that the best gratings are few and are invariably retained for the use of the maker, only the less perfect ones being placed on sale.

The game has been distinctly "worth the candle," however. It was in spectroscopy that the earliest of the real precision measurements were made in the field of physical sciences. More than anything else, physics owes its reputation for being the most exact of the exact sciences to the development of precision measurements made possible by the diffraction grating. Thus an obscure phenomenon, neglected for a century after Grimaldi's early discovery, and even then coming only slowly into scientific use, ultimately developed, as do many similar phenomena, into one of the major factors in the growth of a science.

Questions for Self-Examination

1. What was Newton's opinion on the controversial question of the nature of light?
2. Describe Newton's rings as seen by reflected monochromatic light and account for them.
3. How did Newton "solve" the problem of why part of the light is reflected from glass?

4. What evidence did Thomas Young adduce to support his contention that light was a wave motion?
5. How is the "black spot" in thin films produced and what application is now being made of the phenomenon?
6. What is diffraction and who made the earliest observations of it?
7. State Huygens' principle and give an example.
8. Why do interference bands appear in the overlapping region of light diffracted through two adjacent apertures?
9. Show how double-slit diffraction gives information about the wave-length of light.
10. Describe and account for the diffraction pattern produced by a single slit.
11. What are the relative merits of prisms and gratings as spectrum-producing devices?

Problems on Chapter 33

1. As oil films or soap films grow thin enough to exhibit colors, the first color to appear is blue. Why is this?
2. The film over the left third of the dial of Figure 295 was produced by dipping the glass into a solution of cadmium arachidate 44 times. If the final thickness of the film was .025 of the average wave-length of white light (.000059 cm), how much was deposited at each dipping? The middle section was dipped 26 times. What was its thickness? $3.3 \cdot 10^{-8}$ cm. $8.6 \cdot 10^{-7}$ cm.
3. In observing "Newton's rings," Newton found the radius of the fifth dark ring produced by reflection to be $\frac{611}{6000}$ inch. The rings were produced by laying a lens having a radius of curvature of 91 inches on a piece of plane glass. He deduced that the thickness of the air film at that point was $\frac{5}{88000}$ inch. Check his calculations. What value does this give for the average wave-length of white light? Compare your result with the table on page 368.
4. Sodium light of wave-length .000,058,93 centimeter passes through a grating having 6000 lines per centimeter. Through what angle is the first-order diffracted beam bent? 21°.
5. The headlamps of an automobile are 1 meter apart. The diffraction patterns produced when these lights are observed through a silk umbrella overlap. An observer 100 meters away notes that the fifth-order image of one light falls squarely on the other light. What is the spacing of the threads of the umbrella? 40 threads per cm.
6. Two glass plates in contact at one edge are separated at the other by a piece of tinfoil. When illuminated by light of wave-length λ angstroms, n interference bands are counted across the width of the plates. What is the thickness d of the tinfoil in millimeters?

λ	n	d
6678	30.	0.01002
5893	33.9	0.00999
5780	34.6	0.01
5461	36.6	0.01

7. Parallel light of wave-length λ angstroms, from a point source, passes normally through a narrow slit a millimeters wide and falls normally upon a screen which is d meters distant from the slit. What is the breadth b of the central bright diffraction fringe in millimeters?

	λ	a	d	b
7.	5893	.5	4	9.4
	5893	.4	3	8.8
	5893	.3	2	7.9
	5893	.2	1	5.9

	a	n	α	λ
8.	.5	5	45.91	6700
	.5	5	40.52	5900
	.5	5	39.74	5800
	.5	5	37.55	5500

8. Monochromatic light is placed before the collimator of a spectrometer. A vertical auxiliary slit of width a millimeters is placed before the telescope, and the telescope brought approximately into line with the collimator. The system of diffraction bands then appear as shown in Figure 305. The cross hair is set first on the nth dark fringe to the right of the center and then on the nth dark fringe to the left. The telescope in the meantime turns through an angle of α minutes of arc. What is the wave-length λ of the light in angstrom units?

9. A certain grating is ruled with n lines per centimeter. Parallel light is incident normally upon it. What is the angle α in degrees between the normal to the grating and the line to the violet end (4000 angstroms), and the angle β in degrees between the normal to the grating and the line to the red end (8000 angstroms) of the mth order spectrum? What fraction r of the $(m - 1)$th order spectrum overlaps the mth?

	n	m	α	β	r
9.	2500	5	30°	90°	.8
	2500	4	24°	53°	.68
	2500	3	17°	37°	.5
	2500	2	12°	24°	0.
	2500	1	6°	12°	

	λ	v
10.	6562.14	30
	6560.83	90
	6558.86	180
	6556.24	300

10. The so-called C line of hydrogen has a normal wave-length of 6562.80 angstroms. But the same line emitted by a "nova" (exploding star) shows a wave-length λ. With what speed, v, is the exploded material approaching the earth, in kilometers per second? (Velocity of light = $3 \cdot 10^5$ km/sec.)

11. A double slit is placed over the objective of an astronomical telescope. The two slits are at first placed close together. The telescope is then directed at a double star, and the distance between the slits increased until the interference fringes (at first visible in the telescope) disappear. The distance is then a centimeters. What is the angular separation α of the two stars in seconds of arc? Take 5500 angstroms as the wave-length.

a	α
10	.5672
30	.1891
100	.05672
400	.01418

CHAPTER 34

Spectra

The Birth of Spectroscopy

Newton adapted the word *spectrum* from the Latin to describe an elongated and colored image of the sun which was produced by his prism. His first public use of the word in this connection was in his paper of 1672 before the Royal Society. The term was not particularly descriptive at the time, for the Latin word *spectrum* is not associated in any way with color. It means simply appearance, image, or specter and is derived from the verb *specere*, to look or see. Subsequent use has, however, made the word spectrum one of the richest terms in the vocabulary of physics.

In Newton's time, and indeed for a hundred and thirty years after his first paper on spectra, the only means of identifying portions of a spectrum was by color. The modern association between color and wave-length could not be made, partly because the means of measuring wave-length were not at hand, but more especially because light was generally imagined to consist of flying particles to which no wave-length concept was applicable. But when Sir Thomas Young first took advantage of interference phenomena to secure information about the wave-lengths of light, and found that red light was made up of longer waves than violet, he provided the foundation for the most imposing structure of nineteenth-century experimental science, that of spectroscopy.

In 1802, William Wollaston (1766–1818), saw in a spectrum of the sun something that, by strange chance, nobody had ever seen before. His own account of the discovery is as follows:[1]

> If a beam of day-light [i.e., sunlight] be admitted into a dark room by a crevice $\frac{1}{20}$ of an inch broad, and received by the eye at the distance of 10 or 12 feet through a prism of flint glass, free from veins, held near the eye, the beam is seen to be separated into the four following colours only, red, yellowish-green, blue and violet....
>
> The line *A* that bounds the red side of the spectrum is somewhat confused, which seems in part owing to want of power in the eye to converge red light. The line *B* between red and green, in a certain position of the prism is perfectly distinct; so also are *D* and *E* the two limits of violet. But *C*, the limit of green and blue, is not so clearly marked as the rest; and there are

[1] *Philosophical Transactions*, **92**, 378 (1802).

> also on each side of this limit other distinct dark lines *f* and *g*, either of which in an imperfect experiment might be mistaken for the boundary of these colours....

The chief point of importance in this imperfect account is the discovery that the solar spectrum, as formed by a prism from light which has traversed a slit, is crossed by a series of dark lines (Fig. 312, opposite page 387). These lines have since become known as *Fraunhofer lines*, after a later German observer who discovered them independently and with better comprehension of their nature and significance. Wollaston saw only seven lines. Fraunhofer recorded about six hundred, and later observations with improved instruments have increased the number to more than 10,000.

Fraunhofer Utilizes the Solar Absorption Lines

The lone investigator of the subject between 1814 and 1824 was Joseph Fraunhofer, the same man who made the first measurements of wave-length with a grating. As the optician in a "mathematical institute" in Bavaria, he had occasion to make determinations of indexes of refraction of glass specimens used in the manufacture of achromatic lenses. The indexes were of course different for different colors but they were also different for different parts of the same color as spread out in a spectrum. Fraunhofer's great perplexity was how to specify with sufficient precision just what portion of the spectrum was being used in a given determination of index of refraction. He had accidentally discovered in the spectra of various kinds of flame lamps an extremely narrow isolated section of the yellow part now termed the sodium "line." He found this to be always "exactly in the same place and consequently very useful" in the determination of indexes. Wishing for greater brilliance, it occurred to him to see (45:10).

> whether a similar bright line could be seen in the spectrum of sunlight as in the spectrum of lamplight, and I found, with the telescope, instead of this, an almost countless number of strong and feeble vertical lines, which however, were darker than the other parts of the spectrum, some appearing to be almost perfectly black.

Here was precisely what Fraunhofer needed: an array of sharp landmarks, well distributed throughout the spectrum, which could be used as points of reference when the necessity arose for describing precisely any desired part of the spectrum. These dark lines were, of course, the same that had been seen by Wollaston, but in vastly greater numbers because Fraunhofer's instrumental equipment was much better.

Unlike Wollaston in another respect, Fraunhofer proceeded to explore all the implications of his discovery. His measurement of the wave-lengths of the principal dark lines, now called the Fraunhofer lines, has already been described. Recognizing that the wave-length measurements

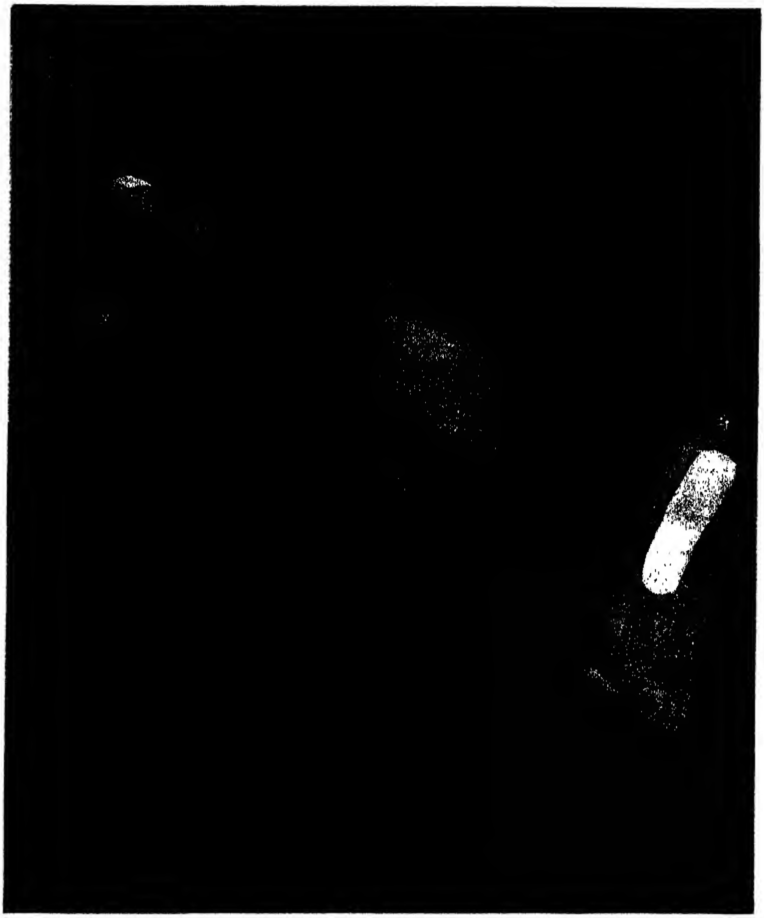


FIG. 265. NEWTON STUDYING THE SOLAR SPECTRUM PRODUCED BY A PRISM (see p. 341)

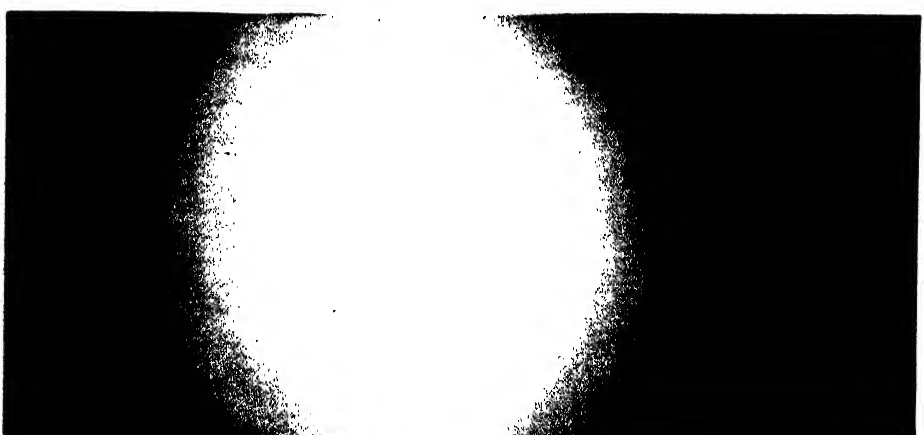


FIG. 266. WHAT NEWTON SAW (see p. 341)

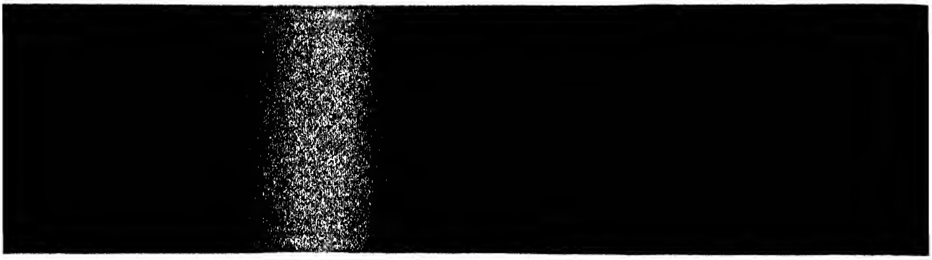


FIG. 311. A CONTINUOUS SPECTRUM, THE SOURCE BEING AN ILLUMINATED SLIT

FIG. 288. THE RAINBOW IS REALLY A "PILE" OF SUPERPOSED COLORED DISKS (see p. 361)

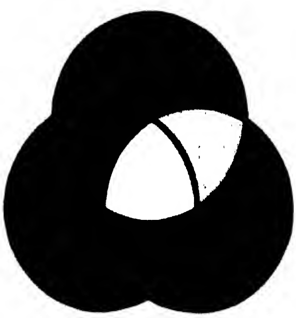


FIG. 280. COMBINATIONS OF PRIMARY HUES (see p. 355)

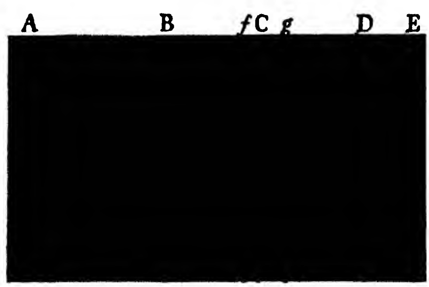


FIG. 312. WHAT WOLLASTON SAW (see p. 386)

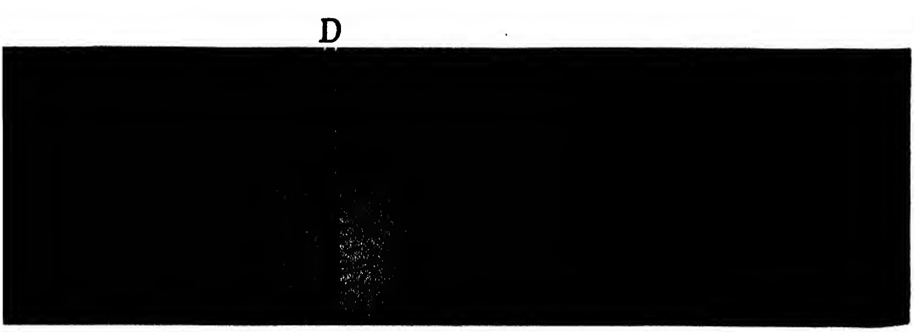


FIG. 313. THE EMISSION AND THE ABSORPTION SPECTRUM OF SODIUM (see p. 388)

facilitated the comparison of the positions of the lines from various sources, he directed his *spectroscope* (the name which came to be applied to his combination of telescope, prism, and angle-measuring device) at the moon and some of the planets and found the lines in the same position. In the stars, however, he found notable differences, though with many lines unchanged. The points of difference made it possible to conclude that the dark lines were not due, at least not exclusively, to the action of the terrestrial atmosphere. It was the later extension of this part of Fraunhofer's work by others which has made it possible to state with certainty the chemical constitution of the heavenly bodies. Not without reason does Fraunhofer's tomb bear the words, "Approximavit sidera."

The Extension of Fraunhofer's Work

Gustave Robert Kirchhoff (1824–87) of Heidelberg resumed where Fraunhofer had left off. Kirchhoff, working in collaboration with R. W. Bunsen (1811–99), had been studying the spectra of flames into which various vaporized substances were introduced. They had observed that these spectra, instead of being unbroken bands of color continuously graded from red to violet, consisted of individual lines scattered across the spectral region, each line possessing the color of that portion of the spectrum in which it was located. These have since received the name *emission line spectra* to distinguish them from the *absorption line spectra* which Wollaston and Fraunhofer had seen in light from the sun. It had been found that a given chemical element always showed the same spectrum, that the spectra of different elements were easily distinguishable from each other, and that within certain limits the spectrum of an element remained unchanged by chemical combination with other elements. Hence Kirchhoff was using spectra as an aid in making chemical analyses of unknown substances.

The simplest of these emission line spectra was that of the metal called sodium. It appeared at first to consist of a single yellow line, but on closer inspection it proved to be double, the two components being very close together. Fraunhofer had seen this line in candle flames, and had noticed that it occupied the same position in the spectrum as one of the most prominent dark lines of the sun, which he had labeled the D line merely for identification. There his observation had stopped. Kirchhoff, besides identifying the line with sodium (in itself a major discovery, though others had anticipated him in this part of the work), proceeded to show the conditions under which this emission line became converted to an absorption line, and then extended the same observation to the spectral lines of other elements. Here is his own description of his procedure (77:355):

> Fraunhofer noticed that in the spectrum of a candle flame two bright lines occur which coincide with the two dark lines D of the solar spectrum. We obtain the same bright lines in greater intensity from a flame in which common salt is introduced. I arranged a solar spectrum and allowed the sun's

rays, before they fell on the slit, to pass through a flame heavily charged with salt. When the sunlight was sufficiently weakened there appeared, in place of the two dark *D* lines, two bright lines; if its intensity, however, exceeded a certain limit the two dark *D* lines showed much more plainly than when the flame charged with salt was not present. (Fig. 313, opp. page 387.)

The Identity of Solar and Terrestrial Lines

To verify this indication that absorption lines were produced by sending light through a flame which was the source of corresponding emission lines, Kirchhoff repeated the experiment, using not the sun, as before, but a laboratory source of white light which would normally produce a continuous spectrum without the dark lines characteristic of the solar spectrum. He thus describes the result (77:355 ff.):

> If an alcohol flame in which salt is introduced is placed between the [white light] and the slit, then in place of the bright lines two dark lines appear, remarkably sharp and fine, which in every respect correspond with the *D* lines of the solar spectrum. Thus the *D* lines of the solar spectrum have been artificially produced in a spectrum in which they do not naturally occur.

Kirchhoff then extended his observations to the spectra of other chemical elements and drew his conclusions as follows:

> If we introduce lithium chloride into the flame of a Bunsen burner, its spectrum shows a very bright, sharply defined line which lies between the Fraunhofer lines *B* and *C*. If we allow the sun's rays of moderate intensity to pass through the flame and fall on the slit, we shall see in the place indicated the lines bright on a darker ground; when the sunlight is stronger there appears at that place a dark line which has exactly the same character as the Fraunhofer lines. If we remove the flame the line disappears completely, so far as I can see.
>
> I conclude from these observations that a colored flame in whose spectrum bright sharp lines occur so weakens rays of the color of these lines, if they [the rays] pass through it, that dark lines appear in place of the bright ones, whenever a source of light of sufficient intensity, in whose spectrum these [dark] lines are otherwise absent, is brought behind the flame. I conclude further that the dark lines of the solar spectrum, which are not produced by the earth's atmosphere, occur because of the presence of those elements in the glowing atmosphere of the sun which would produce in the spectrum of a flame bright lines in the same position. We may assume that the bright lines corresponding with the *D* lines in the spectrum of a flame always arise from the presence of sodium; the dark *D* lines in the solar spectrum permit us to conclude that sodium is present in the sun's atmosphere. Brewster has found in the spectrum of a flame charged with saltpeter bright lines in the position of the Fraunhofer lines *A*, *a*, *B*; these lines indicate that potassium is present in the sun's atmosphere. From my observations, according to which there is no dark line in the solar spectrum coinciding with the red line of lithium, it seems probable that lithium either

> is not present in the sun's atmosphere or is there in relatively small quantity.
>
> The investigation of the spectra of colored flames has thus acquired a new and great importance.... It gives an unexpected interpretation of the Fraunhofer Lines and allows us to draw conclusions from them about the composition of the sun's atmosphere and perhaps also of that of the brighter fixed stars.

It would be hard to exaggerate the significance of these observations and conclusions. August Comte [1] (1798–1857), the famous social philosopher, had remarked only a few years before this that man could never know anything about the chemical composition of the heavenly bodies. This fatalistic pronouncement was confuted almost within his own lifetime. Subsequent progress in spectrum analysis proceeded so rapidly that within nine years after Kirchhoff had initiated it, a previously unknown chemical element, helium, was identified in the sun more than a quarter of a century before its existence on the earth was discovered.

The Doppler Effect in Light

But the spectroscope was to yield other types of information about the stars than merely their chemical constitution. For many years astronomers had desired to secure information about the motions of the stars. The components perpendicular to the line of sight had yielded to direct measurement, but no way was known for determining the components toward or away from the earth ("radial motion"). But the spectroscope opened the way to the measurement of this radial motion. The principle was one that had been applied already to sound, the so-called Doppler effect (page 253). The fact that the apparent pitch of an approaching sounding body is higher than its actual pitch means that the sound waves which it is emitting are crowded up in front of it and that hence the wave-length of the sound is less than normal. For the contrary reason, the wave-length of the sound sent out behind a moving body is greater than normal. The same is true for light, as was pointed out by Doppler himself in 1842 (58:140).

But two obstacles stood in the way of using the information thus rendered possible. One was the inaccuracy of knowledge of the velocity of light and the other the lack of the necessary precision in astronomical instruments. But by 1868 both obstacles had been surmounted and Sir William Huggins (1824–1910) in that year made the first successful measurement of the radial velocity of a star. Since then many such measurements have been made. The velocities of approach or recession of most stars are less than 40 miles per second, though a few exceed this speed (Fig. 314).

One of the interesting applications of the Doppler principle is in the observation of double stars, pairs which are rotating about each other.

[1] J. A. Thomson, *The System of Animate Nature* (New York: Henry Holt, 1920), *1*:15.

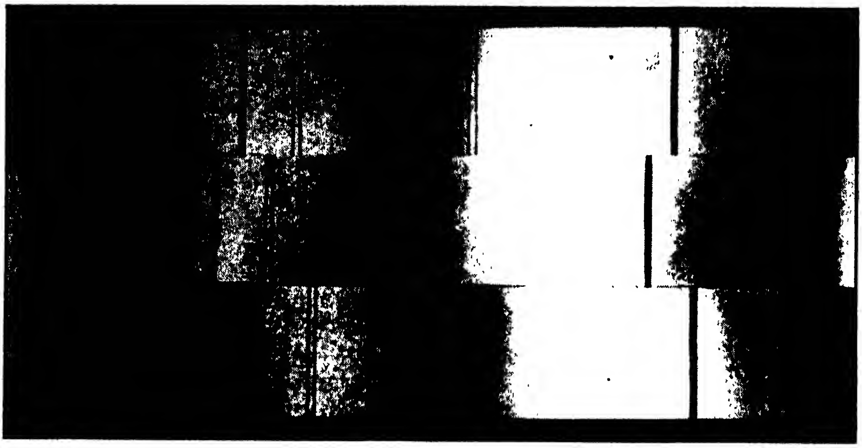


FIG. 314. THE SPEED RECORD OF ONE COMPONENT OF A DOUBLE STAR

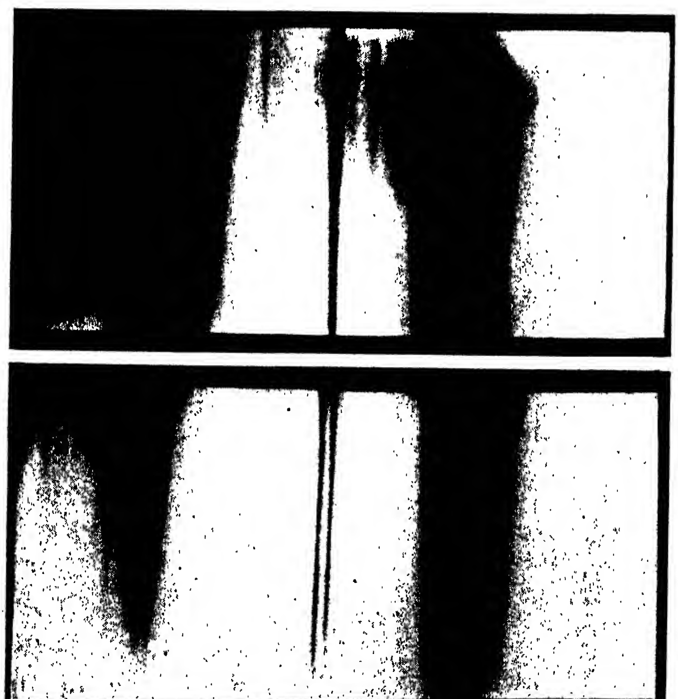


FIG. 315. THE SPEED RECORD OF A DOUBLE STAR
(The change occurred within eighteen hours.)

At certain times in their revolutions, one star is approaching the earth and the other receding from it. Under those circumstances each line in the spectrum of the pair becomes double (Fig. 315), the light from the approaching star decreasing its wave-length slightly, that from the receding star increasing. When the stars have traveled an additional quarter of their orbit, the spectral doublets coincide, and thus the process continues. Some double stars are known to be such only through this spectroscopic evidence. They are too far away to be seen separately by any telescope.

Photography as a New Spectroscopic Tool

During the interval between Fraunhofer's contributions and those of Kirchhoff and Bunsen, something occurred which was destined to have a more profound effect on spectroscopy than all the other developments before or since. In 1826 a French cavalry officer named Niepce made the first photograph on record, using a thin coating of asphaltum on a metal plate. With the aid of an associate named Daguerre he improved the process. After the death of Niepce, Daguerre in 1839 announced a new type of photography which he called the *daguerreotype*, which sprang immediately into common use for portraiture. In those days a visit to the photographer was quite an ordeal. In the first trials, the face of the sitter was dusted with white powder to increase the visibility, and even on a bright day an exposure of from five to seven minutes was required.

Cumbersome as this process was, judged by modern standards, photography was immediately seized upon as a useful way to record spectra. Ever since that time it has been the principal tool of spectroscopy. In 1842 Edmond Becquerel (1820–91) made the first photograph of a complete solar spectrum, a really amazing accomplishment, considering the times, and one which was not surpassed for more than a generation (69:81). Becquerel's photograph is reproduced herewith, the red end being at the bottom.

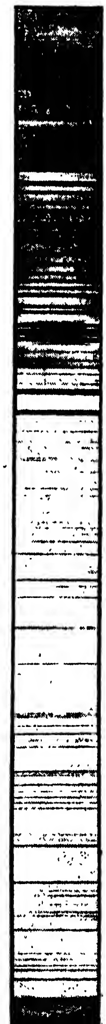


FIG. 316. BECQUEREL'S PHOTOGRAPH OF THE SOLAR SPECTRUM

The Discovery of the Ultra-Violet Region

Becquerel's photograph pointed up several developments which had been taking shape during a number of years preceding. The one which was the most significant will be evident in the fact that, though the red end of the photograph included only that portion of the spectrum which was visible when viewed directly, the violet end extended far beyond the region which was visible. Had Becquerel been looking at the spectrum itself, instead of its photograph, he would have seen nothing above the point *H*. At that point was the violet end of the spectrum, the visible spectrum extending from *H* down through the successive colors ending with red at the extreme bottom. Evidently the camera was capable of catching a section of the spectrum above *H* which was not visible to the eye. Today we call that the *ultra-violet* portion of the spectrum and say that it consists of wave-lengths shorter than 4000 angstrom units.

This was not the first time, however, that evidence had appeared of the existence of invisible radiations outside of the spectrum. As early as 1801 J. W. Ritter[1] (1776–1810) of Jena had reported that beyond the visible violet the spectrum had the power of blackening silver chloride and that, indeed, the blackening was more pronounced there than in the visible part of the spectrum. This observation of its proclivity for stimulating certain chemical reactions led to the ultra-violet's being called for many years the "chemical spectrum."

The identification of the physiological effects of ultra-violet light and the development of its therapeutic use have occurred within the present generation. But its spectroscopic attributes were quite thoroughly explored within a few years of its discovery. The establishment of actual wave-lengths had to await the construction of Angstrom's wave-length tables some thirty years later. But by 1852 Sir G. G. Stokes (1819–1903) of Cambridge University had found that quartz transmitted ultra-violet much more readily than glass. With a spectroscope consisting of a quartz prism and lens he identified an ultra-violet region of the solar spectrum twice as long as had been seen with glass, and extended the ultra-violet spectrum from an electric spark over a range five or six times as long as the entire visible portion of the spectrum produced by the same prism.

The Wave-Length Range of Ultra-Violet

Stokes used fluorescence as his exploring agent in the ultra-violet. Most of the other experimenters in this field were using photography. The same division of practice exists today in the much shorter wave-length region called X-rays. The fluorescent screen or *fluoroscope* is used quite interchangeably with photography to render X-ray beams visible. Two prominent contemporaries of Stokes who used photography to make the ultra-violet spectrum visible were Professor W. A. Miller (1817–70), also of Cambridge University (69:88), and E. E. N. Mascart (1837–1908).[2] Miller used a quartz prism and Mascart a grating to produce the spectra. The difference was important. Whereas both Stokes and Miller had observed that the ultra-violet portions of their prism spectra were five or six

[1] *Gilbert's Annalen*, *7*, 527 (1801); *12*, 409 (1802).

[2] *Comptes Rendus*, *57*, 789; *58*, 1111 (1863–64).

times as long as the visible portions, Mascart's grating spectra indicated that the ultra-violet portion, even though it included all the parts shown by Stokes and Miller, was nevertheless scarcely even as long as the visible portion. We now know that Mascart's results give the more accurate picture of the actual conditions by reason of the "normal" spectra produced by gratings (page 381). The "normality" of Mascart's results will be evident when it is realized that the visible spectrum extends roughly from 7500 angstrom units down to 4000, whereas that portion of the ultra-violet with which these investigators were concerned began at 4000 a.u. and could scarcely have gone below 2000 a.u., if indeed it went that far.

Thus, during the first half of the nineteenth century there developed a realization that the visible spectrum contained only a portion of the radiations emitted by luminous bodies. Knowledge of the extension of the visible spectrum beyond the violet was greatly facilitated by the development of photography, the effectiveness of which lay in the fact that the maximum sensitivity of photographic emulsions lay in the ultra-violet, to which the eye was not sensitive at all. The short wave-length limit (2000 a.u.), which investigators were unable to pass for more than fifty years, is now known to have been due in part to the inability of the gelatin coating of the photographic plates to transmit wave-lengths shorter than 2000 a.u., and in part to the fact that air itself is afflicted with the same inability. Not until photographic emulsions were made without the use of gelatin ("Schumann plates"), nor until entire spectrum-producing systems were enclosed in evacuated chambers, did it become possible to investigate the wave-length region less than 2000 a.u. This "far ultra-violet" (so called to distinguish it from the "near ultra-violet," 4000–2000 a.u.), or *Schumann region*, was not studied until the beginning of the twentieth century.

The Discovery of the Infra-Red Region

During the same span of years that saw the extension of the visible region at the short wave-length end into the ultra-violet, the other obvious extension at the long wave-length end into the so-called *infra-red* was occurring. It was, in fact, a year before Ritter's first observation of the ultra-violet in 1800 that Sir William Herschel (1738–1822) took thermometer readings at different portions of a prismatic spectrum and observed that the temperature was higher at a certain distance beyond the red end than at any other point. Commenting on this observation he said:[1]

> May not this lead us to surmise that radiant heat consists of particles of light of a certain range of momenta, and which range may extend a little farther, on either side of refrangibility, than that of light?... In this case, radiant heat will at least partly, if not chiefly, consist, if I may be permitted the expression, of invisible light; that is to say, of rays coming from the sun, that have such a momentum as to be unfit for vision.... Hence we

[1] *Philosophical Transactions* (abridged), *18,* 675 (1796–1800).

> may also infer, that the invisible heat of red-hot iron, gradually cooled till it ceases to shine, has (only) the momentum of the invisible rays...; and this will afford an easy solution of reflection of invisible heat by concave mirrors.

Herschel later investigated the reflection and refraction of this "invisible light" and demonstrated both the equality of the angles of incidence and reflection and the applicability of Snel's law.[1] It is worthy of note that Herschel described his observation in terms of the "momenta" of "particles" of light. This was two years prior to Young's pioneer work on interference, by which the wave theory of light was to be placed on its first secure foundation, so it would be expecting too much to have Herschel's statement couched in terms of the wave theory. The observation itself was entirely sound, and indeed required only minor changes in phrasing to lend itself to the wave terminology.

As was the case with Ritter's observation of ultra-violet, Herschel's discovery of the infra-red remained largely undeveloped for a generation. This was in part due to the fact that thermometers were too clumsy and insufficiently sensitive to yield better than rough observations in this field. This handicap was removed in 1830 by the invention of the *thermopile* by Leopoldo Nobili (1784–1835). Collaborating with Nobili in the improvement of the thermopile, Macedonio Melloni (1798–1854) began a brilliant series of researches on the infra-red spectrum, culminating in his final publication in 1850, *La Thermochrôse, ou la coloration calorifique.* As the title indicates, Melloni established the existence of different "colors" of radiant heat. After exhaustive tests he found that rock salt would transmit this radiant heat the most readily, just as a few years later Stokes discovered that quartz was almost the only substance that would transmit ultra-violet. Hence Melloni used rock salt for the lenses and prisms with which he studied his infra-red spectra.

Completing the Conquest of the Infra-Red

Photography had been one of the major factors in the development of ultra-violet, but it was not adapted to the exploration of the infra-red. Largely because of this fact, the final stage in the exploration of infra-red came much later than that of the ultra-violet. Nevertheless, the final stages of the two were strikingly similar. In both cases, the conclusive stage arrived with the substitution of gratings for prisms, thus producing "normal"

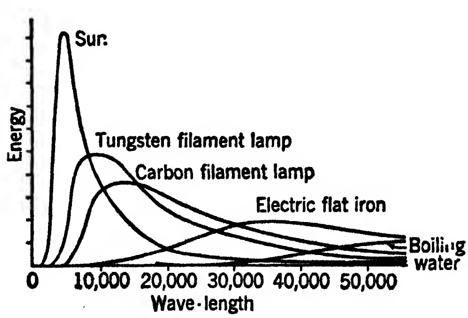


FIG. 317. SPECTRAL DISTRIBUTION OF ENERGY FOR VARIOUS HOT MATERIALS

[1] *Ibid., 18,* 692, 750.

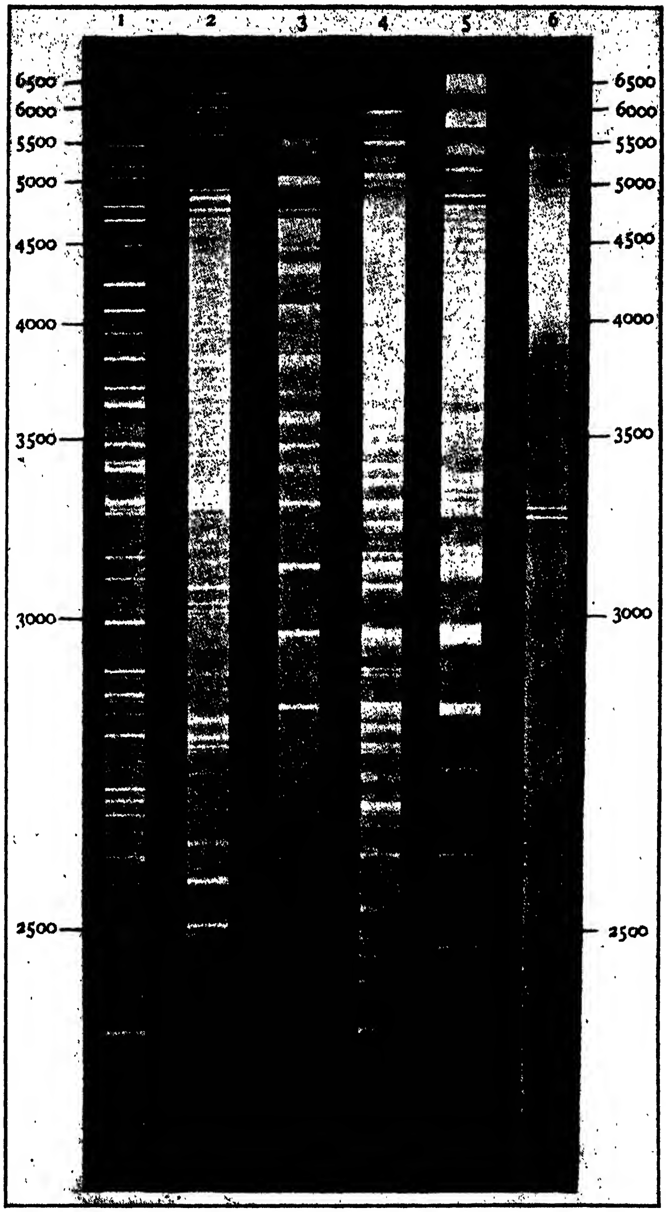


FIG. 318

Spectra shown are: (1) the spectrum of silver produced by an electric arc; (2) that of zinc, by an electric spark; (3) that of carbon monoxide and (4) that of carbon dioxide; (5) a band spectrum of nitrogen, produced by an electric discharge in rarefied gas (e.g., electric signs); (6) the spectrum of copper as it appears in the "green flame" familiar when copper salts are thrown into a fire.

(From *Spectroscopy* by E. C. C. Baly, 1918 edition [Longmans, Green and Company, publishers].)

spectra, making wave-length measurements possible, and eliminating the selective absorption due to passage through the prisms. This step was taken in 1880 by S. P. Langley[1] (1834–1906) of the Smithsonian Institution. Langley had devised a vastly more sensitive temperature indicator than Melloni's thermopile, which he called a *bolometer*. A very fine strip of wire had its resistance increased by exposure at a given point in the spectrum. Measurement of the change in resistance could be made so precisely that temperature changes of one ten-millionth of a centigrade degree were identifiable. The combination of grating and bolometer very soon enabled Langley to write *finis* on the problem of the infra-red. The wave-lengths of the solar spectrum were extended from 7500 a.u. to about 30,000 a.u. (.003 mm) where the solar spectrum appeared suddenly to terminate. Terrestrial sources yielded even longer waves, some of them as great as 500,000 a.u. (.05 mm) as measured by later observers.[2]

Moreover, Langley found that the maximum temperature of the solar spectrum lay at the wave-length corresponding to yellow light, instead of in the red as indicated by Herschel and Melloni. Most important, perhaps, was his discovery that the position of this maximum depended upon the temperature of the source, so that by charting the temperature distribution across the spectrum of a distant source, such as a star, the position of the maximum betrayed the temperature of the source (Fig. 317). Later workers even applied this method to low-temperature sources such as the moon and the planets.

The Origin of Spectra

Thus the study of spectra provided very early several types of information about the astronomical bodies whose distance rendered them otherwise inaccessible. The physical state (whether gaseous, as indicated by emission line spectra, or solid or liquid, as indicated by continuous spectra), the chemical constitution if in the gaseous state, the state of motion toward or away from the observer, and the temperature; all these types of information were adduced from studies of the spectra of distant luminous bodies. Other types of information about astronomical bodies were forthcoming as experimental methods improved, but in the meantime another spectroscopic discovery bore significant fruit of a different kind.

It was realized very early that the spectrum of a gas was not made more complicated by any increase in the quantity of the gas producing the spectrum, nor rendered any more simple by reduction of the quantity. The same array of spectral lines, whether of emission or of absorption, appeared in both cases. A natural conclusion, and one which has since been demonstrated to be correct, was that in such cases each atom or molecule of the gas emitted the same array of spectral lines. The only effect of in-

[1] *Proceedings American Academy, 16*, 342 (1881).

[2] Rubens and Nichols, *Physical Review, 4*, 314 (1897).

creasing the number of spectrum-emitting particles was to make the spectrum more intense.

But if each atom or molecule emits the complete spectrum, a sufficient mechanism must reside in each such particle to produce the complete spectrum. Some spectra are extremely complicated. That of vaporized iron, to take a somewhat extreme example, contains many thousands of lines; hence the atom of iron must be very complicated. Professor H. A. Rowland (1848–1901) of Johns Hopkins University is said to have remarked once that, compared with an atom of iron, a grand piano must be a very simple structure. Notwithstanding the natural implication that a study of spectra should yield information on the structure of the atoms and molecules producing them, spectra withstood all attempts at such interpretation until the twentieth century was well under way. But certain observations paved the way for these twentieth-century discoveries.

Molecular Spectra

One of Wollaston's observations with the prism, presumably at about the time that he saw what later came to be known as the Fraunhofer lines, was the spectrum of the blue portion at the base of the candle flame (30:293). This resembled neither a continuous spectrum nor a line spectrum, but presented a curious fluted appearance which seventy years later received the name *band spectrum*.[1] It is now known that what Wollaston saw was the spectrum of the carbon monoxide produced by incomplete combustion of the material of the candle. A different band spectrum, that of carbon dioxide, is produced when combustion is complete. Figure 318, spectra numbers 3 and 4, shows both these band spectra, photographed more than a century later. It was not until 1882 that it was shown[2] that band spectra originate in gases in the molecular state. The more familiar line spectra were shown to be produced by gases in the atomic state. Thus pure gaseous carbon would give one type of line spectrum, pure oxygen another, but the combination would give a band spectrum of one kind if each molecule consisted of one atom each of carbon and oxygen (carbon monoxide) and of the other kind if of one atom of carbon and two of oxygen (carbon dioxide). The flutings or bands were later found not to be continuous gradations of light, as was at first supposed, but to consist of masses of closely spaced spectral lines (Fig. 319). Studies of the regularities in the sequence of wave-lengths constituting these bands have in recent years yielded an immense amount of information about the structures of the molecules responsible for band spectra. Thus another attribute of spectra has been made to give further information on the condition of the luminous material producing the light (or the non-luminous material absorbing it).

[1] Wüllner, *Poggendorff's Annalen*, *137*, 337 (1879).

[2] Goldstein, *Wiedemann's Annalen*, *15*, 280 (1882).

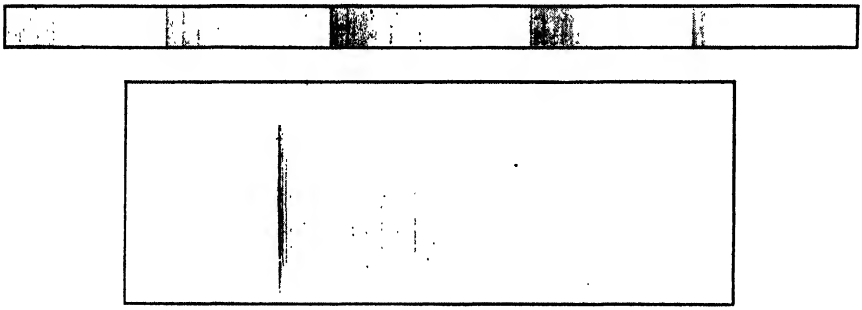


FIG. 319. A BAND SPECTRUM UNDER LOW DISPERSION AND A PORTION OF IT UNDER HIGH DISPERSION
(Spectrogram by O. S. Duffendack.)

FIG. 320. THE FIRST FEW LINES OF THE BALMER SERIES IN THE SPECTRUM OF HYDROGEN
(From *Molecular Spectra*, by Gerhard Herzberg. Prentice-Hall, Inc., publishers.)

Spectral Series

Regularities in the spacing of the spectral lines in line spectra were observed very early. Under certain circumstances, the spectrum of hydrogen had been discovered to consist of a simple sequence of lines, the wave-lengths of the first four being 6562.10, 4860.74, 4340.1, and 4101.2. Figure 320 is a photograph of this series, showing other lines besides those of the above wave-lengths. It is called the Balmer series after J. J. Balmer (1825–98), who first identified the relation between its wave-lengths. In 1885 Balmer published the results of his search for the nature of the regularity of the lines constituting the spectrum. He found that the wave-length sequence could be described by a "common factor" of value 3645.6 multiplied by a series of numbers, one for each line, consisting of $\frac{9}{5}$, $\frac{16}{12}$, $\frac{25}{21}$, and $\frac{36}{32}$ respectively. Stated algebraically

$$\lambda = 3645.6 \times \frac{m^2}{m^2 - 4}$$

in which m assumed the values 3, 4, 5, and 6 successively. The wave-lengths of the lines, computed from this formula, were 6562.08, 4860.8, 4340, and 4101.3. The agreement with the measured values was very impressive. The agreement became still more significant when subsequent discovery of other lines in the series, the existence of which had not been known to Balmer, yielded wave-lengths agreeing with Balmer's formula within a margin which, though not as startlingly small as the first four, was still tolerable. The complete series which Balmer[1] gave in his paper is reproduced herewith:

m	Calculated	Observed
3	6562.08	6562.1
4	4860.8	4860.74
5	4340.	4340.1
6	4101.3	4101.2
7	3969.	3968.1
8	3888.	3887.5
9	3834.3	3834.
10	3796.9	3795.
11	3769.6	3767.5
12	3749.1	3745.5
13	3733.3	3730.
14	3720.9	3717.5
15	3711.	3707.5
16	3702.9	3699.

Balmer's discovery opened a new era in spectrum analysis. His formula served as a model for later spectral formulas and became the foundation of an immense structure on the theory of spectral lines. Most of it has been built during the present century, and cannot be described here. The later portions of it have succeeded definitely in connecting the arrangement of spectral lines with the structures of the atoms emitting the light. Thus the immense body of knowledge of spectra has become available as a tool in the modern attack on the structure of atoms.

Balmer's first paper was confined to the study of spectral lines of hydrogen. He realized the desirability of extending the study to other substances. He also commented on the possibility that other spectral series might be identified by the use of some number other than 4 in the denominator of his formula. Such series were actually found; one corresponding to a value 9, another to a value 1. The first corresponds to a series in the infra-red, the other to a series in the ultra-violet, known respectively as the Paschen and the Lyman series, after the twentieth-century experimenters who established their actual existence.

Spectrum Analysis as an Example of Science in Action

The study of spectra has yielded an incalculable amount of information on regions otherwise inaccessible on account of distance, as with sun and stars, or on account of minuteness and delicacy of structure, as with atoms and molecules. Neither of these two great fields was in prospect when the study of spectra began. If investigators had waited to study spectra until some utility for the study became evident, the development of physical

[1] *Annalen der Physik und Chemie*, **25**, 80 (1885).

science, and with it the development of modern civilization, would have been retarded by at least a century. There have been two eras when cultural civilization stood on a high plane. One was the time of the early Greeks in their day of political independence, the other is the present. Both originated in the impartial and avid pursuit of knowledge. The "practical" man played his part in both instances. He was the skilled artisan in Greek times and the engineer in modern times. Always his contribution, important though it was, has been secondary to that of those who have blazed the trail of new knowledge, without any necessary prevision of where it would lead. It is easy and common to stigmatize the ideal of disinterested search for truth, regardless of its immediate utility, as visionary and futile. In a sense, it is always visionary and it sometimes seems futile. But without it we should still be living in caves. The totalitarian assault on disinterested intellectual endeavor may yet put us back into caves.

Questions for Self-Examination

1. Tell about the first observation (by Wollaston) of solar-spectrum absorption lines and his interpretation of them.
2. Tell about Fraunhofer's rediscovery of solar-spectrum absorption lines and what he did with them.
3. Tell about Kirchoff's extension of previous studies of "Fraunhofer lines."
4. What information may be secured by measuring the shifts and the doublings of spectrum lines from certain stars?
5. Tell the story of the discovery of the ultra-violet portion of the spectrum.
6. Tell the story of the discovery of the infra-red portion of the spectrum.
7. Describe Balmer's discovery and tell what new phase of spectrum interpretation was inaugurated by it.

CHAPTER 35

Polarized Light

Polarization and the Wave Hypothesis

The verb "polarize" is a technical term commonly applied in several unrelated fields. Unfortunately it is not particularly descriptive. Its application to light is no exception. The general connotation of "communicating polarity" (Webster) gives no inkling of the nature of the phenomena in light to which the term is applied. Polarized light, though recently acquiring an increasing utilitarian significance, is far from new. It has for more than three centuries yielded information on the nature of light which could not have been adduced from any other source.

The last two chapters centered in the concept of the wave-nature of light. But though this concept was developed to the point of dealing quite circumstantially with wave-lengths, the question as to the nature of the wave aspects of the disturbance constituting light was not raised. The wave theory of refraction, diffraction, and interference involved no supposition as to how the vibrations were directed in space, nor, more specifically, as to the relation of this direction to the direction of propagation. On these questions, the first and indeed almost the only information available has grown out of the phenomena of polarized light.

When classified according to relation between direction of vibration and direction of propagation, the two principal types of wave are longitudinal and transverse. The most common example of the longitudinal is the sound wave, and a convenient example of the transverse is the wave easily set up in an inextensible string. In the early nineteenth century when the wave theory of light was getting a foothold, the unanimous assumption was that light waves were longitudinal. When it was observed, in ways that will presently be described, that polarization phenomena could not be accounted for in terms of the imagery of longitudinal wave motion, for a time it was assumed that that fact rendered the wave hypothesis untenable. A considerable number of years elapsed before it developed that the very phenomena that were at first supposed to have rung the death knell of the struggling wave theory actually constituted the most incontrovertible evidence of the wave properties of light.

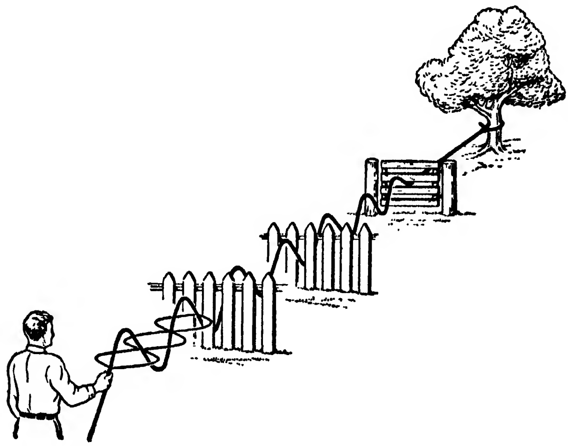


FIG. 321. THE "POLARIZATION" OF WAVES IN A CORD

The Nature of Polarization

The observed phenomena of polarized light changed their rôle from one of a threat to the wave theory to one of fundamental support only when men ceased regarding light as a longitudinal wave and began to picture it as a transverse. The basic concept may perhaps be most easily visualized by considering waves in a cord which is passed through slots, horizontal and vertical, in a series of picket fences (Fig. 321).[1] Let the vibrations impressed at one end be in horizontal and vertical planes and indeed in all intermediate planes, in quick succession. The wave motion thus set up would be termed *unpolarized*. Upon passage through the first fence it would become polarized, with its vibrations in a vertical plane. These vertical vibrations would not be affected by passage through a second fence with its apertures also vertical, but would be eliminated by passage through a third fence with its apertures horizontal.

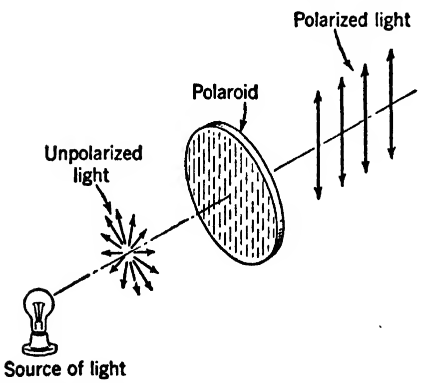


FIG. 322. THE POLARIZATION OF LIGHT

Analogously, the vibrations constituting unpolarized light occur in all possible directions in the plane perpendicular to the line of propagation (Fig. 322). The part of the first picket fence may be played by any one of a variety of

[1] Twelve of the illustrations in this chapter have been furnished through the courtesy of Dr. Martin Grabau of the Polaroid Corporation. He has, in addition, given extended constructive criticism.

polarizing devices which will be described later. After passing through one of these devices, the light exhibits properties which are most aptly described by likening its behavior to that of the portion of the cord beyond the first picket fence. It is then said to be polarized. More specifically it is called *plane polarized,* to distinguish it from the *circularly* and *elliptically* polarized light which will presently be treated. Any device which produces the same effect is termed a *polarizer.*

The unaided eye is unable to detect any difference between polarized and unpolarized light. But if the polarized beam be sent through another device, similar to that which produced the polarization, its state of polarization may be inferred by rotating the *analyzer* (as the second device is termed, from its use in "analyzing" the light). If the *polarizing axis*[1] of the analyzer (comparable to the slots in the fence) be parallel to that of the polarizer, the intensity of the light is unaffected. But if the axis of the analyzer be turned to a position at right angles to that of the polarizer, the light is extinguished. The two cases are diagramed in Figure 323.

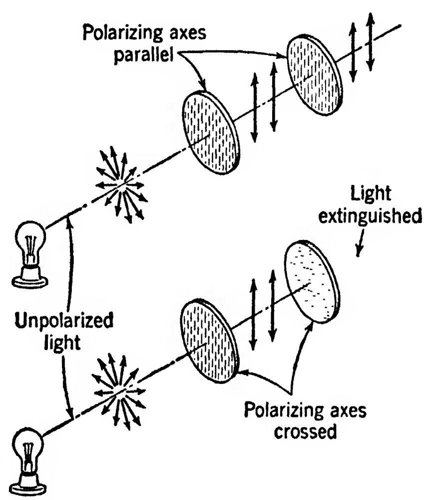


FIG. 323. THE FUNCTION OF THE ANALYZER

It will immediately be evident that a longitudinal wave cannot be polarized. If the cord of Figure 321 had been a spring and the vibrations had consisted of longitudinal displacements, these displacements would have been in no way affected by passage through the successive picket fences, regardless of the orientation of their apertures. Thus the mere fact that light can be extinguished and restored at will merely by turning an analyzer about the plane-polarized ray as an axis of rotation, identifies light unequivocally as a transverse vibration and rules out the possibility of its being longitudinal. This reasoning seems today to be straightforward and elemental, and the conclusion both natural and inescapable. The case was quite otherwise when this field of knowledge first took shape early in the nineteenth century.

Common Polarizing Devices

Until very recently polarizers and analyzers most commonly took the form of what are known as *Nicol prisms,* devised in 1828 by the Scottish physicist W. Nicol (1768–1851). Nicol prisms are still used where the utmost attainable precision is required. But their bulk and expense is causing them to be replaced by more modern polarizing devices. The

[1] The term *axis* used in this connection means a *direction* rather than a particular line.

working principle of the Nicol prism is noted on page 415. The polarizing function may also be performed through reflection at glass surfaces or transmission through glass plates under conditions which will shortly be described. But these, too, are very awkward to use and are correspondingly limited in utility. Recently a group of so-called "Polaroid" products have become available which are proving so convenient and inexpensive as not only to displace other types of polarizers and analyzers, but also to extend greatly the field to which polarized light is applicable. A brief account of the origin of Polaroid products accordingly seems appropriate.

In 1852 a distinguished English physician, W. B. Herapath, discovered a synthetic crystalline material which transmitted polarized light of all colors with astonishingly little absorption. Chemically it came to be known as sulphate of iodo-quinine, but later it was christened "Herapathite." The potential utility of this material was fully recognized, and it was diligently studied all over Europe for a number of years. Unfortunately, however, herapathite crystals were so fragile that they shattered into a useless powder at the slightest touch and hence defied all attempts to make them up into optical units. The discovery was soon all but forgotten.

About seventy-five years later the abandoned problem was taken up and solved by Edwin H. Land, then an undergraduate in Harvard College. Instead of trying to make single crystals large enough for optical use, Land hit upon the idea of embedding myriads of microscopic crystals in a transparent sheet of flexible plastic. Means were devised for aligning these crystals all in one direction. This material is now produced in quantity.

An interesting application of these films is shown in Figure 324. A few club cars of streamlined trains are being equipped with these Polaroid windows. An outer disk is permanently set and an inner may be rotated by a hand-wheel so that passengers may adjust the amount of light admitted. The advantages of such an arrangement over the usual window shade are obvious and striking.

Polarized Light as a Potential Aid in Night Driving

A potential application of polarized light may be found in the field of automotive illumination. Unfortunately it has not yet come into use, though its engineering development is complete. There is always a lag between scientific discovery and its general acceptance for utilitarian purposes. The lag is seldom less than ten years and often runs into centuries. It seems likely that this particular development may come into use within the next ten to fifteen years, though predictions of this kind are hazardous.

If an automobile possessed headlamps which projected polarized light, and an approaching car were equipped with an analyzer with its plane of transmission set crosswise of the plane of vibration of the polarized light

FIG. 324. A POLAROID WINDOW, THE "DENSITY" OF WHICH CAN BE CONTROLLED AT WILL, IS A GREAT IMPROVEMENT OVER A WINDOW SHADE

from the oncoming car, the driver of the second car would not be troubled by glare from the approaching headlights. Though this analyzer extinguishes the light from the oncoming car, as indicated by its shadow across the eyes of the driver, it in no way interferes with his vision of the road. On the contrary, his vision is enormously improved because glare has been eliminated. Figure 325 shows this improvement.

To effect this improvement, the polarizer on the headlamps of the approaching car must be "crossed" with respect to the analyzer of the driver, who would otherwise be subjected to glare. Figure 326 shows a simple way of doing this. If the polarizer and analyzer of each car had their planes of transmission set in the direction upper right to lower left as seen by the driver of that car, they would lie in the direction upper left to lower right as seen by the driver of an approaching car.

The expense of equipping a car in this way would not be at all prohibitive. Securing general adoption of this kind of road illumination is no longer a scientific or an engineering or even an economic problem. Like so many other instances of applied science, it is purely a social problem. The average driver will not assume even the small expense of providing his car with polarized light equipment because it will only be of benefit to the other fellow, not directly to him. The general level of the social conscience

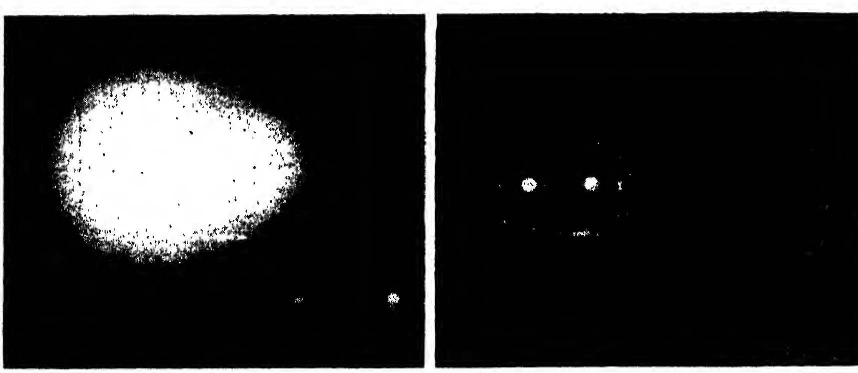


FIG. 325. VISION OF THE ROAD IS GREATLY IMPROVED BY POLARIZED HEADLIGHT ILLUMINATION

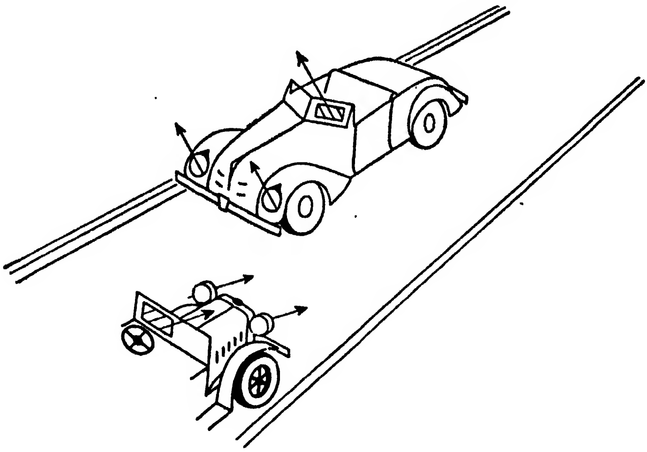


FIG. 326. HOW POLARIZED LIGHT COULD BE APPLIED TO HEADLIGHTS

has not yet reached the point where this motivation is sufficient. We shall have to await legislation, which, in a democracy, is perhaps a rudimentary form of social conscience. Legislation is usually a long process. Perhaps the radically new design that seems to be in prospect for post-war automobiles may incorporate this feature.

Stereoscopic Projection

The stereoscope was described on page 304. It is adaptable only to individual observers. The problem of stereoscopic projection promises to be solved by the use of polarized light, thus extending the effect to audiences. Two stereoscopic scenes are required in this case also, but they are superposed instead of being presented side by side. Both scenes are projected in polarized light, the planes of vibration being perpendicular to each other as indicated in Figure 328. Each member of the audience is provided with a pair of polarizing spectacles, the axis of the lenses being also perpendicular to each other corresponding to the two planes of vibration of the projected scenes. Hence one eye will see one scene and the other eye the other scene. Thus, in Figure 327 a single image will seem to be behind the screen at *B* if the right eye sees only the right image and the left eye sees only the left

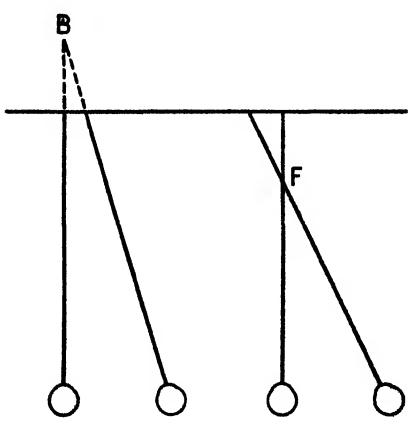


FIG. 327. VARYING DEGREES OF CONVERGENCE OF THE EYES PRODUCE THE ILLUSION OF DEPTH

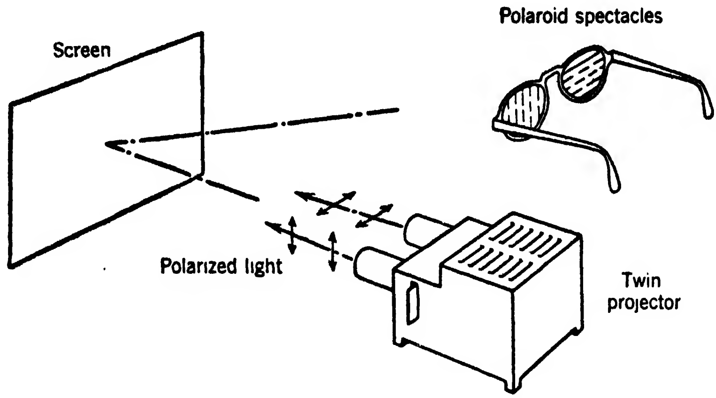


FIG. 328. SCHEMATIC DIAGRAM OF STEREOSCOPIC PROJECTION WITH POLARIZED LIGHT

image of the same object. But it will seem to be in front of the screen at *F* if the right eye sees only the left image and the left eye sees only the right image. In stereoscopic scenes, different portions will be differently displaced, both in direction and in magnitude; thus the illusion of depth or "relief" is produced.

Though there have been other ways of producing the same effect for black-and-white projection, the use of polarized light is the only method of projecting stereoscopic pictures in full color that seems practicable at present. The addition of relief to sound and color, both fairly recent developments, will make the projection of moving pictures almost complete reproductions of the original events.

Rotatory Polarization

One of the later discoveries in this field[1] (1811) was that under some circumstances the plane of vibration of polarized light was rotated by passage through transparent materials. There seem to be three types of such rotation, one observable in such crystalline substances as quartz, another in certain liquids and solutions, and a third in thin crystal sections called "half-wave plates." Discussion of the last type will be reserved until later in the chapter (page 418).

The rotation in quartz occurs only when light traverses the specimen in a particular direction termed the "optic axis" (see page 416). The rotation is considerable, being from 14° to 40° per millimeter of thickness, depending on the wave-length of the light. One of the most interesting features is that quartz having one of its two crystalline forms rotates the plane of vibration to the right (clockwise) while the other crystalline form, a mirror image of the first, rotates it to the left (counter-clockwise).

Certain liquids such as turpentine show the same effect, though in far

[1] 103:436.

less measure. A millimeter layer of turpentine rotates the plane of vibration only between a quarter and three quarters of a degree, depending on the wave-length of the light. Such liquids are said to be "optically active."

Optical activity of liquids and solutions seems to be a fundamentally different phenomenon than that of solids such as quartz, for when quartz is melted it loses its optical activity. A fairly satisfactory theory of optical activity of solids has been built up, but the optical activity of liquids is largely unexplained. It is not really surprising that crystalline solids should possess this property, and that it should be contingent on passage of light in a particular direction through the solid. The regularity of the arrangements of atoms in crystals makes optical activity an effect that is at least not implausible. But the random orientations of molecules in liquids and solutions present quite a different problem. Unless the very passage of the light "lines up" the molecules in a liquid, thus creating a temporary crystalline substance, it is difficult to account for the phenomenon. There seems to be no independent evidence for such an hypothesis.

Double Refraction

Rather paradoxically, the first studies of polarized light did not involve any observation of polarization. They centered in what was called, and is still appropriately termed, *double refraction.* That double refraction involved polarization of the doubly refracted beams was later observed by Huygens (61:92) and by Newton (90:385). The first hint that here was a new field for investigation was dropped by a Danish physician named Erasmus Bartholinus in 1669 (16). Bartholinus' observations may appropriately be described in his own words. He started by describing

> a transparent crystal, recently brought to us from Iceland, which perhaps is one of the greatest wonders that nature has produced.

After describing the shape of this new crystal, Bartholinus continued:

> As my investigation of this crystal proceeded there showed itself a wonderful and extraordinary phenomenon: objects which are looked at through the crystal do not show, as in the case of other transparent bodies, a single refracted image, but they appear double. This discovery and its explanation occupied me for a long time, so that I neglected other things for it; I recognized that I had come upon a fundamental question in refraction.

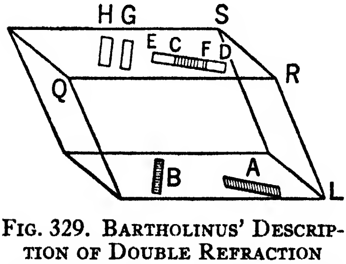


FIG. 329. BARTHOLINUS' DESCRIPTION OF DOUBLE REFRACTION
(From his *Experimenta Crystalli Islandici* of 1669.)

Bartholinus went on to describe the appearance of an elongated rectangular object when placed respectively in the positions B and A (Fig. 329) with respect to the crystal and viewed from above. In the former position, instead of B he saw two images, H and G side by side; in the latter, instead of *A* he saw two overlapping images, *EF* and *CD*. The displacement of the images with respect to each other was always along the bisector of the angle *SRQ*. Subsequently the plane determined by this bisector and the edge *RL* was termed the ***principal section*** of the crystal and came to play an important rôle in the theory of double refraction. It is scarcely necessary to remark that Bartholinus' placement of his refracted images on the upper face of the crystal was nothing more than a device to simplify his diagrams. It was intended merely to show what he saw when he looked down through the crystal at *B* or *A*.

The "Ordinary" and the "Extraordinary" Rays

Bartholinus described another observation in the following terms:

> If we look at objects through [ordinary] transparent media, the image remains fixed and immovable in the same position, however we move the medium. . . . In this case, on the other hand, we can observe that one of the two images is movable.

He described this novel effect with the aid of Figure 330. Double refraction produced two images, *C* and *B* of the fixed point *A*. If now the crystal was rotated about a vertical axis, *C* remained stationary while *B* revolved in a circle around it. He concluded that

> we can distinguish two kinds of refraction, and we designate that one which gives us the fixed image as *ordinary* refraction and the other, which gives the movable image, as *extraordinary* refraction.

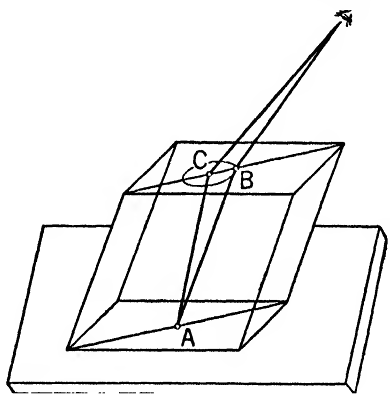


FIG. 330. THE BEHAVIOR OF THE ORDINARY AND EXTRAORDINARY RAYS (From Bartholinus' *Experimenta Crystalli Islandici* of 1669.)

These somewhat naïve terms, ordinary and extraordinary, have persisted to this day.

Here Bartholinus' contribution ended. It will be noted that he did not discover any trace of polarization in the doubly refracted beams. That was first observed by Huygens twenty years later, and developed still further by Newton. Neither did he provide any explanation of double refraction. Though he tried, he had nothing better to propose than a conjecture that it might be occasioned by the disposition of the pores through which the light was transmitted. Here the knowledge of double refraction remained for nearly a hundred and fifty years except for the refinement of Bartholinus' original observations by Huygens and Newton. Huygens especially made some notable advances in comprehension of double refraction on the basis of the wave theory of light. But since the wave theory was not to come into its own until the nineteenth century, Huygens' discoveries received little attention.

Polarization by Reflection

One evening in the year 1808 a whole new chapter was added to the subject of polarized light and incidentally to the comprehension of double refraction. A young French army engineer named Étienne Louis Malus (1775–1812), who had interested himself in optical theory, had his attention drawn to the phenomena of double refraction by the announcement of a prize to be awarded by the Paris Academy for the best mathematical treatment of the subject. While engaged in experimentation with some doubly refracting crystals he idly directed one of them at sunlight reflected into his room from a window of the Luxembourg Palace and was astonished to observe that one of the two images which he expected to see was absent. Polarization by reflection is now well known, but it did not at first occur to Malus to attribute what he saw to the effect of reflection. Instead he speculated on the possibility that passage of the light of the setting sun through long reaches of atmosphere was responsible for this effect. But when after sunset he examined light from tapers reflected from a pane of glass and also from a water surface, and observed the same effect that he had seen in the reflected sunlight, he could not escape associating it with the reflection itself.

Malus' observations can now be formulated in modern terms. Suppose light which is initially unpolarized to be incident on a glass plate. A portion of it will be reflected and that portion is completely plane-polarized for a certain value of the angle of incidence. This *polarizing angle* depends upon the index of refraction of the glass, but for ordinary crown glass it is approximately 57°. The vibrations of the reflected polarized light are perpendicular to the so-called *plane of incidence* (the plane containing the incident and the reflected rays).

According to the electromagnetic theory of light a polarized electromagnetic vibration (light being one example) consists of two parts both

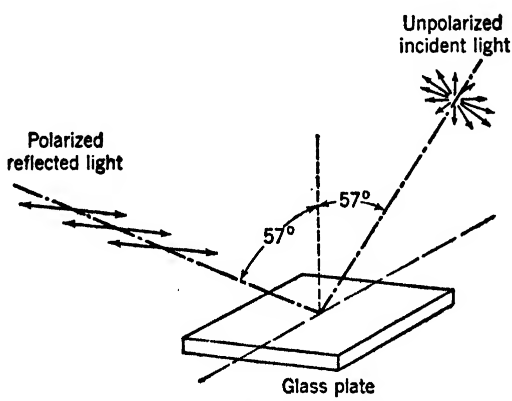


FIG. 331. POLARIZATION BY REFLECTION

transverse to the direction of travel of course, and mutually perpendicular to each other. One of these is a vibratory electrostatic field, the other a magnetic field. There is strong evidence that of these two, only the electric vibration affects a photographic emulsion or the retina of the eye. "Light vibrations" are accordingly to be identified with the electric vibrations of the electromagnetic theory. In the case of light polarized by reflection these lie in a plane which contains the ray and is normal to the plane of incidence. This will be termed the *plane of vibration*.[1] It is indicated by the arrows athwart the reflected ray in Figure 331.

It would be natural for one to look for polarization phenomena in light reflected from an ordinary mirror. They would not be found. More properly speaking, their manifestations would be far from simple and their identification would require optical devices much more pretentious than the mere analyzer that would suffice for light reflected from transparent surfaces such as glass or water. An ordinary mirror is a metallic reflector. The discoveries of Malus do not apply to reflection from metals, as he himself observed. They apply only to reflection from substances that are nonconductors of electricity. This is one of the lines of evidence that light is basically an electromagnetic radiation, though that theme cannot be developed here.

It is also necessary to distinguish between reflection from polished surfaces such as glass and from mat surfaces such as paper, linoleum, wood, asphalt or concrete pavements, etc. This distinction lies at the basis of the reduction of glare by the use of polarizing sun glasses and similar applications described in the following section. A mat material, such as paper, is composed of intertwined fibers which are at least translucent if not transparent. Some of the light reflected from paper comes from the very outer fibers, and the rest from internal surfaces after being shuttled around in the texture of the paper. So we must distinguish between specular (mirror-like) and diffuse reflection. It is the specular reflection from the top surface that produces glare. This, like light reflected from glass, is in a state of partial polarization, and the glare can therefore be reduced by polarizing devices. But the diffusely reflected light, which is the part really useful in making the material visible, is unpolarized, and hence vision by this light is unimpeded by its passage through polarizing spectacles.

Some Utilitarian Aspects of Polarization by Reflection

Light reflected from windows or other glass surfaces frequently obscures by its glare the visibility of objects behind the glass. This glare may be reduced by taking advantage of the state of polarization of the light producing it. The contents of a showcase, or show window, for instance, are frequently obscured by the glare. If the case or window is photographed by

[1] The term *plane of polarization* is sometimes used. This refers to the plane of incidence of light reflected at the polarizing angle. It is at right angles to the plane of vibration. The term came into use before the direction of vibration was known. It will not be used here.

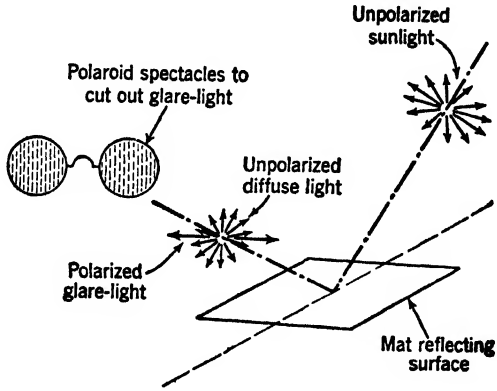


FIG. 332. HOW POLAROID SPECTACLES REDUCE GLARE

a camera with an analyzer in front of the lens, the glare is eliminated. The polarizing axis of the analyzer would have to be set horizontal to cut out the glare light, the plane of vibration of which is in this case vertical. (The plane of incidence is horizontal.) In the same way glare from pavements may be reduced if the driver wears polarizing spectacles with their axes vertical. Figure 332 explains diagrammatically how the polarizing spectacles operate.

Glare may be avoided in another way if the source of light can be polarized. If the plane of vibration of such a source is parallel to the plane of incidence, the light after reflection can contain no horizontal component of vibration perpendicular to the plane of incidence. Figure 333 shows a picture illuminated by unpolarized and polarized light. Reading lamps containing polarizing screens are now available.

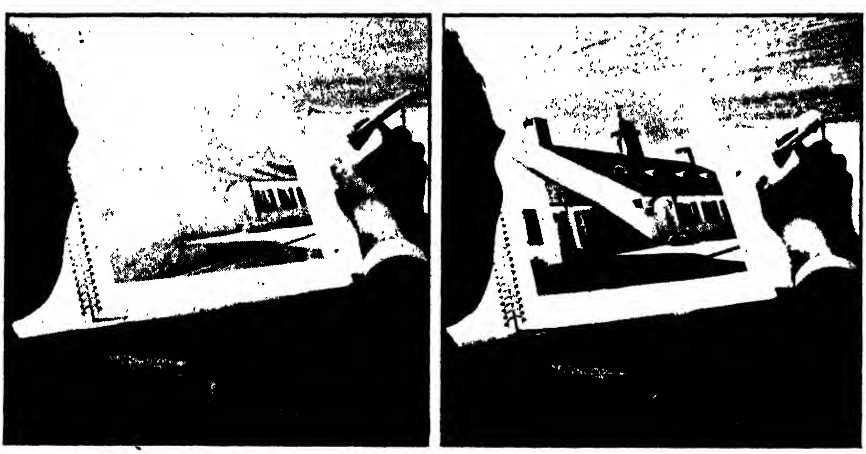


FIG. 333. GLOSSY PRINT ILLUMINATED BY UNPOLARIZED AND POLARIZED LIGHT

Brewster's Law

In 1815 David Brewster (1781–1868), who later became the founder of the famous British Association for the Advancement of Science, discovered that *for light incident at the polarizing angle, the reflected and the refracted rays are at right angles with each other.* This discovery is now known as *Brewster's law.* It furnishes a reason for the otherwise puzzling phenomenon of polarization by reflection. Certain experiments, the nature of which cannot be considered here, indicate that reflection, instead of occurring *at* the surface, occurs after the light has penetrated the medium to a small depth. Consequently it is *after* refraction that reflection really occurs, and the reflected ray splits from the refracted rather than the incident ray. Brewster's law thereupon makes it evident that, if light be a transverse vibration, the reflected ray can consist only of vibrations perpendicular to the plane of incidence.

To show this, consider Figure 334. In this figure, *ab*, *cd*, *ef*, *gh*, represent various transverse directions of the vibrations constituting the incident ray *so*. All these vibrations can be resolved into two component vibrations, one parallel to the plane of incidence and represented by the short cross-lines on the ray *so*, and the other perpendicular to that plane and represented by dots (as though other short cross-lines were seen "end-wise"). Considering the reflected ray which has split off the refracted ray, any vibrations parallel to the plane of incidence would have to have been longitudinal vibrations in the refracted ray from which it separated. But, as has already been seen, if light possessed any longitudinal vibrations, polarization would be impossible. Hence, any ray reflected at the polarizing angle cannot possess any vibrations parallel to the plane of incidence. It is therefore completely plane-polarized, and consists solely of vibrations perpendicular to the plane of incidence.

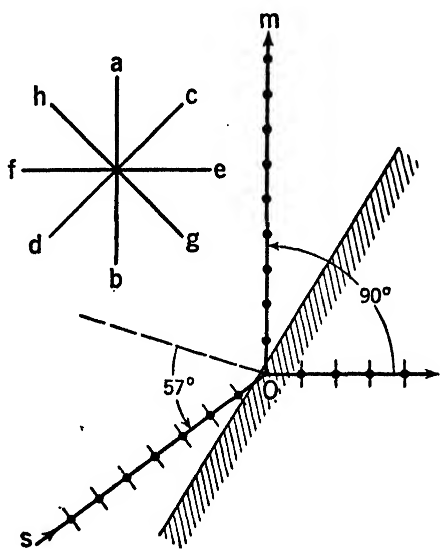


FIG. 334. BREWSTER'S LAW AND POLARIZATION BY REFLECTION

Another Form of Brewster's Law

Brewster's law is frequently stated somewhat differently, namely, that *the tangent of the polarizing angle is numerically equal to the index of refraction of the reflecting substance.* This may not at first sight be obviously the equivalent of the first statement of Brewster's law, but the equivalence will become evident by the examination of Figure 335. By the first form of Brewster's law, *mop* is a right angle. But by construction *qon'* is also a right angle, and the two angles are therefore equal. Since they have *qop* in common, *moq* is equal to *pon'*. That is,

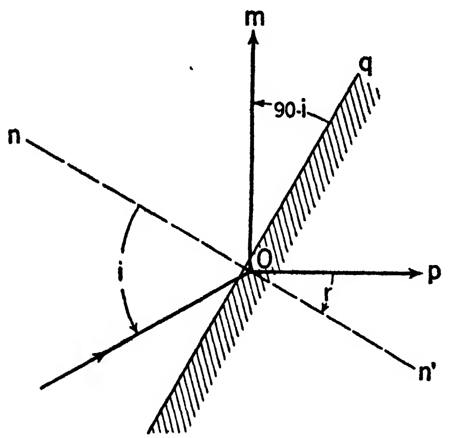


FIG. 335. THE GEOMETRY OF BREWSTER'S LAW

$$90 - i = r \tag{1}$$

hence, $\sin(90 - i) = \sin r$ (2)

or $\cos i = \sin r$

whence, in Snel's law of refraction

$$\frac{\sin i}{\sin r} = \mu \qquad \frac{\sin i}{\cos i} = \mu = \tan i \tag{3}$$

which is the second form of Brewster's law, thus deduced from the first form.

The reason that the polarizing angle of crown glass is about 57° is, therefore, that the refractive index of crown glass is about 1.54, and 1.54 is the tangent of 57°, the circumstance first observed by Malus (page 410). Similarly, as Malus also observed, the polarizing angle for water, being 53°, is to be associated with the fact that the tangent of 53° is 1.33, the refractive index of water.

Since the refractive index of a medium has different values for different wave-lengths, the polarizing angle will, strictly speaking, be different for different wave-lengths and complete polarization cannot be secured for white light by reflection. This becomes evident when reflected white light is observed through an analyzer. Not only can no angle of incidence be found for which complete extinction is securable with the analyzer, but colors appear when the closest approximation to complete extinction is realized. This is a consequence of the complete extinction for one wave-length and the resulting distortion of color values.

Polarization in Double Refraction

Malus had stumbled onto the phenomenon of polarization by reflection while studying double refraction. The side show almost eclipsed the main show, but not quite. He took advantage of reflection to adduce information not otherwise easily obtainable about the polarization of doubly refracted light. He found that the two beams into which light was split by a crystal were both polarized, with the two planes of vibration mutually perpendicular. This is illustrated in Figure 336, in which the plane of vibration of the least refracted ray is seen to be perpendicular to the plane of incidence

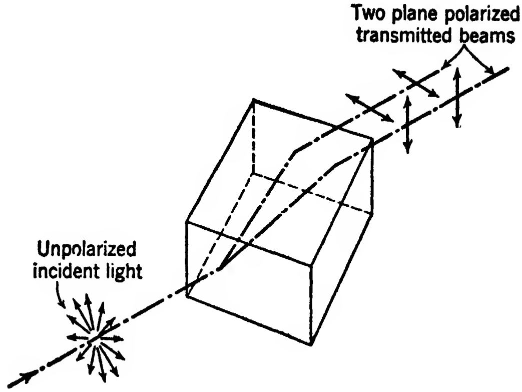


FIG. 336. POLARIZATION ACCOMPANYING DOUBLE REFRACTION

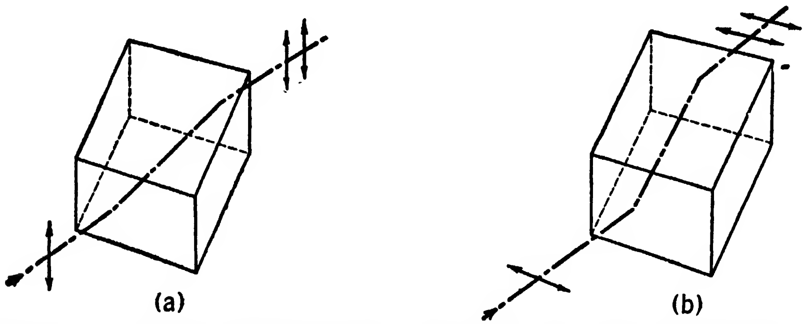
which is arranged to be also the principal section (see page 409) and that of the most refracted ray parallel to the plane of incidence. Here is the way he described his observation (77:318):

> We look through a block of crystal at the image of the flame reflected at the surface of the body or of the liquid. We see in general two images; but by turning the crystal about the visual ray as an axis, we perceive that one of the images diminishes as the other increases in brightness... [For a particular angle of reflection] if we continue to turn the crystal slowly, we shall perceive that one of the two images is extinguished alternately at each quarter of a revolution.

The alternate extinction of the two images which Malus saw is easily accounted for. If the light incident on the crystal, instead of being unpolarized as shown in Figure 336, were polarized (by reflection or otherwise), then the crystal would act as a sort of double analyzer. If the vibrations of the incident light were perpendicular to the principal section, then only the least-refracted ray would penetrate (the crystal being in the position shown in Figure 336). This is indicated in Figure 337 (*a*). If the vibrations were parallel to the principal section, then only the most-refracted ray would penetrate (Fig. 337 (*b*)). In either case, turning the crystal would produce the effect which Malus described, the alternate eclipse of the two images. Thus a crystal acts as a sort of screen or filter to the light, separating light which passes through it into two components. The Nicol prism referred to on page 403 is simply an Iceland spar crystal cut and recemented in such a way as to reflect one of the two rays off to one side. Hence the light which emerges is the other — plane-polarized — ray.

The Velocities of Doubly Refracted Rays

A clue to the reason for this separation of the vibrational components of light by passage through a crystal may be found in the fact that the



(*a*) with the plane of vibration perpendicular to the plane of incidence. (*b*) with the plane of vibration parallel to the plane of incidence.

FIG. 337. DOUBLE REFRACTION WITH THE INCIDENT LIGHT POLARIZED

two beams are refracted by different amounts upon entering the crystal.[1] Refractive effects have already been associated with the change of speed of the light upon entering the refracting material. The greater the deviation of a ray of light upon entering glass, the more pronounced is the diminution of its speed. Thus, in Figure 336, light consisting of vibrations parallel to the principal section of the crystal must have been slowed down more upon entering than light consisting of vibrations perpendicular to the principal section. Otherwise the two beams would not have taken different directions in the crystal.

It was Thomas Young to whom it first occurred to associate the different states of polarization of the two rays produced by double refraction with the presumable difference in the velocities of the rays. This is scarcely surprising since he stood almost alone in support of the wave hypothesis of light (page 368). A few years earlier a man named Ernst Chladni, who had made notable contributions to the science of sound, had observed that sound traveled 25 per cent faster along the fibers of a block of wood than it did at right angles to them. Apparently Young took his cue from this observation, for he quoted it as a precedent for the concept of a connection between speed and direction of vibrations of a wave.[2] With some allowance for figures of speech, crystals may be said to resemble wood in possessing a "grain" and consequently light vibrations parallel to that "grain" might conceivably travel at a different speed than those perpendicular to it. This optical "grain" of a crystal is now termed its optic axis.

When light is passed through a doubly refracting crystal in various directions, it is found that there is always at least one direction ("uniaxial crystals") for which no double refraction occurs. This direction is, by definition, the ***optic axis*** of the crystal. Some crystals have two optic axes and are termed "biaxial." The absence of double refraction in the case of

[1] The fact that one of the rays, the extraordinary, actually does not obey Snel's law makes its own contribution to the case, in addition to the circumstance mentioned here.

[2] Young, in an article on "Chromatics" in the *Encyclopaedia Britannica* for 1817.

light traveling parallel to the axis of a uniaxial crystal is readily accounted for. The vibrations of a transverse wave which is traveling parallel to the axis are naturally all perpendicular to that axis and hence all are propagated with the same speed.

If a crystal is so cut that the surface is parallel to the optic axis, instead of perpendicular, as in the foregoing section, and light is incident normally, the double image effect is again absent, though for a different reason. If the incident light is polarized with the plane of vibration perpendicular to the optic axis, the speed through the crystal will have the value characteristic of that condition. For Iceland spar this is the ordinary ray and is the slower of the two (Fig. 338 (*a*)). If the plane of vibration of the incident polarized light is parallel to the optic axis, the speed through an Iceland spar crystal will be greater, characteristic of the extraordinary ray (Fig. 338 (*b*)). In neither case will the ray be deviated from its original direction upon entering the crystal. If now the incident light is unpolarized, the component vibrations corresponding to the ordinary ray will traverse the crystal more slowly than those corresponding to the extraordinary. Both the preceding effects will be present simultaneously (Fig. 338 (*c*)). Though the extraordinary ray gains on the ordinary, there will be no lateral separation and hence no doubling of the image.

Indexes of Refraction of Crystals

Since index of refraction is defined as the ratio of the velocity of a light wave in air to that in the medium under consideration, it will be evident that a doubly refracting crystal possesses two different indexes of refraction, one for the ordinary ray and the other for the extraordinary. That for the ordinary ray is a constant for a given substance regardless of direction in the crystal. But that for the extraordinary is variable, depending for its value on the direction of the ray in the crystal. This value lies between two limits. The lower limit (for Iceland spar; the upper for quartz and certain other crystals) has the same value as the index for the ordinary ray and is realized when the ray travels parallel to the optic axis. The

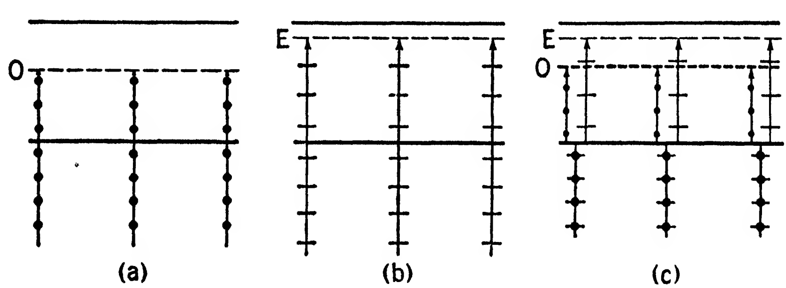


FIG. 338. PASSAGE OF LIGHT THROUGH A CRYSTAL WITH ITS FACE CUT PARALLEL TO THE OPTIC AXIS

(Optic axis in plane of paper.)

upper limit (for Iceland spar; the lower for quartz) is realized when the ray travels perpendicularly to the optic axis, as in Figure 338. For oblique passage through a crystal, the index for the extraordinary ray has some value intermediate between these two extremes. The extremes are called the ***principal indexes of refraction.*** The values of the two principal indexes for several crystalline materials are shown in the accompanying table:

Crystal	Index	
	Ordinary	Extraordinary
Iceland Spar	1.658	1.486
Beryl	1.581	1.575
Sapphire and Ruby	1.769	1.760
Ice	1.309	1.310
Mica (Muscovite)	1.561	1.592
Quartz	1.544	1.553

The table shows why double refraction is so much more prominent in Iceland spar than in other crystals. It also shows the existence of two types of double refraction. One, exemplified by the first three crystals, involves higher indexes for the ordinary ray than for the extraordinary. The other, exemplified by the last three, involves lower indexes for the ordinary ray than for the extraordinary. The geologist terms the former type of crystal negative, the latter positive. As in the case of glass, the values of the indexes of crystals are different for different wave-lengths. The above values are for the yellow light emitted by incandescent sodium, the wave-length conventionally used. Because the indexes are different for different colors, when white light is used double refraction phenomena are often attended by brilliant color effects.

The Half-Wave Plate

The passage of light in a direction perpendicular to the optic axis of a crystal was discussed in the second preceding section. A special case of this is involved in the so-called half-wave plate, a crystal of such a thickness that the speedier of the two doubly refracted but undeviated rays emerges from the crystal one half wave-length ahead of the slower. The half-wave plate possesses the peculiar property of rotating the plane of vibration of incident polarized light through an angle which depends on the orientation of the plate. (Recall the discussion of rotation of plane on page 407.)

To account for the behavior of the half-wave plate, let the incident light be polarized, the plane of vibration making an angle, say somewhere between zero and 90° with the principal section of the crystal. As the ray progresses through the crystal, two component vibrations, one parallel to the principal section and the other perpendicular to it, will travel with different speeds. The slower will emerge a half wave-length behind the faster. A "crest" of the slower wave, which we shall imagine to have been

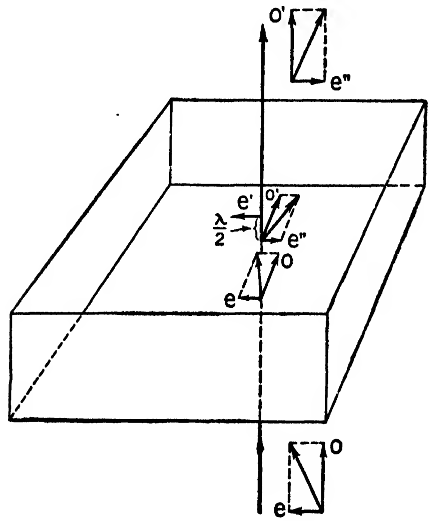


Fig. 339. The Rotation of Plane of Vibration by Passage of Plane-Polarized Light Through a Half-Wave Plate

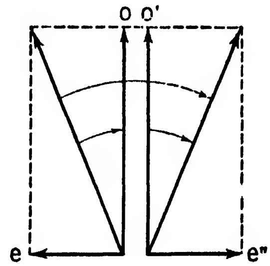


Fig. 340. Rotation of the Plane of Vibration by a Half-Wave Plate

initially associated with a "crest" of the faster, is upon emergence associated with a "trough" of the faster. Figure 339 will show that in consequence the plane of vibration of the emergent light is not the same as the plane of the incident light. The angle between the two is twice the angle between the plane of the incident light and the principal section of the crystal.

In Figure 339 let *e* and *o* represent the components of an instantaneous surge of a light vibration entering a crystal. Upon emergence these two components now represented by e' and o' have become separated because the *e* component traveled more rapidly through the crystal. That portion of the *e* component vibration which will combine with o' will be the portion now abreast of it, namely e'', a half wave-length behind e'. The resultant of o' and e'' will then lie in a different plane from the incident vibration. The angle between the two planes may be seen to be twice the angle made by the plane of vibration of the incident light with the principal section of the crystal, taken to be parallel to *o* (Fig. 340).

The Quarter-Wave Plate and Circularly Polarized Light

If the foregoing plate had been half as thick as it was, the two component vibrations would have emerged a quarter of a wave-length apart instead of half a wave-length. The *e* vibrations (e' and e'' of Figure 341) in the emerging ray, separated now by $\lambda/4$, would then be mutually out of phase by 90° instead of by 180° as in the previous case. The resultant of harmonic motions of the same period but 90° out of phase is an elliptical

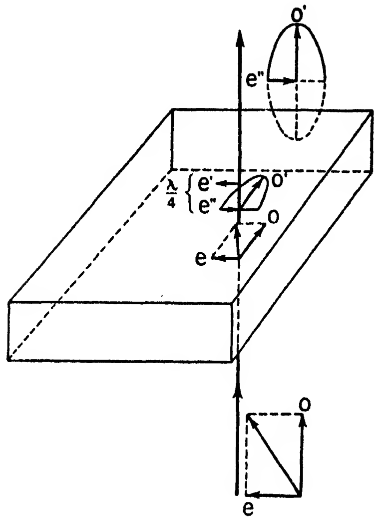


FIG. 341. THE CONVERSION OF PLANE-POLARIZED LIGHT INTO ELLIPTICALLY POLARIZED LIGHT BY PASSAGE THROUGH A QUARTER-WAVE PLATE

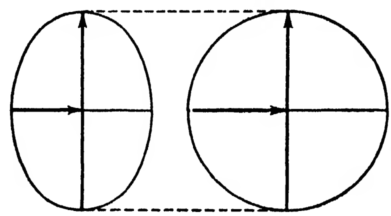


FIG. 342. ELLIPTICAL AND CIRCULAR POLARIZATION

motion. Hence, the plane vibrations of the incident beam would be converted into elliptical vibrations in the emergent beam by passage through a quarter-wave plate.

If the *o* and *e* (Fig. 341) components entering the quarter-wave plate had been equal, the elliptical vibrations constituting the emergent light would have had equal axes. This is the condition for a circle. Equality of the *o* and *e* components could be secured by so orienting the plate that its principal section would make an angle of 45° with the plane of vibration of the incident light. Light so produced is appropriately said to be circularly polarized (Fig. 342). When one views elliptically polarized light through a plane-polarizing analyzer, the intensity changes as the analyzer is rotated, though there is of course no position which produces extinction. But with circularly polarized light no such fluctuation of intensity is produced by rotating the analyzer. Thus an unaided analyzer is incapable of distinguishing between circularly polarized and unpolarized light. One way of identifying circularly polarized light is to add a quarter-wave plate (not to be confused with the one producing the circularly polarized light if it is so produced) to the analyzer. This will convert circularly polarized light into plane-polarized light but will not affect unpolarized light. The conversion of circularly into plane-polarized light is a consequence of the fact that two quarter-wave plates are the equivalent of a half-wave plate.

Colors in Thin Crystals

If a piece of mica be viewed by plane-polarized white light through an analyzer, a color pattern will be observed, the form and colors of which will depend upon the specimen. Rotating the analyzer through a quarter-turn will cause all colors to change to their complements.

To understand this, suppose that one point on the mica is just thick enough to constitute a half-wave plate for the violet portion of the light. Since red light possesses a wave-length approximately twice that of the violet, this same point on the mica will constitute a quarter-wave plate for the red portion of the light which traverses the crystal. Thus the violet portion of the emergent light would be plane-polarized and the red portion elliptically or circularly polarized. The analyzer could be so oriented as to extinguish the violet portion of the light, but not the red portion. The appearance of color is a consequence. Since the colors extinguished for one position of the analyzer are precisely those which have their maximum intensity for the position 90° away, a quarter-turn of the analyzer produces complementary colors. The action of the intervening

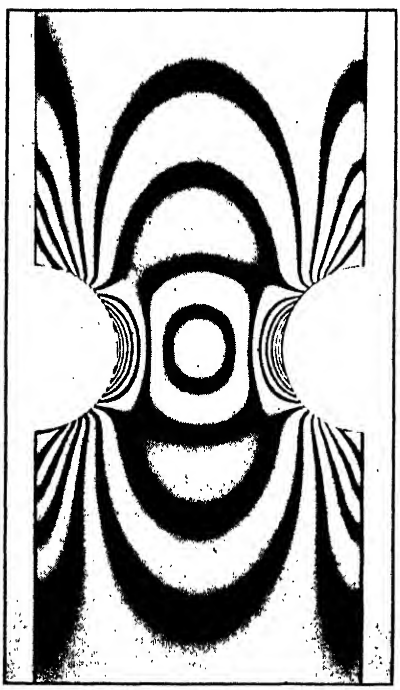


FIG. 343. STRAIN PATTERN IN A BAR WITH SEMICIRCULAR GROOVES, SUBJECTED TO TENSION

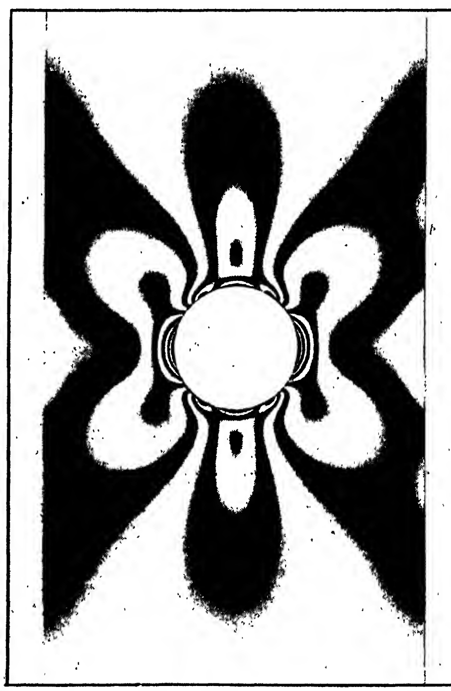


FIG. 344. STRAIN PATTERN IN A BAR WITH CENTRAL CIRCULAR HOLE SUBJECTED TO TENSION

(From *Photoelasticity*, by M. M. Frocht. John Wiley & Sons, Inc., publishers.)

colors of the spectrum complicates the case somewhat, but the general effect is understandable through considering the extremes.

A crystal which is too thin to act like a half-wave plate in violet light cannot show marked color effects, since no color can be entirely extinguished for any position of the analyzer. On the other hand, a crystal so thick as to be, say a $\frac{9}{2}$-wave plate for violet light would be a $\frac{7}{2}$-wave plate for green and a $\frac{5}{2}$-wave plate for red. Since the action of all these is the same as that of a half-wave plate, the effect would be the extinction or the transmission of all three colors simultaneously; in other words, a substantial dimming of color effects. Thicker crystals would accentuate this dimming, ultimately eliminating color effects entirely. Hence color phenomena can be observed only with crystals which are thin enough to be the equivalent of a small number of half-wave plates, but which are thick enough to be the equivalent of at least one half-wave plate. Marvelously beautiful moving color patterns can be observed by viewing in polarized light a mass of crystals forming from a supersaturated solution of a subcooled liquid. The fact that cellophane is doubly refracting is being used to draw attention to kaleidoscopic color displays in advertising, produced by polarization. The first commercial examples of this new display medium were exhibited at the 1939–40 fairs in New York and San Francisco.

Double Refraction by Strain

If a piece of glass or cellophane be put under strain it will exhibit colors when viewed in polarized light. The strain has effected a certain degree of "lining up" of the molecules with the result that the specimen acts somewhat like a crystal. The effect is the same whether the strain is produced mechanically, as by squeezing a specimen in a vise, or thermally as by rapid and unequal cooling of a glass specimen, thus producing internal stresses. The latter is the basis for testing whether optical glass has been properly annealed. If it has, color patterns fail to appear when suspected specimens are examined in polarized light. The former is the basis for the technique known as *photoelasticity*. The distribution of stresses in structural steel, in gear teeth, and similar metal specimens may be determined by constructing transparent models, subjecting them to stresses equivalent to those which the actual specimens will experience when in use, and studying the resulting patterns produced in polarized light. Photoelasticity has developed to such a point that it is now a whole branch of engineering in its own right.

Figures 343 and 344 show in this way the strain patterns due to stresses in transparent specimens subjected to tension. Though the distribution of stresses lends itself to mathematical calculation in some instances, the photoelastic method solves many problems not amenable to mathematical treatment.

Questions for Self-Examination

1. How does polarized light differ from ordinary light?
2. Describe how polarized light promises to eliminate the glare of approaching automobile headlights.
3. Describe the principle of stereoscopic projection with the aid of polarized light.
4. Describe what happens when polarized light travels through "optically active" materials.
5. State the principal facts about double refraction.
6. State the principal facts about polarization by reflection and mention some applications.
7. State Brewster's law in two forms and demonstrate their equivalence.
8. Describe how a half-wave plate rotates the plane of polarization of a polarized beam passing through it.
9. Describe how a quarter-wave plate converts into elliptically polarized light a plane polarized beam perpendicularly incident on it.
10. Why are colors produced by the passage of polarized light through thin layers of doubly refracting material?
11. Describe the phenomenon of photoelasticity.

ELECTRICITY

CHAPTER 36

Electric Currents

Joseph Priestley, in his *History of Electricity*, said (105):

> The electric fluid is no local or occasional agent in the theatre of the world. Late discoveries show that its presence and effects are everywhere.

This was a remarkable statement to have been made as long as nearly two centuries ago. But even Priestley could scarcely have imagined such a complete and literal fulfillment of his description as occurs today in every city. For he was dealing exclusively with what we now call "static" electricity, that is electricity at rest. *Current* electricity, electricity "flowing" or in motion, the kind that supplies virtually all our electric power requirements, had not even been identified at the time that Priestley wrote.

There is somewhat the same distinction between static and current electricity that there is between water lying stagnant in a pool, and water circulating through the pipes of a city water system. The utility of electricity, like that of water, is enormously enhanced by provisions made for its circulation. Those provisions involve, in both cases, the expenditure of energy to create motion and pressure (or their electrical equivalents) and the reappearance of a part of that energy at the places where the water or electricity is utilized.

The energy is commonly supplied to electric distribution systems today in the form of mechanical power, through steam or internal combustion engines driving electrical "generators." But another source of electrical energy is chemical, examples being familiar in automobile and flashlight batteries. It was, indeed, in the operation of an ancestor of the flashlight battery that current electricity was first identified. This was described in 1800 by Alessandro Volta (1745–1827), of the University of Pavia in Italy. In a letter to the president of the Royal Society of London entitled, "On the Electricity excited by the mere contact of conducting substances of different kinds," he said (77:430):

> We set up a row of several cups or bowls (small drinking glasses or goblets are very suitable) half-full of pure water, or better of brine or of lye; and we join them all together in a sort of chain by means of metallic arcs of which one arm *A* which is placed in one of the goblets is of . . . copper and the other *Z*, which is placed in the next goblet is of tin or better of zinc. . . .

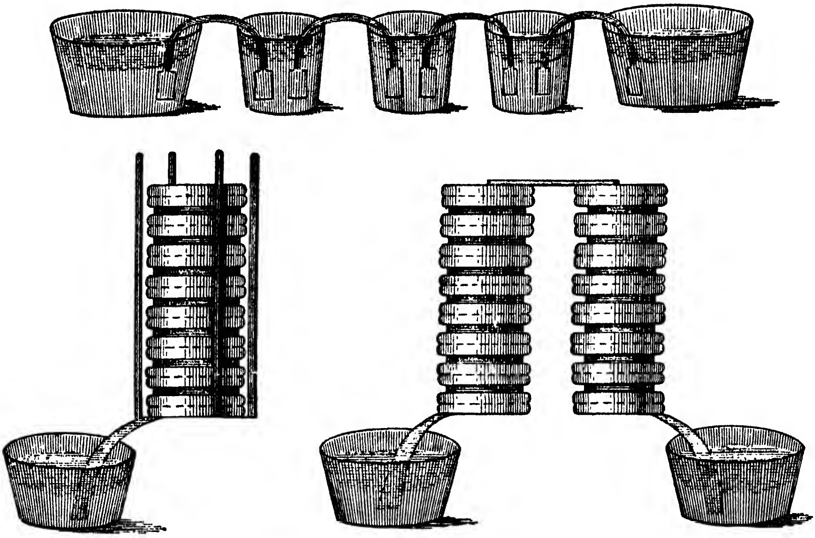


FIG. 345. VOLTA'S ELECTROLYTIC BATTERIES
(*Philosophical Transactions of the Royal Society*, 90, 430, [1800].)

> The two metals of which each arc is composed are soldered together somewhere above the part which is immersed in the liquid.

This is clearly the prototype of the modern *electrolytic battery*, or simply *battery* as it is commonly called today. Volta described also a somewhat more compact form, which has come to be known as the *voltaic pile*, as follows (77:428):

> I provided myself with several dozen small round plates or discs of copper, of brass or better of silver, an inch in diameter more or less (for example, coins) and an equal number of plates of zinc.... I further provided a sufficiently large number of discs of cardboard, of leather, or of some other spongy matter which can take up and retain much water, or the liquid with which they must be soaked if the experiment is to succeed....
>
> I place horizontally on a table or base one of the metallic plates, for example one of the silver ones, and on this first plate I place a second plate of zinc; on this second plate I lay one of the moistened discs; then another plate of silver, followed immediately by one of zinc, on which I place again a moistened disc. I thus continue... to form from several of these steps a column as high as can hold itself up without falling.

The ordinary flashlight battery of today is a fairly direct descendant of the voltaic pile, though a whole century of evolution has intervened between the two. That evolution will not be traced here.[1] It consisted primarily in finding ways to eliminate the corrosion and formation of gases which characterized the action of the early batteries and which reduced

[1] See C. J. Brockman, *The Journal of Chemical Education*, 4, 770 (1927).

FIG. 346. SIR HUMPHRY DAVY (1778–1829)

their effectiveness after relatively short operation. All batteries worked on the principle that the energy of certain chemical reactions became available in electrical form, and that this electrical energy would continue as long as the supply of reagents lasted. The principle of the storage battery was involved in the discovery in 1801 [1] that one such chemical reaction was reversible and that a depleted supply of reagents could, under certain conditions, be restored by sending a current through the exhausted battery in a reverse direction. Storage batteries are sometimes termed *secondary batteries*, in contrast to the ordinary non-reversible type, correspondingly termed *primary batteries*. The conventional type of battery, both primary and secondary, often consists of two or three or more units called *cells*. Thus we may purchase a two-cell or a three-cell flashlight battery, or purchase the cells separately. The usual automobile storage battery consists of three cells, connected in series like Volta's first form. Strictly speaking, the term "battery" applies only to an assembly of two or more cells, but it is often loosely used to describe a single cell as well.

Volta's experiments had grown out of some observations of a contemporary named Luigi Galvani (1737–98). Galvani had attributed the

[1] By Nicholas Gautherot, *Philosophical Magazine*, *24*, 185–86 (1801).

twitching of a dead frog when placed in contact with two dissimilar metals to some agency in the frog's organism which he called "animal electricity" (77:224). Volta demonstrated that the electricity was not at all of "animal" origin, as Galvani assumed, but appeared whenever humid contact occurred between dissimilar metals. Galvani's influence on the science of electricity, in spite of his error, is attested by the contemporary use of the noun "galvanometer" and the verb "galvanize," both words having grown out of the nineteenth-century term "galvanic current," or simply "galvanism," for what we now call electricity.

Electric Heat and Light

In 1801 the heating effect of the electric current was observed, and various metals were fused by its agency. The first commercial application was in 1808 when electric ignition of gunpowder used in blasting was introduced.

In 1802 was exhibited the first electric light produced by current electricity. The production of light by electrostatic devices of one kind and another had for a century been receiving desultory attention with no practical result. But in 1802 in Paris, a piece of charcoal was connected to each end of a voltaic pile of 120 elements and the two brought into contact and separated. There resulted (86:342)

> a brilliant spark of an extreme whiteness that was seen by the entire society.

This accomplishment is ordinarily attributed to Sir Humphry Davy in 1809, at which time he gave the demonstration on a considerably larger scale, probably in ignorance of its having been done before. This was the prototype of the "carbon arc" which was very common from seventy-five to one hundred years later. But at the time that it was first exhibited the consumption of electrical energy was so huge in comparison with the amount normally available that the arc light (as it came to be called from the circular arc which marked the path of the electric discharge) was only a scientific curiosity.

The Beginnings of Telegraphy

In the same year (1802) occurred the first instance of electroplating ("galvanizing"), a process which was ultimately to develop into a major industry. There were also at this time some faint stirrings of the idea of electric telegraphy. Several proposals had been made during the preceding century, which involved the conduction of static charges. None of them could have been made practicable except over distances so short that other methods of signalling were more effective. After two projects, one in 1795 and another in 1802 (86:312, 361), about which not enough is known to form a basis for evaluation, the first major attempt to use current electricity for signalling was made in Munich (110a). It involved the principle of decomposition of water. The two instruments were connected

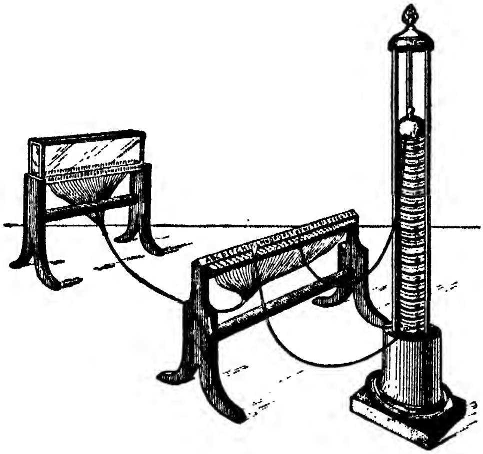


Fig. 347. Soemmerring's Electrochemical Telegraph of 1809

with thirty-five wires, one for each letter and number, the receiving end of each wire dipping into a tube of water. As current was sent over any one of the wires a stream of bubbles at the immersed end signalled the corresponding letter. The inventor succeeded in making his model work through two miles. The original apparatus is in the Deutsches Museum in Munich. In 1811 the number of wires was reduced to two and a code established to transmit the letters. Partly due to the shortcomings inherent in the method and partly due to official inertia, the scheme came to nothing. Another attempt in England was wrecked in 1816 by the official statement (86:439):

> Telegraphs of any kind are now wholly unnecessary and no other than the kind now in use will be adopted.

The "kind now in use" consisted of a row of semaphore towers within sight of each other, which relayed messages from one to the other from end to end of the line.

It was possible for the foregoing preliminary applications of the new agent, electric current, to be made immediately. But real progress in this, as in all other fields of physics, could not be made until units were established and it thereby became possible to pass from the preliminary qualitative development of the subject to the quantitative phase. The bridge from the crude beginnings of the use of electricity in heat, light and communication a hundred and forty years ago to the extensive applications typified in the episode with which the subject of electricity was introduced can, without serious exaggeration, be said to be the establishment of the unit of electric current.

It is now a commonplace that two parallel electric currents exert a force on each other. It is a force of attraction if the currents flow in the same direction and one of repulsion if they flow oppositely. The force is implemented by an intermediary called *magnetism*, but that fact is not of particular importance in the present connection. This force between two parallel currents had been observed long before magnetism was known to be associated in any way with an electric current.[1]

But the real development of this observation into a means of defining a unit of electric current occurred at the hands of André Marie Ampère (1775–1836) in 1820.[2] Ampère was developing some of the implications in a discovery just made by Hans Christian Oersted (1770–1851), of the University of Copenhagen, which will be described presently. Ampère did not content himself with observing merely that there was a force between parallel currents, but he established the fact that its magnitude depended on the strengths of the currents. To do this he had to make his own instruments, for none existed at the time.

Ampère made one wire easily movable on a delicately poised suspension (*CD* of Fig. 348). The other (*AB*) he fixed just below it, connecting each with a voltaic pile. He verified the fact that the two wires were attracted to each other if the currents were in the same direction and were mutually repelled if the currents flowed in opposite directions, and discovered that the force became greater when either of the two currents was increased and less when the distance between them was increased.

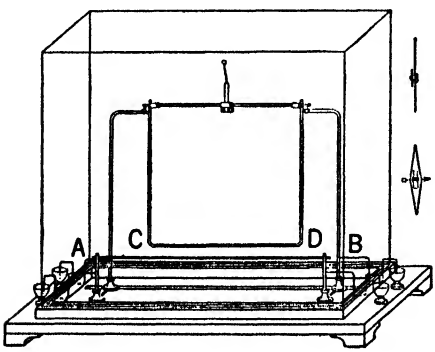


Fig. 348. Ampère's Current Balance from His Own Drawing of 1820

Within a very few weeks a pair of Ampère's contemporaries, Jean Biot (1774–1862), and a young assistant named Félix Savart (1791–1841), made a great addition to Ampère's observation. They secured the information which made it possible to compute the magnitude of the force from knowledge of the currents and distances involved. They discovered[3] that the force per unit length on one wire due to the current in another neighboring, "infinitely" long straight wire was directly proportional to the product of the currents and inversely proportional to the distance between the wires. Stated algebraically, I and i being the two currents in amperes and r the distance apart in meters,

$$f \propto \frac{Ii}{r} \quad \text{or} \quad f = c\frac{Ii}{r} \text{ newtons per meter.} \tag{1}$$

[1] *Annales de Chimie et de Physique*, *39*, 209 (1801).
[2] *Annales de Chimie et de Physique*, (2) *15*, 59 (1820).
[3] *Annales de Chimie*, *15*, 220 (1820).

In the second form of the equation, c is a numerical constant, the value of which could not be determined at the time of Biot and Savart since no unit of current had then been established. Today it is possible to specify its value. Stating the case more accurately, if a value be arbitrarily assigned to c, the unit of current, now called the *ampere*, is thereby determined. This value of c is 2×10^{-7} when r is in meters and f in newtons per meter of one of the wires.

Perhaps the most useful parallel to draw in a preliminary consideration of concepts leading to electrical units is that of a water system. The *rate of flow* of water in a pipe (measured, for example, in gallons per second) is quite completely analogous to *current* in an electrical circuit. It is quite appropriate that the unit of current which, in the electric circuit, might be said to correspond to the unit rate of flow of one gallon per second in the water system, has been named the *ampere.* An electrical instrument which registers amperes is termed an *ammeter*, a contraction of "ampere meter."

Thus equation (1) may be regarded as defining the ampere. For if the two currents be made equal, which could be done even though no unit had been established, and adjusted to such a value that the force between every meter of length of two long wires a meter apart were $2 \cdot 10^{-7}$ newtons, then, substituting in equation (11):

$$2 \cdot 10^{-7} = 2 \cdot 10^{-7} \cdot \frac{Ii}{1} \quad \text{or} \quad I = i = 1. \tag{2}$$

Hence the ampere may be defined as *that constant current which, if maintained in two long, straight, parallel conductors, one meter apart in air, produces a mutual force of 2·10⁻⁷ newtons per meter of length.* In an actual

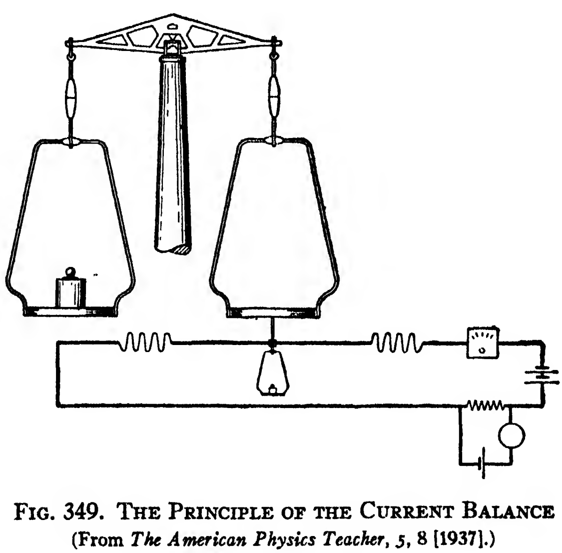


FIG. 349. THE PRINCIPLE OF THE CURRENT BALANCE
(From *The American Physics Teacher*, 5, 8 [1937].)

experimental determination of the magnitude of the ampere, the wires would be placed close together instead of a meter apart, in order that the greater force thereby produced between them could be more readily measured. They would also be coiled into circles, each coil consisting of many turns, to reduce the bulk of the apparatus. These alterations involve minor modifications in the defining equation but the principle of the definition of the ampere is not affected. A pair of coils so mounted that their mutual attractions or repulsions can be counterpoised by weights and therefore measured is termed a *current balance,* and is the device used when there is occasion to establish the absolute value of the ampere directly from mechanical units.

The Relation Between Current and Quantity

If an attendant in a pumping station should observe that his flow-meter indicated, say, one hundred gallons per second, and that ten hours was required, at that rate, to empty the reservoir, he could readily compute that the reservoir had initially held 3,600,000 gallons. The type of calculation involved would be

$$\textit{quantity} = \textit{rate of flow} \times \textit{time}.$$

He could doubtless have determined the volume of water in some other way, such as by measuring the interior of the reservoir, but the former method would be possibly the most convenient one. Analogously, a convenient way to specify *quantity of electricity* is by the relation

$$\textit{quantity} = \textit{current} \times \textit{time}.$$

If the current is measured in amperes and the time in seconds, the quantity is represented in units called *coulombs.* Expressed algebraically

$$Q = It \tag{3}$$

where Q represents quantity in coulombs, I current in amperes, and t time in seconds. Obviously, *a current of one ampere flowing for one second will transfer one coulomb of electricity.*

Electrolysis

Another chapter in the early history of current electricity belongs logically with the work of Volta, though chronologically it was part of the subsequent wave of progress. The discovery of the evolution of hydrogen and oxygen when an electric current was passed through water has already been commented upon, as has the now common phenomenon of electroplating. Certain unexplained aspects of these phenomena came to the attention of Michael Faraday (1791–1867) in 1832 and resulted in his discovery of the *laws of electrolysis* which go by his name. Faraday coined the term *electrolyte* (39:*1*:197) to describe

> bodies decomposed directly by the electric current, their (chemical) elements being set free.

The process of decomposition itself he termed *electrolysis*, and the surfaces at which the products of electrolysis appeared he called *electrodes*.

Faraday's laws of electrolysis are now usually phrased somewhat as follows:

1. The mass of the substance liberated from an electrolyte is proportional to the quantity of electricity driven through the solution.
2. The masses of the substances liberated from an electrolyte by a given quantity of electricity are proportional to their atomic weights.
3. The masses of the substances liberated from an electrolyte by a given quantity of electricity are inversely proportional to their respective valences.

The term *valences* introduced in the third law will presently be discussed.

The point of great present significance in Faraday's laws will become apparent by associating them with the idea that all matter consists of atoms. This idea had become quite explicit in the beginning of the science of chemistry during the generation preceding Faraday's work. If the proportionality set forth in the first law extended down to such small portions of substance as the individual atoms involved in electrolysis, which was a natural assumption, then it followed that *every atom of a given substance had associated with it the same quantity of electricity.* By the second law the same number of atoms of different substances must be associated with a given quantity of electricity, and hence *the same quantity of electricity was associated with every atom participating in a given manifestation of electrolysis* regardless of the substance involved. This statement was subject, however, to modification by the implications of the third law, which will now be developed.

The Atomicity of Electricity

Before proceeding to the third law however, the epochal nature of the first two should be emphasized. Faraday himself said (39:*1*:249):

> It is impossible perhaps, to speak on this point without committing oneself beyond what present facts will sustain; and yet it is equally impossible, and perhaps would be impolitic, not to reason upon the subject. Although we know nothing of what an atom is, yet we cannot resist forming some idea of a small particle, which represents it to the mind; and though we are in equal, if not greater, ignorance of electricity... yet there is an immensity of facts which justify us in believing that the atoms of matter are in some way endowed or associated with electrical powers, to which they owe their most striking qualities, and amongst them their mutual chemical affinity.

As will presently appear (Chapter 48), the idea was born at this time that electricity itself is atomic and that one or more atoms of electricity are associated with every atom of substance, but the idea was not completely developed for the better part of a century. If it were possible in some way to count the number of atoms N in a given mass of substance, and if each

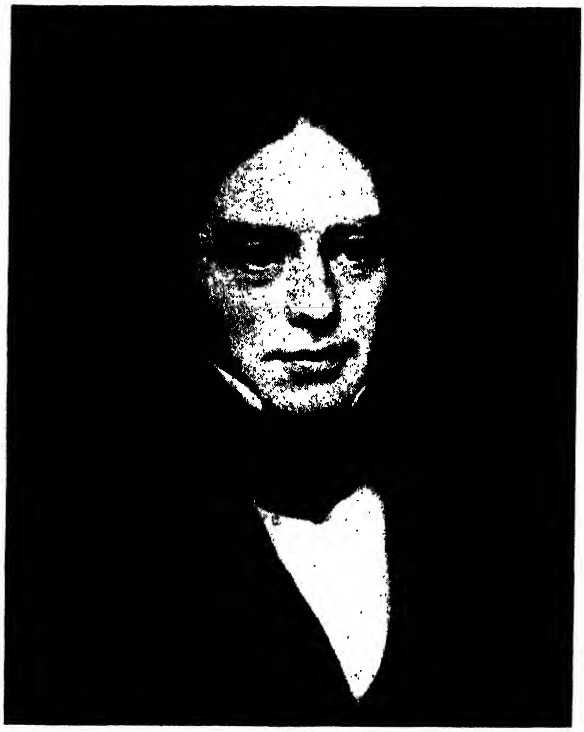


FIG. 350. MICHAEL FARADAY (1791–1867)

atom had a quantity of electricity e associated with it, then the quantity of electricity E necessary to deposit the given mass by electrolysis would be

$$E = Ne. \tag{4}$$

If the mass thus deposited were M grams and each atom weighed m grams, then

$$M = Nm, \tag{5}$$

and therefore

$$\frac{E}{M} = \frac{Ne}{Nm} = \frac{e}{m}. \tag{6}$$

That is to say that even though Faraday could not know either the mass m of an atom or the quantity of electricity e associated with it, his laws made it possible to state what the *ratio* of those two magnitudes was. This was a long step in advance.

The third law made its own considerable contribution to the new picture The possibility of the existence of an atom of electricity was hinted at above. This atom, for the present purely hypothetical, will be called the ***electron.*** The third law allows for the possibility of more than one electron's being associated with an atom. If, for example, two electrons were associated with an atom, then the mass of substance deposited by a given quantity of electricity would be half as great as it would be if there were only one. If there were three electrons on an atom, then only one third as much substance would be deposited, etc. The term ***valence,*** therefore, means a number numerically equal to the number of electrons associated with each atom deposited out of an electrolyte. One would have to consider the possibility that in a given instance of electrolysis, atoms of all kinds of valence might have to be taken into account, except that experience has demonstrated conclusively that this is not the case. In a given instance, all the atoms collecting at one electrode will have the same valence.

Faraday tied the eighteenth-century discoveries of Volta in with nineteenth-, even twentieth-century physics. Noting that the solutions which Volta used for his batteries were all electrolytes and that electrolytic action was necessarily a part of the operation of these batteries, he went on to point out a close association between these facts of electrolysis and some of the basic phenomena of chemical transformation. On the basis of this association he said (39:*1*:248):

> It touches by its facts more directly and closely than any former fact, or set of facts, have done, upon the beautiful idea that ordinary chemical affinity is a mere consequence of the electrical attractions of the particles of different kinds of matter.... I have such conviction that the power which governs electro-decomposition and ordinary chemical attraction is the same.

Thus was laid the groundwork for the interplay between physics and chemistry which a century later was so to vitalize the relations between the two as to make of them virtually a single science.

Questions for Self-Examination

1. Tell about Volta's invention of the electrolytic battery.
2. What was the point at issue in the controversy between Volta and Galvani?
3. Mention the first significant applications of current electricity to the production of heat and light.
4. How was electrolysis utilized in one of the earliest attempts at telegraphy?
5. Define the ampere and the coulomb.
6. What is a current balance?
7. State Faraday's laws of electrolysis. How do they indicate the existence of "atoms" of electricity?

Problems on Chapter 36

1. What is the force per meter between the two conductors of the extension cord to a 100-watt lamp if the conductors are .25 centimeters apart? Is it one of repulsion or attraction? .000,05 newton.
2. Two straight parallel wires each carrying I amperes of current are separated a distance s millimeters. Find the force f per meter length between them in newtons.

I	s	f
10	1	.02
20	1	.08
20	2	.04
100	8	.25

3. The ratio E/M of the charge on the monatomic hydrogen ion to the weight of the ion has been measured as $9.21 \cdot 10^5$ coulombs per gram. Assuming that the charge is due to the excess or deficiency of one electron in the atom, compute the mass in grams of the hydrogen atom. Take the charge of an electron as $1.591 \cdot 10^{-19}$ coulomb. $1.656 \cdot 10^{-24}$
4. From the electrochemical equivalent of univalent copper (.000659 gms/coulomb), find the mass in grams of the copper atom. $1.048 \cdot 10^{-22}$
5. In an electron stream, the measured value e/m of the charge of each electron to its mass is $1.77 \cdot 10^6$ coulombs/gm. What is the mass of the electron in grams? $8.991 \cdot 10^{-28}$

CHAPTER 37

Interaction Between Currents and Magnets

Oersted's Observation

A very large part of the applications of electricity depends upon the interaction between currents and magnets. That there should be some such interaction had been strongly suspected for nearly a century before its discovery actually occurred. Beginning with records as early as 1735 of some magnetizing effects of lightning strokes,[1] a long search, at first casual but later systematic,[2] was made to find some relationship between electricity and magnetism. It was not until 1819 that the long-sought discovery was made by Hans Christian Oersted (1770–1851). The story of the discovery was told as follows by one of Oersted's pupils (62:*2*:395) in the quaint phraseology of one not entirely at home with the English language:

> Once after the end of his lecture, as he had used a strong galvanic battery in other experiments, he said; "Let us now once, as the battery is in activity, try to place the wire parallel to the [suspended magnetic] needle." As this was made he was quite struck with perplexity to see the needle making a great oscillation.... Then he said; "Let us now invert the direction of the current"; and the needle deviated in the contrary direction. Thus the discovery was made, and it has been said, not without reason, that he tumbled over it by accident. He had not before any more idea than any other person that the force should be transversal.

Besides Oersted's observation of a reaction between a magnet and an electric current, which was the principal discovery and a momentous one, he observed a peculiarity in the nature of that reaction to which the above account scarcely does justice: that the magnet, instead of being attracted or repelled as would have seemed more natural, set itself perpendicularly across the wire which carried the current. This was regarded as very mysterious and did, in fact, lead to some very far-reaching conclusions.

Oersted's observation was interesting enough, but one is scarcely prepared for the furor which it aroused, not merely among men of science but even among those whose knowledge of science was confined to hearsay. Figure 352 shows a device used by an itinerant lecturer of the time, William

[1] *Philosophical Transactions*, *39*, 74 (1735); *41*, 614 (1740). See also 117a:10.

[2] *The Electrical Engineer*, *13*, 27 (1892). See also 116a:54.

FIG. 351. OERSTED'S EXPERIMENT
(From the Oersted medal of the American Association of Physics Teachers. Courtesy of J. Rud Nielsen.)

Sturgeon (about whom we shall soon hear more), in his demonstrations of Oersted's discovery before popular audiences. It has its counterpart in almost every physics laboratory today.

Extensions of Oersted's Discovery

Seldom, if ever in the history of science, has the publication of a discovery precipitated such a landslide. Oersted's paper, published in the

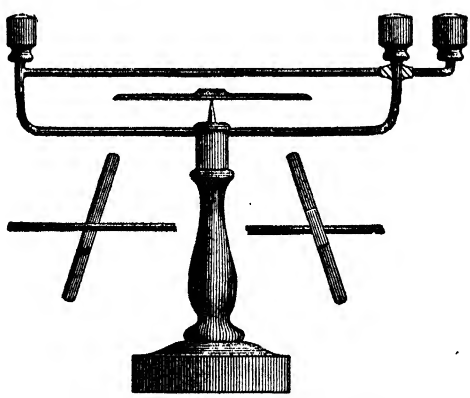


FIG. 352. STURGEON'S OERSTED DEMONSTRATION
(From *Transactions of the Royal Society of Arts*, *43*, plate 3 [1825].)

FIG. 353. ANDRÉ MARIE AMPÈRE (1775–1836)

form of a pamphlet dated July 21, 1820, was reported to the French Academy of Sciences on the following September 11, by D. F. J. Arago (1786–1853). The result was like a starting gun at a hundred-yard dash. More than a dozen capable experimenters immediately set to work on one aspect or another of the field thus opened up. Of these the most energetic proved to be Ampère. Nearly all the weekly meetings of the Academy for the next four months were devoted to the development of the implications of Oersted's discovery, and Ampère was the principal contributor.

Ampère's first paper was given on September 18, just one week after Oersted's observation had been reported to the Academy. In it he refined and generalized Oersted's results. Oersted had stated that a pivoted magnetic needle would set itself

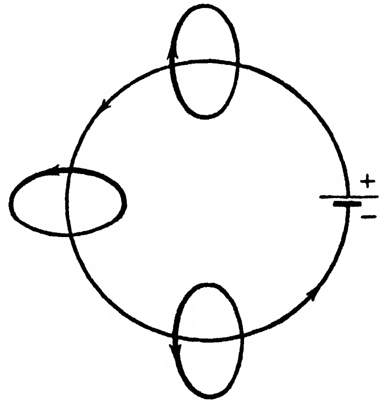


FIG. 354. CIRCULAR MAGNETIC FIELD AROUND A CURRENT

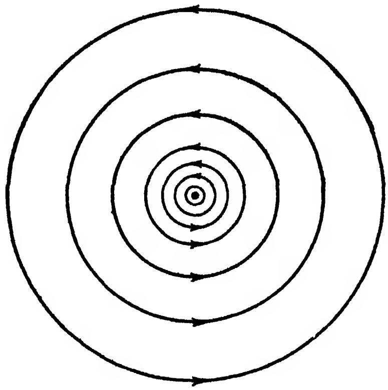


FIG. 355. THE RELATIVE DIRECTION OF CURRENT AND RESULTING MAGNETIC FIELD (CURRENT OUT OF THE PAGE)

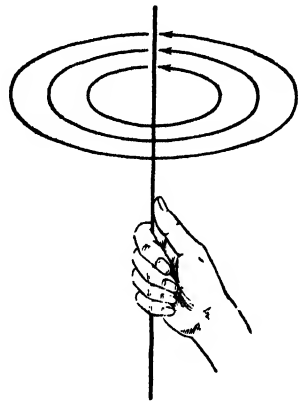


FIG. 356. RIGHT-HAND SCREW RULE

perpendicularly to a neighboring wire carrying a current, and that its deflection would take place one way for one direction of the current and the opposite way for the other direction. Ampère clarified this by formulating in a graphic way a relation between the direction of the current and that of the motion of the needle. His original imagery has been modified into a more useful form, customarily stated as the *right-hand screw rule*:

> *Encircle the wire with the fingers of the right hand, thumb extended in the direction of flow of the current. The fingers then point in the direction of deflection of a north pole.*

If, as is often the case, it is the deflection of a needle that is observed, the direction of the current being unknown, the same rule will apply, conversely phrased.

Thus it appeared that an array of tiny compass needles, free to move, would form a pattern of concentric circles around a current. Arago, the friendly competitor of Ampère who was quoted above, demonstrated this very nicely by dipping a wire carrying a current into a pile of iron filings, and showing that while the current flowed, the filings adhered in a tuft, with their long dimensions tending to be perpendicular to the wire. The circumstance is described today by saying that there is a circular magnetic field around a long straight wire carrying a current. It can be made evident with the aid of iron filings, the direction of the lines showing the direction of the field, and their density (extent of their crowding) showing the strength of the field.

This circularity of the magnetic field around a current was very neatly demonstrated by Faraday within a few months of Oersted's discovery. Faraday showed that, when the proper experimental conditions were observed, a constant and rapid rotation of a magnet about a wire would occur. He also demonstrated the converse, the rotation of a wire about a magnet (Fig. 357). Such an effect, if it had not been conclusively demon-

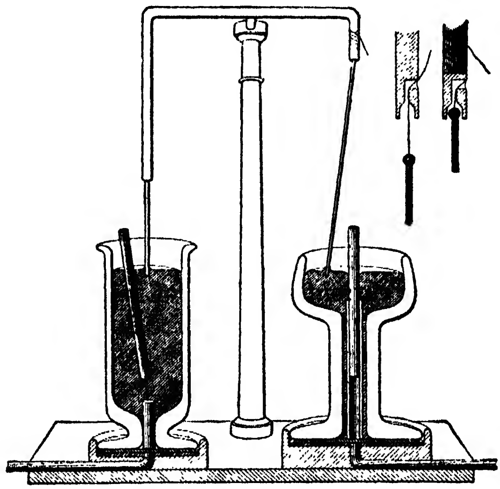


FIG. 357. FARADAY'S ROTATING MAGNET AND WIRE
(From the *Quarterly Journal of Science*, *12*, plate 7 [1822].)

strated, would probably have been termed altogether incredible. But it was only one of the implications of Oersted's own observation of the transversality of the force between a magnet and a current.

The Direction of Flow of a Current

Nearly a century earlier, a Frenchman named Charles Dufay (1698–1739) had observed that there appeared to be two kinds of electricity which he called *vitreous* and *resinous* respectively. The vitreous he so named because it most commonly occurred on glass (Latin *vitrus*) and the resinous because it was the kind observed on amber and similar resinous materials. His nomenclature was soon changed by Benjamin Franklin to *positive* and *negative* respectively, and by those names they are known to this day. Dufay continued (77:399):

> The characteristick of these two electricities is that a body of the *vitreous electricity*, for example, repels all such as are of the same electricity; and on the contrary attracts all those of the *resinous electricity*.... Amber on the contrary will attract electrick glass and other substances of the same class and will repel gum-lac, copal, silk, etc.... From this principle one may deduce the explanation of a great number of other *phaenomena*. And it is probable, that this truth will lead us to the further discovery of many other things.

Dufay's expectation was fully realized. Subsequent developments have verified his idea of the existence of two distinct varieties of electricity, and the concept has occupied a central rôle in "the further discovery of many other things." But his implication that both kinds of electricity moved with equal facility through conductors has not been borne out. It is apparent now that, though both kinds are ubiquitous and exist normally in equal and therefore mutually neutralizing measure in all substances, only the negative kind (resinous) is mobile. It is solely the motion of negative electricity that constitutes a current through a wire. One of the consequences of this fact is that the actual direction of motion of the electricity constituting a current is from the *negative* terminal to the *positive*. This is necessitated by Dufay's observation, abundantly confirmed, that *charges of like sign repel and of unlike sign attract each other.*

Unfortunately Ampère, in his later study of electric currents, had no means of knowing that only negative electricity was mobile. At that time nobody knew whether a current consisted of negative flowing as just described, or of positive flowing in the opposite direction, or of both simultaneously. It was correctly supposed that the external effects, such as the magnetic field and the development of heat, would be the same in any one of the three cases and thus that it made no difference as far as these phenomena were concerned (the only ones known at the time) which of the three alternatives was chosen. So Ampère chose one of them arbitrarily, simply to avoid repetitious qualifications. The one that he happened to seize upon was that the current would be considered to flow from the positive pole of a battery to the negative, hence that it consisted of the motion of positive electricity. This was exactly contrary to what is now known to be the case. But it was nearly three quarters of a century before the real nature of the electric current became known. During that time Ampère's convention became so firmly rooted in the literature of electricity that it is now virtually impossible to dislodge it. Even today the most recent books solemnly define the direction of electric flow as from positive to negative, though everybody knows better. To do otherwise would involve so many changes in terminology that it is not worth while to disjoint the whole literature of the subject merely for the sake of the small number of relatively obscure phenomena whose description would thereby be simplified.

The Measurement of Current

The current balance could conceivably be used in the regular practice of measuring currents. Actually it is not sufficiently rapid or convenient to be acceptable for such use. It is a primary standardizing instrument, not a rough-and-ready meter for daily use.

The first ammeter (the name is a contraction of "ampère meter") was devised in the autumn of 1820 while Ampère was presenting his first papers before the French Academy on the outgrowths of Oersted's work. It originated in Halle, Germany, at the hands of one Johann S. C. Schweigger.[1]

The year following his great discovery Oersted had remarked that there seemed to be some relation between the deflection of a suspended magnetic needle and the magnitude of the current causing that deflection.[2] This

[1] Schweigger's *Journal für Chemie und Physik*, *31*, 1–17 (1820).

[2] *Philosophical Magazine*, *57*, 44 (1821).

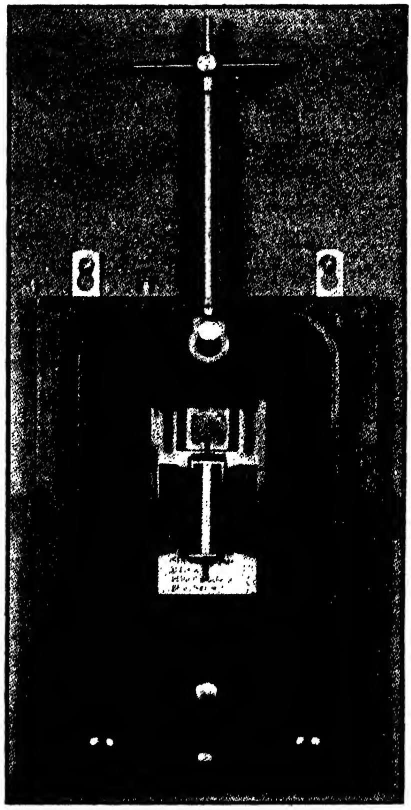


Fig. 358. A Modern Galvanometer of the d'Arsonval Type
(Courtesy of Leeds and Northrup.)

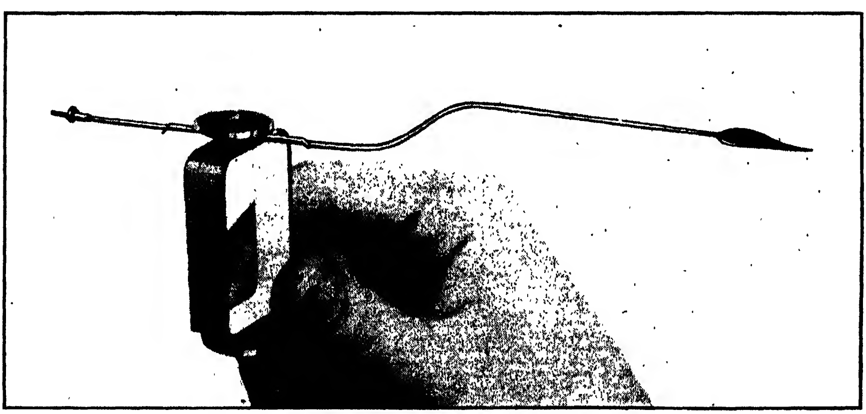


Fig. 359. Coil and Needle of Ammeter
(Courtesy of Leeds and Northrup.)

implied that the deflection might be utilized to give information about the magnitude of the current, a fact which Schweigger had apparently detected before Oersted did. But the great difficulty with an ammeter of that kind, consisting of a single wire and a magnetic needle, was its insensitivity. Schweigger, observing that the deflection produced by a current flowing above the needle in one direction was in the same direction as that produced by an oppositely directed current underneath the needle, simply looped his wire around the needle, thus doubling the effect. A second loop made the response of the needle four times as great for a given current, a third loop six times as great, etc.; whence the name "electromagnetic multiplier." Schweigger's first contrivance was a very humble affair, consisting of two turns of wire within which was placed a compass needle; it is shown in Figure 360. Unimpressive though this appears to a casual observer, probably more electrical history has grown out of this than out of any other single device, with the single exception of the compass itself. During the next fifty years this instrument, consisting, as is evident, of a stationary coil and a moving magnet, became highly refined, largely by William Thomson (Lord Kelvin), as a device for receiving signals over the first Atlantic cable. The nature of the improvements need not concern us, however, since this type of *galvanometer*[1] is now not very commonly used. A modification, consisting of nothing more than making the coil movable and the magnet stationary, has displaced it not only for general laboratory use but also as a receiving instrument in submarine telegraphy, as will appear shortly.

FIG. 360. THE FIRST GALVANOMETER
(From *Edinburgh Encyclopedia* [1831], 3d supplementary vol., Plate 522.)

[1] A generic term, coined by Ampère, in honor of Galvani, applying to the basic type of current-measuring instrument.

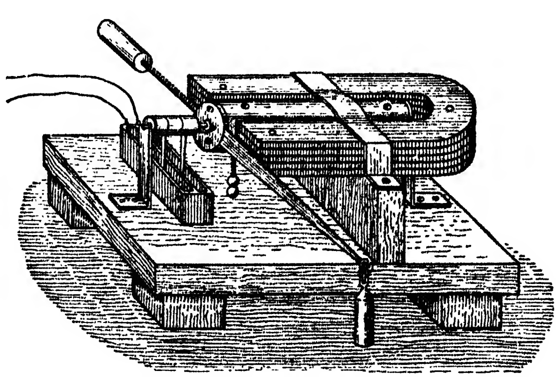


FIG. 361. THE FIRST D'ARSONVAL GALVANOMETER

The Evolution of the Modern Galvanometer

Schweigger's "electromagnetic multiplier" indicated current by the deflection of a movable magnet under the action of the current in a fixed coil. Obviously, if the magnet had been fixed and the coil made movable, the equal reaction could have been utilized to the same effect. Sturgeon actually did this in 1636, but his device was too crude to be accurate and nearly sixty years elapsed before a practical instrument of this kind was developed.[1] When it was done, the new *d'Arsonval galvanometer* displaced the older style quickly and completely. It will repay examination because it embodies the working principle of many of the electrical instruments in common use today. A strong and substantially uniform magnetic field is furnished by a heavy permanent magnet of modified horseshoe shape. A coil, sensitively but ruggedly suspended between the poles of this magnet, experiences a torque when traversed by a current. In Figure 361 this torque was measured by the position of a sliding weight necessary to counterbalance it, a true "current balance." In later models a pointer and scale were added and a spiral spring restored the coil and the needle to the zero point when the current ceased to flow.

The moving-coil galvanometer had been brought to a high state of perfection by Lord Kelvin under the name of the "siphon recorder" and universally used as a receiving instrument in submarine telegraphy ten years before the same device, independently invented, came into use under the name d'Arsonval galvanometer. But the siphon recorder was patented and the royalties furnished its inventor with a very tidy income. This may explain why the modern form of the moving-coil galvanometer grew out of the later French instead of out of the earlier English invention.

The Evolution of the Ammeter

The term *galvanometer* has come to designate a sensitive current-measuring instrument intended primarily for laboratory use. Though it may possess a scale, the scale does not read in any specified units. The variety of conditions under which a galvanometer may be used generally involves a new calibration for each situation. When, however, the conditions of use become standardized and especially when extreme sensitivity is no longer required, two minor modifications will convert this type of galvanometer into a conventional ammeter.

Suppose, then, that it were found, as would invariably be the case, that the currents to be measured in a representative instance of non-laboratory use were too large for the galvanometer at hand. Figure 362 suggests the way in which the solution of this problem evolved. The first step would be to use two or more identical galvanometers, connected in such a way that the current would distribute itself equally between them. The reading of any one of the instruments, multiplied by the number of instru-

[1] *La Lumière Électrique*, *2*, 462 (1880) and *4*, 310 (1881).

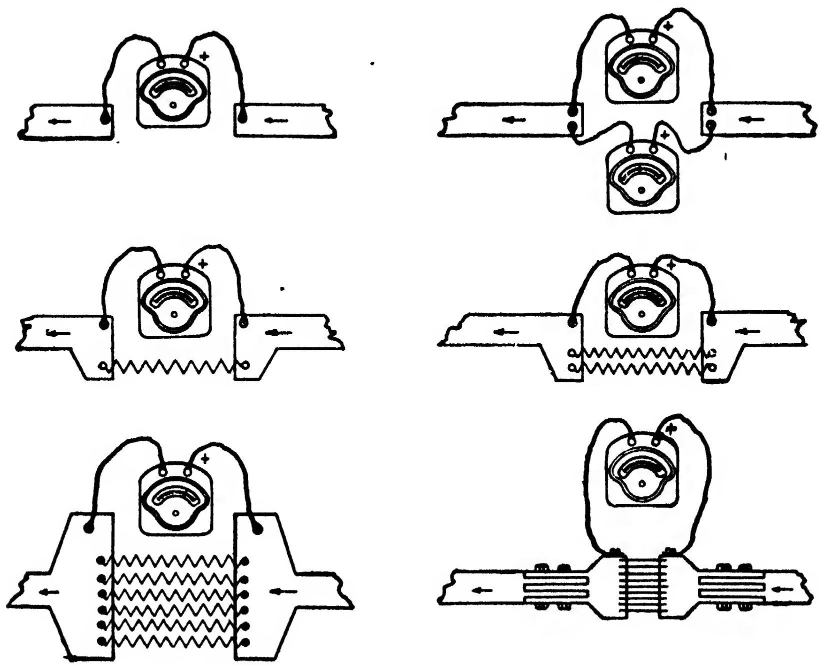


FIG. 362. THE EVOLUTION OF THE GALVANOMETER SHUNT
(Courtesy of Leeds and Northrup.)

ments, would then give the total current. Or, to avoid using so many complete instruments, one galvanometer connected in the same way with merely the separate coils of as many others as were required would secure the same result. Or, finally and most simply of all, a single wire which would have the same effect as all the extra coils could be selected by trial and connected across the terminals of the galvanometer. Only a predetermined fraction of the current would then flow through the galvanometer coil, the remainder flowing through the by-passing wire connected to its terminals.

Such a conductor, termed a *shunt,* plays an important part in every ammeter. By changing the resistance of this shunt, any desired fraction of the total current may be made to pass through the coil, the scale being graduated to indicate the *total current,* not merely the part flowing through the coil. The method of establishing these graduations is described on page 470. After a given shunt is provided, the scale may be graduated and the galvanometer becomes a full-fledged ammeter. The coil and needle are shown in Figure 359, page 445.

The Dynamometer *Instrument*

Another form of ammeter, somewhat less common than the one just described, is illustrated in Figure 363. It is called the dynamometer type,

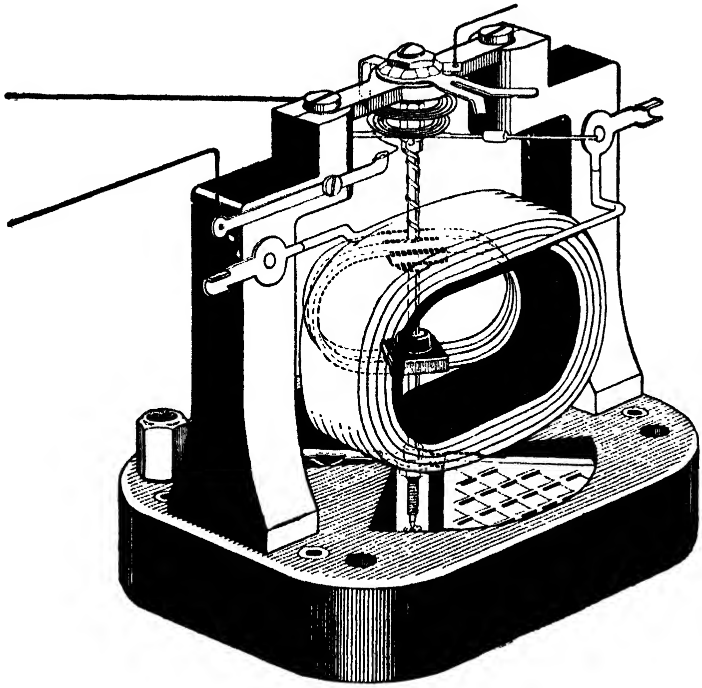


Fig. 363. Dynamometer Type of Ammeter
(Courtesy of Weston Electrical Instrument Company.)

the name apparently having been selected for no good reason. The instrument consists of two coils, one fixed, the other movable, instead of a coil and a magnet. If the same current flows in both coils, the movable coil will experience a torque and will turn against the restoring force of a spiral spring fastened to it. The attached pointer will indicate current in the same way as before.

One of the uses of the dynamometer type of instrument is in connection with the alternating current which is today almost universally used. In the former type of ammeter, reversal of current causes reversal of deflection. Hence, rapid alternations (120 reversals per second in the usual A.C.) would cause the needle simply to tremble somewhat while remaining at the zero position. With the dynamometer instrument, however, simultaneous reversal of the current in both coils produces re-reversal of the deflection of the needle. Each surge of current acts in the same direction, notwithstanding the successive reversals, and the needle maintains a position on the scale which is determined by the average value of the fluctuating current.

The dynamometer instrument can usually be distinguished from the d'Arsonval instrument by the character of the scale. The d'Arsonval scale is uniform, or nearly so, while the dynamometer scale is "compressed" at the low end and "stretched out" at the high end. (See Fig. 364.) The reason is evident. Doubling the current doubles the torque on a d'Arsonval instrument; but in a dynamometer instrument, the torque is doubled twice, once due to the fixed coil and again due to the movable. Hence the scale of such an instrument must be graduated proportionally to the squares of the currents indicated.

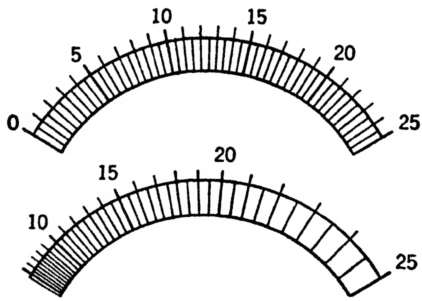


FIG. 364. SCALES OF TWO TYPES OF AMMETER

There are other types of electric measuring instruments than those which have now been described, differing both as to construction and as to purpose. The voltmeter, for measuring potential difference, has not yet been described, nor has the wattmeter for measuring electric power. Before these can be developed, however, it will be necessary to become acquainted with another major principle of current electricity, Ohm's law.

Questions for Self-Examination

1. Describe Oersted's discovery of electromagnetism.
2. Tell about Ampère's extension of Oersted's discovery and state his "right-hand rule."
3. How did Faraday demonstrate the existence of a circular magnetic field around a current?
4. What can be said about the direction of flow of electric currents?
5. Describe the principle of an ordinary galvanometer.
6. How may an ammeter be made out of a galvanometer?
7. Describe the "dynamometer" type of ammeter. Of what use is it?

CHAPTER 38

Potential Difference

The Concept of Potential Difference

A basic prerequisite to the establishment of a system of units is a clear identification of the entities for which units are required. Up to the time of Ampère, not even a beginning had been made at clarifying the concepts of electricity and magnetism. Ampère himself made the first step in this direction. One of his memoirs on the actions between currents began:[1]

> Electromotive action is manifested by two sorts of effects which I believe I should first distinguish by precise definitions. I shall call the first *electric tension*, the second *electric current.*

Ampère then proceeded to distinguish the two "effects." His names for both have survived. That for the first may be seen on the signs attached to the steel towers which support electric power lines. "DANGER! High *tension* line." The same electrical attribute is more commonly termed *potential difference* or, less formally, *voltage.* Ampère's second term, *current*, is now the most common word in electrical literature.

Possibly it is "carrying coals to Newcastle" to enlarge, in this electrical age, on these two concepts. And yet one hears sufficiently often such expressions as "a current of so-many volts" tc give ground for the suspicion that while twentieth-century youth may have the technical patter, there is often lacking a corresponding comprehension of the terms which are used so glibly. It was a mark of rare discrimination on Ampère's part to have discovered and emphasized the distinction between potential difference ("tension," "voltage") and current. It is just as clearly the mark of a lack of discrimination, inexcusable in the twentieth century, to confuse the two by associating one with the other, as in the above expression.

Potential difference is sometimes described as the electrical condition which produces a flow of electricity. This indicates a certain similarity between potential difference and pressure difference between parts of a water system. That is, we may picture an electric current being kept in motion from one point to another in a circuit by the potential difference between those points much as a water current is kept in motion through a pipe by the pressure difference between the ends of the pipe. This analogy is often invoked and is indeed very helpful. But in spite of its utility, it is

[1] *Annales de Chimie et de Physique* (2), *15*, 59 (1820).

not quite complete. The necessary amendment to it may be provided by a slight shift in emphasis. Some water systems pump water into an elevated tank and the subsequent distribution occurs from this point. The elevation of the tank produces pressure in the system, but that is now an incidental and in a certain sense an irrelevant feature. The elevation of the tank necessitates an expenditure of power to keep water flowing into it, and emphasizes what is, in fact, the crucial point, that it is to this expenditure of power that the pressure and hence the flow of the water in the pipes is due when other conditions permit such flow to occur.

The Definition of the Volt

Potential difference is much more closely analogous to energy than to pressure. Actually, the word *potential* may be regarded as an abbreviation of the term *potential energy*, the kind of energy that the elevated water possesses, the rate of change of which is responsible for the subsequent flow. Analogously, electrical potential is due to energy of one kind or another, imparted to an electrical system, the resulting rate of expenditure of this energy being responsible for the current if other conditions are favorable. Hence, the unit of potential difference, named the *volt* in honor of Alessandro Volta, is defined as *the difference of electrical potential between two points on a wire carrying a current of one ampere when the power dissipated between these points is one watt.* Expressed algebraically where E represents difference of potential in volts, P power in watts, and I current in amperes:

$$E = \frac{P}{I}, \quad \text{or} \quad P = EI \tag{1}$$ [1]

The watt has already been encountered as a unit of power in the section on mechanics (page 152). It is much more commonly used as a unit of electrical power than as a unit of mechanical power, though it is the same unit in either case. Small electric accessories, for example, electric lights, are usually rated in watts and heavy electric machines in kilowatts. Hence the use of the watt along with the ampere in defining potential difference is simply a return to a unit already defined in mechanics and which is even more familiar in its present electrical setting.

The Measurement of Potential Difference

Equation (1) suggests the basic or "absolute" method of measuring potential difference. Measure the power P in watts required to maintain a current of I amperes between two points in a circuit. The ratio P/I then represents the potential difference between those two points. It would be possible to use this method to measure potential difference. In practice, voltage measurements can be made much more rapidly and conveniently. As in the case of current measurement (page 444), the "absolute" method is not acceptable for ordinary routine work.

[1] Another definition of the volt will be found on page 457.

The common instrument for measuring potential difference, known as a *voltmeter*, is really an ammeter, though with some minor modifications adapting it to the new purpose. The current flowing through any wire is proportional to the potential difference between its terminals. This was discovered in 1827 by Georg Simon Ohm (1789–1854), whose work will be studied more in detail in the following chapter. Ohm's discovery applies to the current flowing through an electrical instrument as well as through any other conductor of electricity. Hence the deflection of an ammeter will indicate, not only the current flowing through it, but also the potential difference between its terminals. Hence an ammeter may be graduated in volts and used as a voltmeter. To adapt an ammeter to the measurement of voltage in the ranges ordinarily encountered, its electrical characteristics must be considerably modified in ways that will be described immediately. The point of principal importance is that most instruments ordinarily termed voltmeters are in reality modified ammeters, depending for their operation on the proportionality of voltage to current in a given conductor.

On page 448 the use of a so-called "shunt resistance" was described as a step involved in converting a galvanometer to an ammeter. The corresponding step in converting a galvanometer into a voltmeter is the introduction of a "series resistance." The idea may be developed in somewhat the same way. Refer to Figure 365, which corresponds to Figure 362 on page 448. This time it is the value of the voltage between two terminals that is desired instead of the current. A galvanometer is connected to the terminals with the expectation of using it as a voltmeter. Suppose it were found, as would certainly be the case, that the voltage is so large as to drive a current through the galvanometer greater than it can carry. The first step is to use two or more identical galvanometers connected in such a way that the voltage will distribute itself equally between them. Comparison of Figures 362 and 365 will show that this involves a different arrangement of the meters for the measurement of voltage than for the measurement of current. The galvanometers are now said to be connected in *series*. The

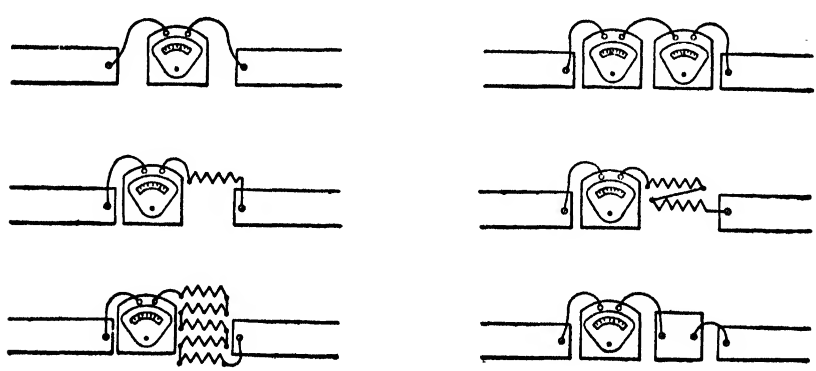


FIG. 365. THE EVOLUTION OF THE SERIES RESISTANCE IN THE VOLTMETER

reading of any one of these series-connected instruments, multiplied by the number of instruments, would then give the total current. But, to avoid using so many complete instruments, one galvanometer connected in the same way with merely the separate coils of as many others as are required will secure the same result, or, as is actually done, a single conductor having a resistance equal to the sum of the resistances of the required number of coils. This constitutes the series resistance which, incorporated into the case with the galvanometer, converts it to a voltmeter. The scale may now be graduated to indicate the *total voltage* between the original terminals, not merely the part of it which exists between the terminals of the galvanometer proper.

Connection of Electrical Instruments

Figure 366 shows how an ammeter would be connected to measure the current delivered to a group of three electrical devices, for example, three electric lights. All of the current delivered to these lights has to flow through the ammeter. No change in voltage will directly affect the reading of the ammeter. Figure 367 shows in the same way how a voltmeter would be connected to measure the voltage applied to the same three lamps. No change in the current delivered to the lamps will directly affect the reading of the voltmeter.

One of the first principles of instrument design is that the use of the instrument shall not affect the value of the entity being measured. This requirement is met in the ammeter by the shunt providing, as it does, an easy channel for the flow of the current so that interposing the ammeter in the circuit introduces no appreciable obstacle to the flow of current. It is met in the voltmeter by the series resistance. This being high, as it always is, prevents the diversion of any appreciable current through the voltmeter in excess of that required to actuate the instrument itself.

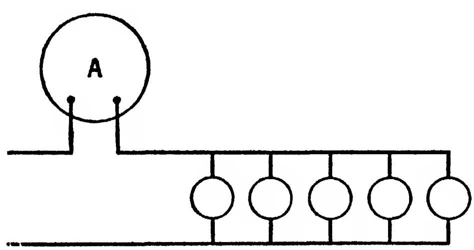


FIG. 366. CONNECTING AN AMMETER

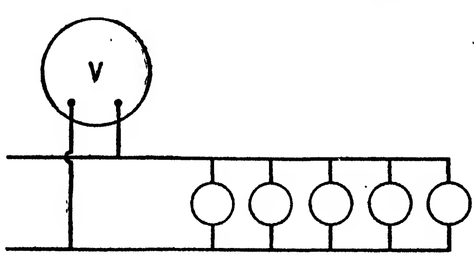


FIG. 367. CONNECTING A VOLTMETER

If a voltmeter should mistakenly be connected as one would an ammeter, so little current could traverse it that the lamps would probably not even glow dimly, but otherwise no harm would be done. If the opposite mistake should be made however, connecting an ammeter as one would a voltmeter, a casualty would immediately result. Such a heavy current would flow through the ammeter as to burn it out in a fraction of a second.

The Measurement of Electric Power

Figure 368 is a combination of Figures 366 and 367. It shows an ammeter and a voltmeter connected to the same circuit, indicating respectively the current through the three lamps (and incidentally through the voltmeter besides) and the voltage across the lamps. Now the second of equations (1) on page 452 shows that the product of volts and amperes represents electric power in watts. Hence the arrangement of Figure 368 permits the calculation of the power consumption in an electric circuit, merely by finding the product of the readings of an ammeter and a voltmeter connected into that circuit.

But with the aid of a dynamometer type of instrument (page 449) power consumption in watts may be measured directly instead of requiring two readings multiplied by each other. With each of the two coils possessing separate terminals, one coil may be equipped with a shunt to function like an ammeter, the other with a series resistance to function like a voltmeter. With each coil then connected independently into the circuit in the manner appropriate to its function, the deflection will be proportional to the product of volts and amperes, the scale may be graduated to read directly in watts, and the determination of power consumption will be correspondingly simplified.

It is scarcely necessary to warn the reader against confusing the wattmeter as above described with the "wattmeter" (incorrectly so-called) used to meter electric service. The latter records watt-hours and multiples, not watts. The product of power by time is electrical energy, not power. The principle of action of the watt-hour meter is, moreover en-

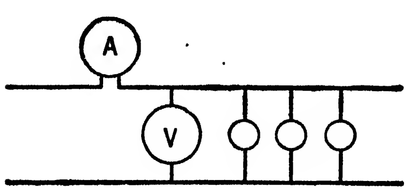


Fig. 368. Measuring Electric Power Consumption with Voltmeter and Ammeter

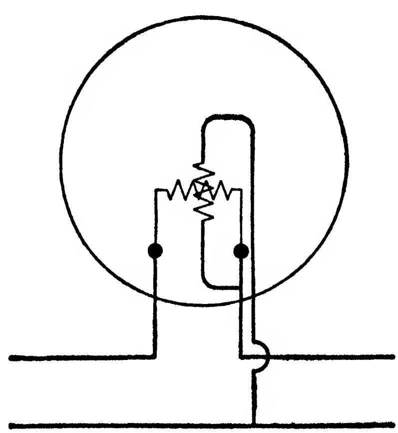


Fig. 369. Dynamometer Connected as a Wattmeter

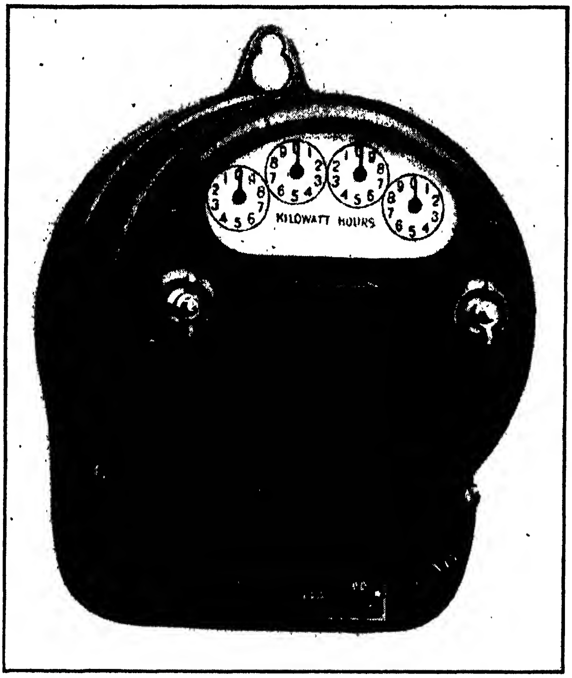


FIG. 370. WATT-HOUR METER
(Courtesy of Weston Electrical Instrument Company.)

tirely different from that of the wattmeter. The former is a small electric motor whose speed varies with the power consumption, driving a set of gears which records the energy used.

Thus, the commodity which the user of electricity purchases from his power company is electrical *energy*, not a quantity of electricity, as is often supposed. Just as much electricity leaves the user's premises as arrives there. All the user does is to strip the electricity of its energy and return it to the power house to be supplied with energy again. The wattmeter, properly so called, merely indicates the momentary rate at which such energy is being used. That is, it registers power consumption in watts. The watt-hour meter measures the product of power by time, that is, energy. If the dials of this instrument were graduated in watt-seconds, it would indicate energy in joules, the energy and work unit already familiar in mechanics and heat. Instead common practice has established the watt-hour, otherwise unnamed, as the commercial unit of electric energy. Its magnitude is, of course, 3600 joules.

Electromotive Force

Equation (1) (page 452) may be expressed somewhat differently after having multiplied both sides by t.

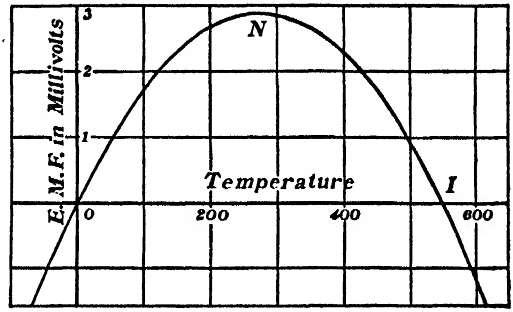


FIG. 371. THERMO E.M.F. AS A FUNCTION OF TEMPERATURE

Thus, $$Pt = EIt$$

whereas $$W = EQ \quad \text{or} \quad E = \frac{W}{Q}, \qquad (2)$$

where W represents work, the product of power and time, and Q represents quantity of electricity which is the product of current and time by equation (5) (page 434). Equation (2) is sometimes used as a defining equation for the volt in place of equation (1). Stated in words, the volt, by this definition, is *the difference of electrical potential between two points when one joule of work is expended in moving one coulomb between these points.* This definition renders more plausible the term *potential* (contraction for *potential energy*), when used in electricity, since potential difference is now being defined in terms of work. When a battery, a generator, or other source of electric energy is involved, the potential difference which it is capable of developing when no current is being delivered is commonly termed *electromotive force,* commonly abbreviated to *e.m.f.* The name is both awkward and inappropriate, but seems to have fastened itself too firmly in electrical terminology to be dislodged. The foregoing definition of the volt implies that an electric charge, placed between two terminals under a difference of potential, experiences a force. This implication is involved in the statement that moving such a charge requires the expenditure of work, which can only be the case if force is exerted in accordance with the familiar definition of work as force multiplied by distance. Work and distance being given, the force is a necessary corollary. The subject of forces between charged bodies and forces on charges placed in an electric field will be developed in a later chapter.

Thermo-Electromotive Force

In 1821, Thomas J. Seebeck (1770–1831) discovered that heating the junction of two dissimilar metals produced a current. We now know that it is not the magnitude of the current thus produced that is significant, but rather the voltage. This is termed a *thermo-electromotive force.* It varies between wide limits for different pairs of metals, but is always a very small fraction of a volt.

Figure 371 is a curve with temperature as abscissas and thermo-e.m.f. as ordinates. The right intersection of the curve with the temperature axis represents the temperature of the heated junction of the two metals and the left that of the cooled junction. The curve is in fact a parabola, the vertex of which represents the so-called *neutral temperature*. The e.m.f. produced by heating one of the junctions increases as the junction's temperature rises, reaches a maximum at the neutral temperature, decreases to zero at the *temperature of inversion*, and then reverses. At the reversal, the temperature of the heated junction is as far above the neutral temperature as that of the cooled junction is below it.

The *thermoelectric effect*, as this group of phenomena is termed, is applied in many ways, one of the most important of which is the measurement of temperatures that are outside the range of ordinary thermometers. By placing *thermocouples* (the term applied to junctions of dissimilar metals used this way) at the foci of astronomical telescopes, the temperatures of planets can be measured and those of stars compared with each other.

The Photoelectric Cell

On page 297 the so-called photovoltaic cell was mentioned. This is a variant of what is more commonly called the *photoelectric cell*. Just as the thermoelectricity was an electric effect of application of heat, so photoelectricity is an electric effect of the application of light. If light is projected onto certain metals, or especially certain compounds which have been found to be more sensitive than metals, electrons are liberated. If the surface is sealed into an evacuated tube along with another electrode which may be charged positively with a battery, it will gather up these electrons and a weak current will flow through the cell whenever light strikes it.

The time required for photoelectric cells to act is extremely short, less than 3×10^{-9} seconds. Hence they are effective for even such rapid-fire operations as are involved in the sound-track of motion-picture films and even the much more exacting performance of television.

The photovoltaic cell referred to above acts somewhat differently. It requires neither an evacuated tube nor a battery. It is, indeed, its own battery, developing an electromotive force which is proportional to the intensity of the light striking it. The active material is a thin layer of cuprous oxide formed on a base of copper. Over the oxide is deposited an exceedingly thin and fairly transparent layer of some metal to act as an upper electrode. When illuminated, this cell delivers current, the top layer of metal being the positive terminal and the base plate of copper the negative. The currents are tiny, but with the aid of an amplifier (page 581), can be made as heavy as required.

The Electrolytic Cell

But perhaps the most widely useful of all the ways of exciting an electromotive force between dissimilar conductors is the way discovered by Volta. The statement was made on page 428 that the ordinary flashlight battery is a fairly direct descendant of the voltaic pile. The "dry cell," two or more of which make up the ordinary flashlight battery, consists of a zinc cup which contains the chemicals and also acts as the negative electrode. The positive electrode is a carbon rod centered in the cup. Between the two is packed a mixture of manganese dioxide, granulated carbon, graphite, and plaster of paris. This powdery aggregate is soaked with a mixture of ammonium chloride and zinc chloride. The top is then sealed with pitch.

The "storage cell" is an electrolytic cell in which it is possible to reverse the chemical reactions which take place during discharge by forcing an electric current through it in the reverse direction. There are two principal types, each possessing its particular merits and shortcomings. The most common is the lead-acid type, consisting, as the name indicates, of lead electrodes immersed in dilute sulphuric acid. In the process of "charging," one of the plates (the positive) becomes coated with brown lead peroxide. It is not customary to make the positive plates of lead directly, but to give them the form of grids upon which is placed a paste of lead oxide. This increases the area and improves the efficiency of the cell. The e.m.f. of such a cell is about 2.1 volts. It changes a little with varying degrees of charge, but the change in voltage is in far lower proportion than the accompanying change in specific gravity of the acid. Hence it is customary to test the state of charge of such a cell with an hydrometer rather than with a voltmeter. The specific gravity of a fully charged cell of this type is about 1.250, which diminishes to 1.150 as the cell approaches a state of complete discharge.

The plates of the other type of storage cell, the Edison cell, contain nickel oxide (positive) and finely divided iron (negative). The electrolyte is potassium hydroxide and the container is steel instead of glass or hard rubber as in the lead-acid cell. The electromotive force of a completely charged cell is about 1.4 volts per cell, diminishing to one volt as the cell becomes discharged, averaging 1.2 volts per cell. A voltmeter rather than an hydrometer is used to test the state of charge of this type of cell.

The Edison cell is far more rugged mechanically than the lead-acid cell, and deteriorates far less rapidly. It is in fact practically everlasting with ordinary care. But it can deliver only a fraction of the current that the lead-acid cell can and hence is not usable for automobile starting purposes where a hundred or more amperes are frequently required. Also the fluctuation of its voltage, both with state of charge and with temperature is a disadvantage for many purposes.

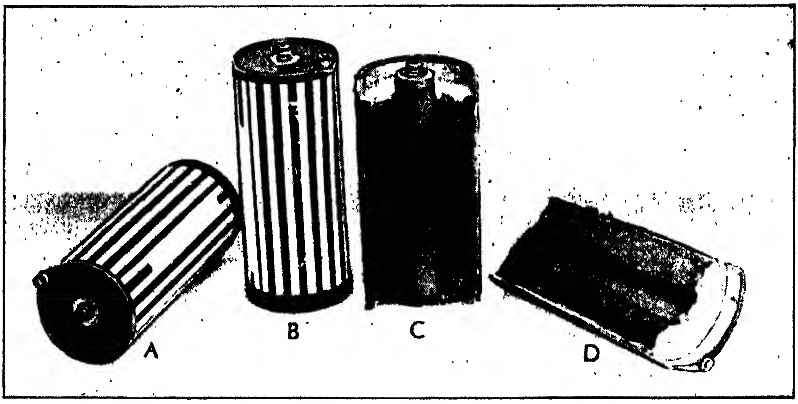


Fig. 372. Cutaway View of a Dry Cell

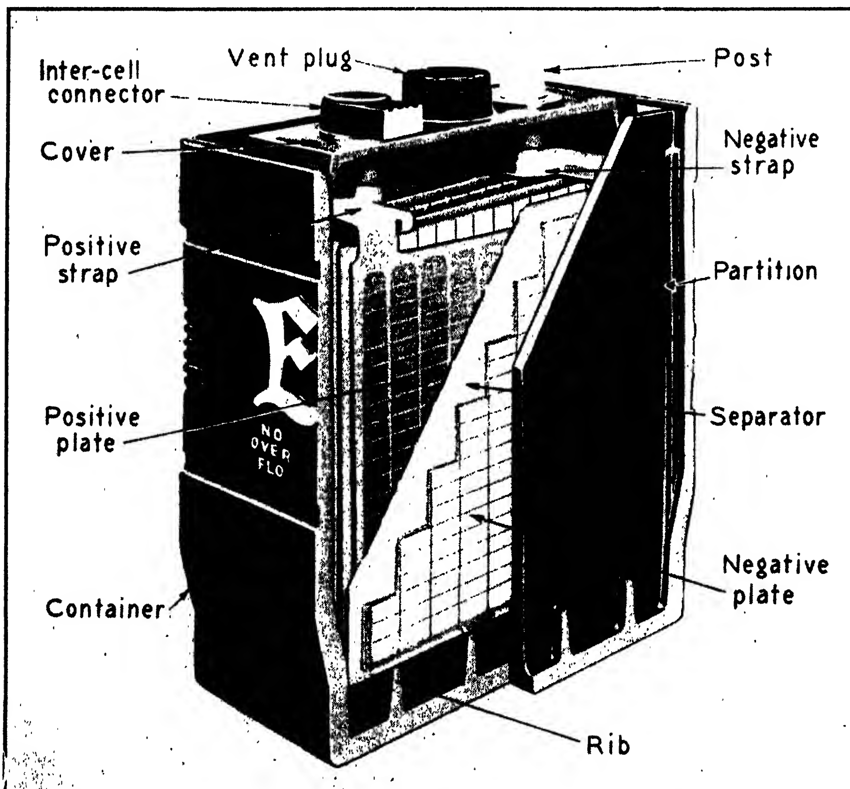


Fig. 373. Cutaway View of Lead-Acid Cell
(Electric Storage Battery Co.)

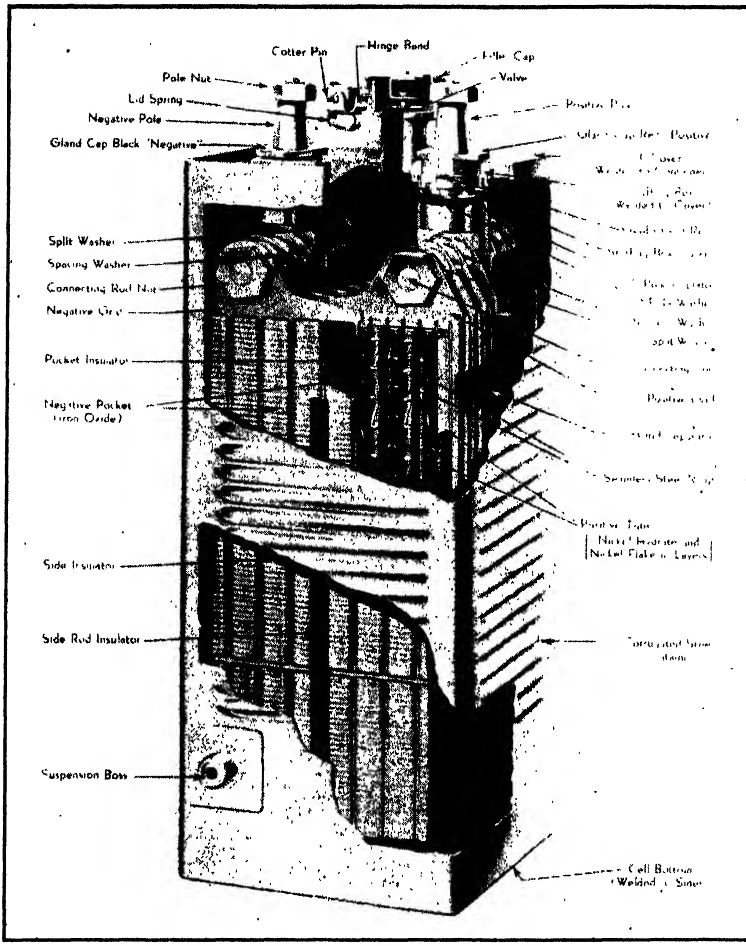


FIG. 374. CUTAWAY VIEW OF THE EDISON STORAGE CELL
(Thomas A. Edison, Incorporated.)

Questions for Self-Examination

1. What is the general meaning of the term *potential difference*?
2. Define the volt, as the unit of potential difference, in two ways.
3. What is the principle of the voltmeter?
4. Show how to connect a voltmeter and an ammeter in a circuit.
5. Distinguish between a wattmeter and a watt-hour meter.
6. What is *thermo-electromotive force* and how does it act?
7. Compare two main types of photoelectric cell.
8. Compare the main types of electrolytic cell.

Problems on Chapter 38

1. An X-ray tube takes I milliamperes at E kilovolts. What is its power consumption P in watts? (The prefixes milli- and kilo- have their usual meanings.)

	I	E	P
1.	1	100	100
	1.5	110	165
	2	120	240
	2.5	130	325

	P	E	I	d
2.	100,000	2	50	1.
	100,000	30	33	.26
	100,000	75	1.3	.16
	100,000	100	1.0	.14

2. High voltage is always used to transmit electric power over long distances. If P kilowatts are transmitted at E kilovolts, what is the current I in amperes and what are the relative diameters d of wire required for a given efficiency?

3. How many amperes I does a 110-volt electric motor developing P horsepower require when its efficiency is e per cent?

	P	e	I
3.	.125	25	3.4
	.500	50	6.8
	10.	90	75.
	100.	97	700.

	w	H
4.	25	2100
	60	5200
	100	8600
	150	13000

4. How many Calories of heat H are developed in an hour by a lamp rated at w watts?

5. How many Calories H per second are developed in an electric furnace which takes I amperes at E volts, and what horsepower P is required to supply the furnace?

I	E	H	P
10	100	.24	1.3
30	50	.36	2.0
50	40	.48	2.7
100	25	.60	3.4

CHAPTER 39

Ohm's Law

Electrical Resistance

In Chapter 17 was developed the idea that different materials conduct heat with different degrees of readiness. The concept of heat conductivity, first clearly formulated by Joseph Fourier in 1822, was extended to include electrical conductivity by a German experimenter, Georg Simon Ohm (1789–1854), in 1826. Ohm's work ranks on a par with the identification of current and potential difference as essential entities in electrical science.

Fourier's work had attracted a great deal of attention. Since it came so soon after the discoveries of Ampère, it is rather surprising that the work did not stimulate others besides Ohm to the recognition of possible parallels between heat and electricity. Fourier had found that the rate at which heat was conducted became greater as the difference of temperature increased, as the area of the conductor was increased, and as its length was decreased. He found also that the rate of heat conduction depended on the material of which the conductor was composed. But Ohm seems to have been alone in seeing the possibility of a similar dependence of electric current on the dimensions and material of wire. Though Ohm's concept of electrical conductivity was anticipated by Henry Cavendish fifty years earlier (26:*1*:77), the significance of Cavendish's work was not to be recognized for many years. Even more important was Ohm's idea that potential difference might play the same rôle in the flow of electricity that temperature difference had been found to play in the flow of heat.

Ohm's first undertaking was to compare the current-conducting powers of wires of the same size, but of nine different materials. Putting these wires into his circuit, he adjusted their lengths until he secured the same deflection of his galvanometer. This instrument, it may be remarked in passing, he had made himself for the purpose. Though it was in principle simply Oersted's magnetic needle and single wire, it was so carefully built that it was probably the first really precise current-measuring instrument. With it Ohm found the lengths of equal-sized wires required to permit the same current to flow.

The "batteries" of Ohm's day were exceedingly erratic, and Ohm's work did not yield dependable results until, following a suggestion by

Kirchhoff, he began to utilize thermocouples as his source of current.[1] These could be kept sufficiently steady, and from the time that he began to use them, Ohm secured dependable results.

Ohm then tried the effects of different lengths of the same wire and found that the galvanometer deflection diminished as the length of wire increased. He then took wires of the same material but of different sizes and found that his galvanometer deflection was the same for each wire when the lengths of the wires were proportional to their cross-sectional areas. These observations established the dependence of the flow of electricity on the dimensions of the wire.

Thus Ohm established three facts. The ease with which a wire conducts electricity (1) depends on the material of which the wire is composed, (2) is inversely proportional to the length, and (3) is directly proportional to the cross-sectional area of the wire. These observations were quite analogous to those of Fourier on the conduction of heat and have been abundantly verified since Ohm's time. But today it is usual to look at electrical conductivity from the reverse side and to concentrate attention on the difficulty, instead of the ease, of transmission of electricity. The reciprocal of the values which Ohm deduced has received the name *resistance*. Stating the above facts in these terms one would say that the resistance R of a wire (1) depends on the material, (2) is directly proportional to the length l, and (3) is inversely proportional to the cross-sectional area a. Stated algebraically:

$$R = \rho \frac{l}{a}, \tag{1}$$

where ρ is a constant, the so-called *resistivity*, characteristic of the material and having values for different materials proportional to reciprocals of the values given in Ohm's table. By giving l and a the value 1 (meter) it will be evident that the resistivity of any material can be found by measuring the resistance between opposite faces of a cube of the material one meter on an edge. A table of resistivities and temperature coefficients (see the next page) of some common metals is given herewith.

It is of passing interest that electrical conductivities and the thermal conductivities of metals were later found to be very closely proportional, the so-called Wiedemann-Franz law (compare page 185). Thus, the parallel between the flow of heat and the flow of electricity proved to be more than a mere analogy.

At one point in his paper Ohm made another remark which is worthy of notice. He said (77:472):

> I cannot avoid mentioning here . . . an observation that the conductivity of metals is increased by lowering the temperature. I took a 4-inch brass conductor and brought it into the circuit; it gave 159 divisions. When I heated it in the middle with an alcohol flame, the force gradually decreased

[1] *Journal für Chemie und Physik*, *46*, 137 (1826).

RESISTIVITIES AND TEMPERATURE COEFFICIENTS

Material	Resistivity (ρ) $\times 10^8$	Temperature Coefficient (α) $\times 10^3$
Aluminum	2.8	3.9
Brass	7.	2.
Copper	1.7	3.9
Carbon	5000.	.5
Gold	2.4	3.4
Iron	10.	5.
Lead	22	3.9
Mercury	95.8	.9
Nickel	7.8	6.
Platinum	10.	3.
Silver	1.6	3.8
Tin	11.5	4.2
Tungsten	5.6	4.5
Zinc	5.8	3.7

by 20 or more divisions;... but when I placed on it a layer of snow, the force increased by 2 divisions.

In 1833 it was found that the change in resistivity occasioned by change of temperature was very nearly proportional to the number of degrees of rise or fall. This discovery was made in 1833 by the Russian physicist H. F. E. Lenz,[1] whose work in another field will be encountered elsewhere. This dependence upon temperature makes it necessary in equation (1) to specify the temperature, which is conventionally taken at 0° C. Then the resistance of the wire in question at any other temperature, $t°$ C., is given by the relation:

$$R = R_0(1 + \alpha t) \quad (2)$$

where R_0 is the resistance at 0° C. and α is a constant termed the *temperature coefficient of resistivity*. The values of this constant for some common metals are shown in the above table.

The Discovery of Ohm's Law

But Ohm's greatest contribution was still to be made, in the form of what has come to be known as *Ohm's law*. In the course of his experiments, he had noticed that if, maintaining his circuit otherwise unaltered, he changed the "electric force" (the voltage of his source), the deflection of the galvanometer changed proportionally. But if, maintaining the voltage unaltered, he changed the resistance of the wires constituting the circuit, the deflection of the galvanometer changed in inverse proportion to the resistance. He generalized on this in a statement the modern equivalent of which would be the following (94:36):

[1] *Poggendorff's Annalen, 34*, 418 (1833).

> The current flowing in any circuit is directly proportional to the algebraic sum of all the impressed voltages and inversely proportional to the total resistance.

Expressing the foregoing proportionalities algebraically, I representing current, E potential difference, and R resistance,

$$I \propto \frac{E}{R}, \quad \text{or} \quad I = k\frac{E}{R} \tag{3}$$

The value of the proportionality factor k appearing in the second equation will be determined by the choice of the unit of resistance, which may now be made.

Equation (3) can be restated

$$R = k\frac{E}{I} \tag{3'}$$

But the unit of current, the ampere, has already been defined as has also the unit of potential difference, the volt. If, now, a conductor is so adjusted that a current of one ampere flows through it when the potential difference between its terminals is one volt, and its resistance is termed one unit, k assumes the value unity and drops out of the equation; whence

$$R = \frac{E}{I}, \tag{4}$$

Thus the unit of resistance, now called the *ohm*, is the *resistance of a conductor which will permit one ampere to flow when its terminals are maintained at a potential difference of one volt.*

Equation (4) or either of its algebraic variations

$$I = \frac{E}{R}, \quad \text{or} \quad E = IR, \tag{4'}$$

is the form in which Ohm's law is usually stated. The formulation of this law was the high point of Ohm's work.

Potential Distribution in a Circuit

Ohm's law, like all really great generalizations, possesses many implications not apparent on the surface. Such byways invite exploration. Some of them Ohm himself explored. Others have been opened up only with the later development of physics.

Consider the simplest of all electrical circuits: a uniform wire, say, of 3 ohms resistance connecting the terminals of a battery which, when its terminals are connected by the wire, will maintain a potential difference of 9 volts. The existence of this potential difference could be observed with the aid of a voltmeter — an instrument whose construction will shortly be examined — connected as shown in Figure 375. The current flowing in all parts of the wire is given by Ohm's law as:

$$I = \frac{E}{R} = \frac{9}{3} = 3 \text{ amperes.}$$

Ohm asked himself the question as to the distribution of potential along the wire under these circumstances. He concluded that if the right-hand terminal of the voltmeter should be moved away from the battery along the wire, the potential difference registered by the instrument would be seen to diminish steadily from the maximum value of 9 at the position shown in the figure, down to 6 at the position B,[1] $4\frac{1}{2}$ at C, 3 at D, and to approach zero as the moving contact approached the other terminal of the voltmeter at E. If the foregoing potential differences are represented by distances perpendicular to the plane of the wire as in Figure 376, a graphical picture of the potential distribution results. If the circuit is divided at the battery and spread out in a plane, the result is the triangle FGH, in Figure 377 taken from Ohm's book. This way of regarding a circuit sheds

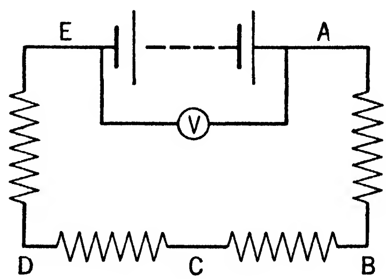


FIG. 375. A SIMPLE CIRCUIT

[1] The points B and D trisect the resistance and the point C bisects it.

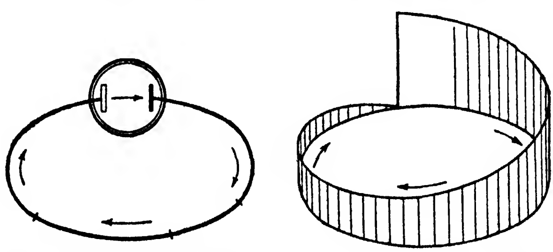


FIG. 376. POTENTIAL DISTRIBUTION IN A SIMPLE CIRCUIT

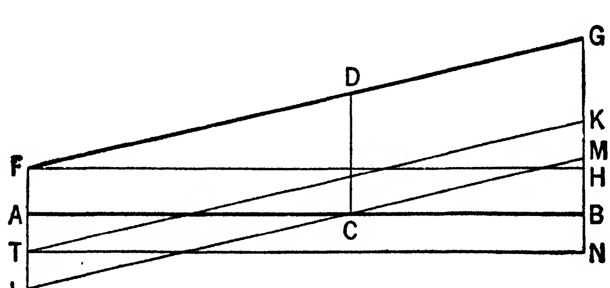


FIG. 377. OHM'S REPRESENTATION OF POTENTIAL DISTRIBUTION IN A SIMPLE CIRCUIT

a certain amount of light on the properties of electric circuits. But what is far more important, it can be made to yield a general rule, which is in effect another interpretation of Ohm's law second only in importance to the interpretation heretofore placed upon it.

Ohm's Law Applied to Portions of a Circuit

In its original form, Ohm's law was made to apply to a circuit in its entirety. V represented "the algebraic sum of all the impressed voltages" and R the "total resistance" (page 466). It may now be shown that the law is also applicable to *any individual part* of a circuit. For example, consider the part $BCDE$ of Figure 375. The current is 3 amperes, the resistance of the entire wire being 3 ohms, that of BE (two thirds of it) is 2 ohms. The observation was made that the voltmeter would read 6 when the terminals were connected to the circuit at B and E. If Ohm's law applies to this section, substitution of $I = 3$, $V = 6$, and $R = 2$ in the equation

$$I = \frac{e^*}{r} \tag{5}$$

* In applying Ohm's law, capital letters will be used for quantities associated with the circuit as a whole; small letters for quantities involved in portions of the circuit. For example, in the above case, I represents the current in the whole circuit, r the resistance of a portion of it, and e the potential difference across that portion.

should produce an equality, as it will be seen to do. Ohm's law can also be shown to apply to the sections CE and DE.

The crucial difference between the total-circuit and the partial-circuit uses of Ohm's law centers in the interpretation of E (or e). In the total-circuit case E stands for the algebraic sum of the voltages maintained by the actual sources of electric power in the circuit. In the partial-circuit case e stands for the fall of potential along the wire caused by its resistance. If the portion of the circuit under consideration includes any source of electrical power, the corresponding voltage (termed "electromotive force" when appearing in this rôle) has to be added in with due attention to algebraic sign. This is not, however, the usual case.

Conductors in Series

These two interpretations of Ohm's law make it possible to find how the aggregate resistance of a group of conductors depends on the individual resistances of which the aggregate is composed. Basically there are two ways in which conductors may be joined together: in *series*, that is, end to end such that the entire current passes through each wire serially; and in *parallel*, that is, branched, so that the current divides itself between the conductors.

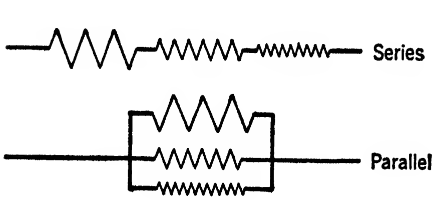


FIG. 378. THE TWO BASIC WAYS OF COMBINING CONDUCTORS

There are other ways of connecting conductors together, but these two will cover all the cases to be encountered here. Naturally, the aggregate resistance will depend on how the individual wires are put together.

Consider, then, a case of three conductors in series, of resistances r_1, r_2, and r_3, respectively (Fig. 378). Let a current I be flowing through them. To each wire the partial-current form of Ohm's law may be applied. Thus:

$$e_1 = Ir_1, \qquad e_2 = Ir_2, \qquad e_3 = Ir_3.$$

But the sum of these successive potential differences, v_1, v_2, and v_3, must equal the outside electromotive force V which is driving the current through these wires. Hence,

$$E = e_1 + e_2 + e_3 = Ir_1 + Ir_2 + Ir_3 = I(r_1 + r_2 + r_3). \tag{6}$$

But this is Ohm's law for the total circuit, the general form of which is

$$E = IR. \tag{7}$$

A comparison of (6) with (7), both expressing the same fact, shows that

$$R = r_1 + r_2 + r_3. \tag{8}$$

Hence, *the aggregate resistance of any number of conductors in series is the sum of their individual resistances.*

Conductors in Parallel

The case of parallel conductors can be treated similarly (Fig. 378). Again apply the partial circuit form of Ohm's law to each conductor. In this case the value of the current is different in each wire, but the sum of the currents in each wire must equal the total current flowing. Thus

$$I = i_1 + i_2 + i_3. \tag{9}$$

But the potential difference across each wire must be the same, since all three wires branch from the same two points, and this potential difference must equal the outside electromotive force which drives the total current through the circuit. That is,

$$i_1 = \frac{E}{r_1}, \qquad i_2 = \frac{E}{r_2}, \qquad i_3 = \frac{E}{r_3}. \tag{10}$$

Substituting (10) into (9)

$$I = \frac{E}{r_1} + \frac{E}{r_2} + \frac{E}{r_3} = E\left(\frac{1}{r_1} + \frac{1}{r_2} + \frac{1}{r_3}\right). \tag{11}$$

But this is Ohm's law for the total circuit, the general form of which is

$$I = \frac{E}{R} \quad \text{or} \quad I = E\frac{1}{R}. \tag{12}$$

A comparison of (11) with (12), both expressing the same fact, shows that

$$\frac{1}{R} = \frac{1}{r_1} + \frac{1}{r_2} + \frac{1}{r_3}. \tag{13}$$

Hence, *the aggregate resistance of any number of conductors in parallel is the reciprocal of the sum of their reciprocals.*

Shunt and Series Resistances in Electrical Instruments

The use of shunts to bring galvanometer sensitivities within the ranges required by workaday ammeters was described on page 448. The use of series resistances in voltmeters for the same purpose was described on page 453. Ohm's law can be used to determine the relation between the resistance of a galvanometer and that of the shunt or series accessory to adapt the resulting instrument to use within a given range.

Suppose, in the case of an ammeter, that the range of a galvanometer of resistance g ohms is to be multiplied A times by the provision of a shunt, the resistance s of which is to be calculated. This means that a given difference of potential must send A times as much current through the completed ammeter as through the original galvanometer. The resistance g must hence be A times the aggregate of g and s in parallel. That is,

$$g = A \frac{1}{\frac{1}{g} + \frac{1}{s}}$$

or, solving for s,

$$s = \frac{g}{A - 1}. \tag{14}$$

Thus, multiplying the current range of a galvanometer by factors of 10, 100, etc., requires shunts having resistances $\frac{1}{9}$, $\frac{1}{99}$, etc., times that of the original instrument.

In the case of a voltmeter, suppose that the range of a galvanometer of resistance g ohms is to be multiplied V times by the provision of a series resistance r ohms, the value of which is to be calculated. This means that V times as much voltage will be required to send a given current through the completed voltmeter as through the original galvanometer. The total resistance of the voltmeter must hence be V times that of the galvanometer alone. That is,

$$g + r = Vg,$$

or, solving for r,

$$r = g\,(V - 1). \tag{15}$$

Thus, multiplying the voltage range of a galvanometer by factors of 10, 100, etc., requires series resistances 9, 99, etc., times the resistance of the original instrument.

The Measurement of Resistance

Besides the instruments for measurement of current, potential difference, and power, those for measurement of resistance require attention. One way to measure resistance and in some respects the simplest is with the aid of a voltmeter and ammeter. If V volts are thus observed to drive I amperes through the conductor under test, then by Ohm's law

$$R = \frac{V}{I}.$$

The method is especially adaptable to extremely high and extremely low resistances, instruments of appropriate ranges being selected.

But for accurate determinations of resistance over ordinary ranges the so-called "Wheatstone's bridge" is the standard device. It did not originate with Sir Charles Wheatstone (1802–75). He himself says that his own (presumably independent) treatment of it was anticipated by ten years by a man named S. H. Christie. Notwithstanding this, Wheatstone's name has always been associated with the instrument. Since his description of it leaves little to be desired by way of clarity, it will be reproduced here.[1] He called it a *Differential Resistance Measurer.*

> Figure 379 represents a board on which are placed four copper wires *Zb, Za, Ca, Cb,* the extremities of which are fixed to brass binding screws. The binding screws *Z, C* are for the purpose of receiving wires proceeding from the two poles of a battery, and those marked *a, b* are for holding the ends of the wire of a galvanometer. By this arrangement a wire from each pole of the battery proceeds to each end of the galvanometer wire, and if the four wires be of equal length and thickness, and of the same material, perfect equilibrium is established, so that a battery however powerful will not produce the least deviation of the galvanometer from zero. . . . But if a resistance be interposed in either of the four wires (The terminals *cd* and *ef* are provided for this purpose), the equilibrium of the galvanometer will be disturbed. . . . It may be restored by placing an equal resistance in either of the adjacent wires.

[1] *Philosophical Transactions, 133,* 323 (1843).

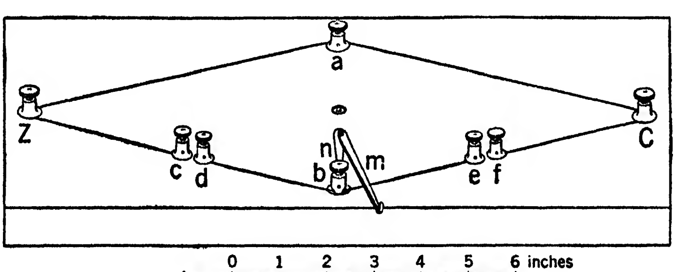


FIG. 379. WHEATSTONE'S ORIGINAL BRIDGE DIAGRAM

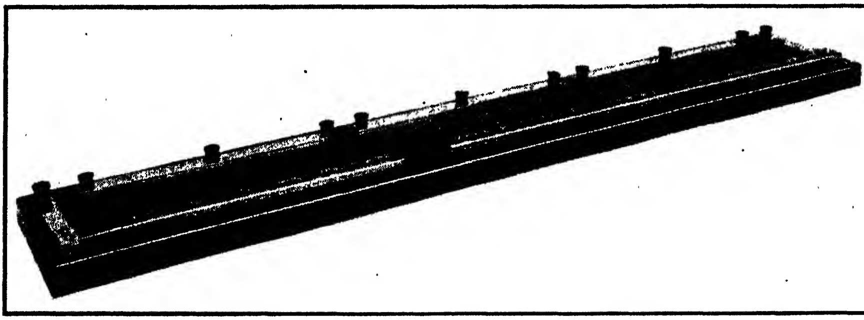


FIG. 380. A SLIDE-WIRE BRIDGE
(Courtesy of Leeds and Northrup.)

The Theory of Wheatstone's Bridge

The use of Wheatstone's bridge involves initially adjusting for equilibrium, introducing the unknown resistance in one branch, and placing a variable but known resistance in the opposite branch. When the latter has been altered until the galvanometer needle returns to zero, the value to which it has been adjusted tells the resistance of the unknown. It is not necessary, moreover, that the resistances Za and Ca be equal, though Wheatstone implies that it is. All that is necessary is that their ratio be known. Then when equilibrium is attained the ratio between the unknown and the standard resistances is the same.

The foregoing principle is applied in two forms, the slide-wire bridge and the box bridge. In the former a single uniform wire replaces ZaC, a sliding-knife-edge contact performing the function of the terminal a. After approximate adjustment of the known to near-equality with the unknown, the final balance is secured by changing the position of a. Then

$$\frac{cd}{ef} = \frac{Za}{Ca}, \tag{16}$$

and if the ratios of the lengths Za and Ca are observed along with the known resistance ef, the value of the unknown is immediately calculable. With the box bridge, pairs of resistance spools carefully adjusted to decimal ratios take the place of ZaC. Among other advantages of this type, one standard resistance can do duty for all decimals of its value as well. This type of bridge frequently contains within one small box the battery, galvanometer, and standard resistance in addition to the ratio coils replacing ZaC. It is used almost universally in commercial testing.

To justify equation (16) note that when no current flows through the galvanometer the points a and b must be at the same potential. Hence, the fall of potential from Z to b must be the same as that from Z to a and similarly for the fall from b to C and from a to C. Hence, by Ohm's law

$$i_{Zb} \cdot cd = i_{Za} \cdot Za \quad \text{and} \quad i_{Cb} \cdot ef = i_{Ca} \cdot Ca. \tag{17}$$

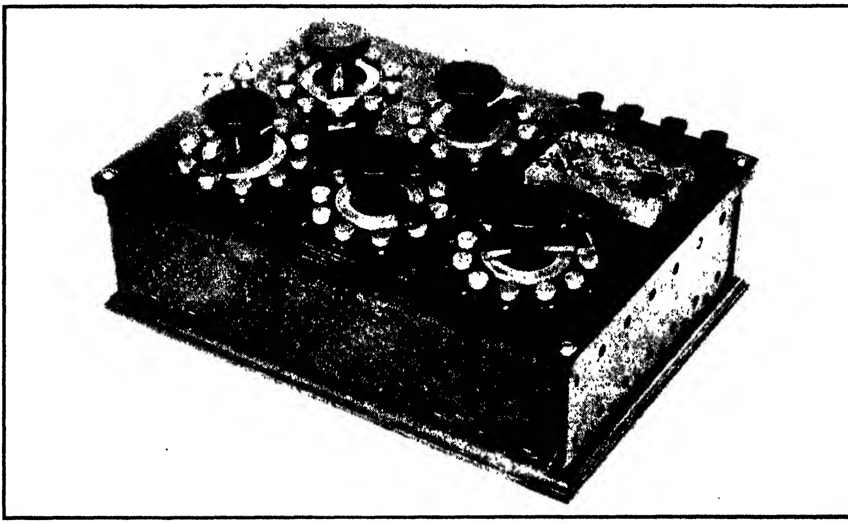


FIG. 381. A BOX BRIDGE
(Courtesy of Leeds and Northrup.)

But since all the current through *cd* passes also through *ef* and similarly for the other two,

$$i_{Zb} = i_{Cb} \quad \text{and} \quad i_{Za} = i_{Ca}. \tag{18}$$

Therefore, substituting equations (18) into equations (17)

$$i_{Cb} \cdot cd = i_{Ca} \cdot Za. \tag{19}$$

Dividing (19) by the second of (17),

$$\frac{i_{Cb} \cdot cd}{i_{Cb} \cdot ef} = \frac{i_{Ca} \cdot Za}{i_{Ca} \cdot Ca}$$

or

$$\frac{cd}{ef} = \frac{Za}{Ca}$$

which is equation (16).

The Potentiometer

It is frequently necessary to measure electromotive forces, defined on page 457 as the potential difference maintained by a source of electrical energy when no current is being delivered. The use of an ordinary voltmeter involves a flow of current, which precludes the use of such an instrument when precise measurements of e.m.f. must be made. In such a case a ***potentiometer*** is used, an instrument which has many uses wherever precise electrical measurements are made. Its principal use is as a device for checking the accuracy of voltmeters, ammeters and wattmeters

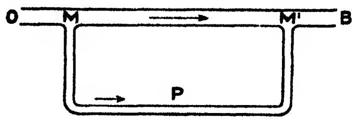


FIG. 382. DIVIDED FLOW IN WATER SYSTEM

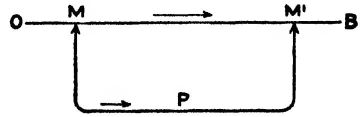


FIG. 383. DIVIDED FLOW OF ELECTRIC CURRENT

As an aid in comprehending the principles involved in the potentiometer an analogy is useful. In Figure 382 let *OB* represent a section of pipe carrying a current of water. Between the points *M* and *M'* a pressure difference exists, which, other things remaining unchanged, will increase with the distance between the points in question, the higher pressure being at *M*. If the pipe is tapped at *M* and *M'* and a branch pipe *MPM'* attached, a current flows as indicated, the flow of water is produced by the pressure difference.

Pressure difference in a water system is analogous to potential difference in an electric circuit; a current flows in a conductor only when a potential difference exists. If, in the preceding paragraph, the terms *conductor*, *electricity*, and *potential differencc* are substituted, respectively, for *pipe*, *water*, and *pressure difference*, and reference is made to Figure 383 instead of Figure 382, the same statements hold true without other modification. The paragraph referred to will then read: Let *OB* represent a section of a conductor carrying a current of electricity. Between the points *M* and *M'* a potential difference exists, which, other things remaining unchanged, will increase with the distance between the points in question, the higher potential being at *M*. If the conductor is tapped (contact made) at *M* and *M'* and a branch conductor *MPM'* attached, a current flows as indicated; the flow of electricity is produced by the potential difference.

Referring next to Figure 384, let a rotary pump be inserted in the branch *MPM'* and let it rotate in the direction indicated by the arrow. By the action of the pump, a pressure difference will be maintained between *N* and *P* with the higher pressure at *P*. It is easy to conceive of the pump as being driven at constant speed so as to maintain a steady pressure difference, and to imagine such strength of current in *OB* that the pump exactly balances the pressure difference between *M* and *M'*. In this situation no current flows through the branch *MPM'* because the tendency to flow in one direction is exactly neutralized by the tendency to flow in the opposite

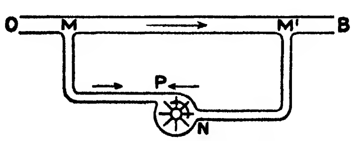


FIG. 384. WATER ANALOGY OF POTENTIOMETER

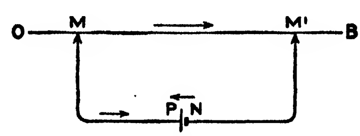


FIG. 385. PRINCIPLES OF POTENTIOMETER

direction. The absence of current could be shown by some kind of flow meter inserted in the branch *MPM'*.

With the same substitution of terms as before, the identical reasoning applies to Figure 385. Here a battery is represented as the source of potential difference and is analogous to the pump as the source of pressure difference.

In the diagram, Figure 386 the cell *W*, which may be an ordinary dry cell, causes a current to flow through the resistance *OB* in the direction indicated by the arrow. The result is a potential difference between any two points on *OB*, such as *M* and *M'*. With the current flowing as indicated, *M* is the higher potential and hence is positive with respect to *M'*.

Now consider the circuit *MEGM'* in which *E* represents any source of steady potential difference and *G* represents a galvanometer. The points *P* and *M* are brought to the same potential by connecting them with a conductor; then if the potential of *N* is as much below *P* as that of *M'* is below *M*, obviously *N* and *M'* must be at the same potential. The net result is that no current can flow in this circuit, and the galvanometer, which is merely a current-indicating instrument, shows no deflection. If, with the value of *E* remaining constant, the contact at *M'* were shifted a little to the right or left of *M'*, a current would immediately flow because the potential difference between *M* and *M'* would then be either greater or less than the fixed value of *E*. The galvanometer would no longer read zero but would indicate the current in *MEGM'*.

Therefore, with the aid of a galvanometer, two points *M* and *M'* can be found on a conductor *OB*, between which the potential difference is equal to that of a known potential *E*. It is immaterial where the two points on *OB* are located, provided that the resistance between them has the correct value, and *M* is connected to the terminals of like sign of *P* and *W*.

Having found two such points, and identified the potential difference between them by the use of a "standard cell," a different potential difference "tapped off" from two other points will be in proportion to the length of wire spanned. Hence, if a cell of unknown electromotive force be substituted for the standard cell, and the adjustment of *MM'* be repeated, the

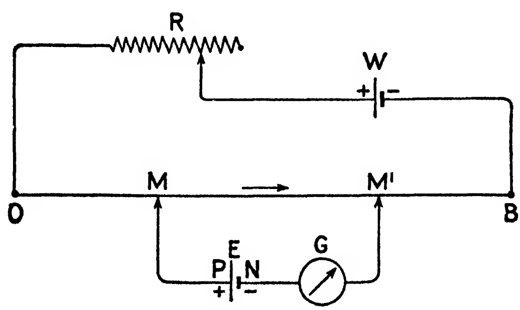


FIG. 386. POTENTIOMETER CIRCUIT

second value of MM' may be read, whence the unknown voltage may be immediately computed from the proportion stated above. This is the principle of the use of the potentiometer to compare two potential differences, a procedure that is obviously involved whenever it becomes necessary to standardize voltmeters.

If a potentiometer is used to make an accurate measurement of potential difference across a resistance of accurately known value, the current that is flowing can be determined to a corresponding accuracy with the aid of Ohm's law. This is the principle of the use of the potentiometer to compare two currents, a procedure that is obviously involved whenever it becomes necessary to standardize ammeters.

Kirchhoff's Rules

Many modern applications of electricity involve wiring which is too complex to be easily analyzed into combinations of simple circuits. Such complex circuits are called *electrical networks*. Their treatment requires a method somewhat more powerful than the simple form of Ohm's law. Gustav Kirchhoff (1824–1887) discovered two principles, so simple as to seem almost obvious, which are effective in treating electrical networks. They are called *Kirchhoff's rules*. The first will be called the *point rule*, the second the *loop rule*.

The point rule is merely a formal statement of the fact that at any point in a circuit in which current is flowing steadily the total current flowing *away* must equal the total current flowing *toward* that point. If this were not true, there would be an accumulation of electricity at some point, a condition which is contrary to the above assumption of a steady current. If the latter currents be termed positive and the former negative, then the point rule may be succinctly stated thus. *At any point in a conducting network, the algebraic sum of the currents is zero.* Algebraically,

$$\Sigma I = 0 \tag{20}$$

The loop rule is a formal statement of the fact that around any closed loop of a network the aggregate e.m.f. of all sources of electrical energy is equal to the sum of the differences of potential along the conductors connecting them. These differences of potential are, by Ohm's law, the products of resistances and corresponding currents. They are frequently referred to as "*IR* drops," and quite appropriately so. If oppositely directed e.m.f.'s be given opposite signs such that the aggregate e.m.f. is positive, then the loop rule may be stated thus. *Around any closed loop in*

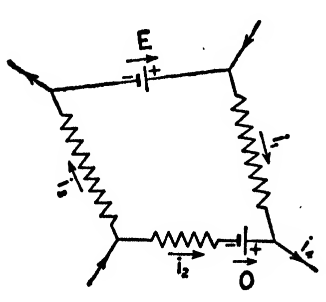


FIG. 387. SINGLE LOOP IN A NETWORK

a network the algebraic sum of all e.m.f.'s present in the loop, minus the sums of all the IR *drops around the loop, is equal to zero.* Algebraically,

$$\Sigma E - \Sigma IR = 0. \qquad (21)$$

Consider, for example, the right-hand junction of Figure 387. It consists of three conductors in which the indicated currents are flowing. By the point rule,

$$i_1 + i_2 - i_2 = 0. \qquad (22)$$

From the loop rule applied to Figure 387, another equation may be stated, namely,

$$(E - e) - i_1r_1 - i_2r_2 + i_3r_3 = 0. \qquad (23)$$

The problem involved in the solution of a network usually consists in computing the currents, given the voltages and resistances. For the case before us this appears at first impossible, since there are four unknowns, i_1, i_2, i_3, and i_4, and only two equations. However, Figure 387 shows only a single loop. Consideration of adjacent loops would give other equations involving the same unknowns (as well as others). When the entire network was taken into account, it would be found that there were in fact a larger number of equations than there were unknowns. Hence not only could the values of the unknown be found, but the extra equations could be used to furnish a check.

The signs assumed for the currents in the various branches of the network must be consistent in the two equations. If they are, and the actual resulting currents are in the directions assumed, all values of i will be positive. If any value of i comes out negative, that merely means that a wrong guess was initially made as to the direction of that particular current. It is common, indeed, to assume positive currents to be flowing clockwise, even in those branches in which an intuitive estimate indicates otherwise, and to rely entirely on the signs of the final results to give information on actual directions of the currents. This was not done, however, in the foregoing example and the corresponding equations.

Though the foregoing presentation of Kirchhoff's rules was limited to the case of steady currents, they may be modified to apply to alternating currents in circuits containing not only resistances and sources of alternating e.m.f.s, but other electrical devices, called *inductances* and *capacitances*, the nature of which will become evident in succeeding chapters.

Questions for Self-Examination

1. Tell what Ohm discovered about (*a*) resistivities of materials, (*b*) dependence of resistance on length and cross-section and temperature, and (*c*) aggregate resistance of wires in series.
2. Ohm was led to formulate his principles of electrical conduction by analogy with Fourier's statements about heat conduction. State both sets of principles in

parallel columns and describe how Ohm verified his electrical principles experimentally.

3. State Ohm's law. Define resistivity and temperature coefficient.
4. Interpret Ohm's law as applied to the whole circuit and to parts of a circuit, and discuss potential distribution in a simple circuit.
5. Prove by the use of Ohm's law that the aggregate of a group of resistances in series is the sum of the individual resistances.
6. Prove by the use of Ohm's law that the aggregate of a group of resistances in parallel is the reciprocal of the sum of the reciprocals of the individual resistances.
7. Describe the function of series and parallel resistors in voltmeters and ammeters.
8. Describe "Wheatstone's bridge."
9. State the principle of the potentiometer. To what uses is it ordinarily put?
10. State Kirchoff's two rules.

Problems on Chapter 39

1. Three resistances of *a*, *b*, and *c* ohms respectively are connected in series. A battery of electromotive force *E* and internal resistance *d* is connected in series with these. What is the potential difference across each resistance and the battery?

	a	*b*	*c*	*d*	*E*	PD_a	PD_b	PD_c	PD_e
(1)	100.	200.	300.	0.	600	100.	200.	300.	600.
(2)	95.	150.	250.	5.	250	48.	75.	130.	250.
(3)	4.	10.	20.	6.	8	.8	2.	4.	6.8
(4)	.1	.2	.4	.05	100	13.	27.	53.	93.
(5)	.02	.03	.05	.1	100	10.	15.	25.	50.
(6)	3.	4	5.	1.	2	.46	.62	.77	1.8
(7)	6.	7.	8.	1.	3	.82	.95	1.1	2.9
(8)	5.	8.	11.	2.	10	1.9	3.	4.2	9.2
(9)	20.	30.	40.	5.	10	2.1	3.2	4.2	9.4
(10)	15.	23.	31.	5.	10	2.	3.1	4.2	9.3
(11)	5.	7.	9.	1.	3	.68	.95	1.2	2.9

2. Three resistances of *a*, *b*, and *c* ohms respectively are connected in parallel. What is their joint resistance *r*?

	a	*b*	*c*	*r*
(1)	9	12	18	4.
(2)	40	100	200	25.
(3)	40	$66\frac{2}{3}$	100	20.
(4)	33	48	88	16.
(5)	20	30	60	10.
(6)	9	10	11	3.3
(7)	1	2	3	.55
(8)	7	10	12	3.1
(9)	12	18	20	5.3
(10)	5	12	15	2.9
(11)	3	4	5	1.3

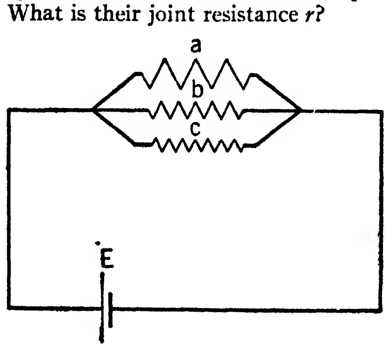


FIG. 388

3. The parallel resistances of the preceding problem are connected in a circuit as shown, containing leads whose total resistance is *e* and a battery of electromotive

force E and internal resistance d. Find the current I in the circuit and in each branch.

	e	d	E	I	I_a	I_b	I_c
(1)	5.	1.	10	1.	$\frac{4}{9}$	$\frac{1}{3}$	$\frac{2}{9}$
(2)	20.	5.	100	2.	$\frac{5}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
(3)	27.	3.	200	4.	2.	$\frac{6}{5}$	$\frac{4}{5}$
(4)	80.	4.	500	5.	$\frac{80}{33}$	$\frac{5}{3}$	$\frac{10}{11}$
(5)	2.	0.	120	10.	5.	$\frac{10}{3}$	$\frac{5}{3}$
(6)	1.5	.4	2	.38	.14	.13	.12
(7)	.3	.2	20	19.	10.	5.2	3.5
(8)	3.	5.	5	.45	.20	.14	.12
(9)	5.	5.	10	.65	.29	.19	.17
(10)	3.	3.	25	2.8	1.6	.67	.54
(11)	1.	1.	15	4.6	1.9	1.5	1.2

4. A low-range voltmeter may be quite conveniently used as an ammeter for the measurement of very small currents. If an instrument so used has a resistance of r ohms and shows a scale reading V, what current is flowing?

r	V	I
100	10.	.1
50	1.	.02
50	.4	.008
10	.2	.02

5. A galvanometer whose resistance is r ohms is shunted by a resistance s ohms. What proportion p of the total current passes through the galvanometer?

r	s	p
500	2	.004
100	10	.091
100	5	.048
90	10	.1

6. A wire l meters long and of diameter d millimeters has a resistance R ohms. What is the resistivity ρ of the material of which it is made? Can you determine, by reference to tables, what the material is?

l	d	R	ρ
10	1	.2	$1.6 \cdot 10^{-8}$
10	1	.9	$7.1 \cdot 10^{-8}$
10	1	1.4	$11. \cdot 10^{-8}$
10	1	4.	$31. \cdot 10^{-8}$

7. The temperature coefficient of resistance of platinum is .00366. The resistance of a platinum thermometer is r ohms at temperature t°C. Its resistance becomes R ohms at an unknown temperature. Find the unknown temperature T in degrees C.

r	t	R	T
300	0	2000	1550
300	0	1000	640
300	0	500	180
300	0	100	− 180

8. A resistance thermometer is found to have a resistance of R_0 ohms at 0°C. and a resistance of R_1 ohms at T°C.
   a. Find the temperature coefficient of resistance α of the metal of which the thermometer is made. Can you determine by reference to a table what the metal is?
   b. If the thermometer shows a resistance of R_2 ohms for a determination of the boiling point, find the temperature t at which this liquid boils. Knowing that this liquid is a chemical element, can you determine by reference to tables what it is?

R_0	R_1	T	R_2	α	t
10	14	98	12.4	.0041	59
25	40	97	20.	.0062	− 32
100	152	96	257.	.0054	290
200	223	95	422.	.0012	917

9. The tungsten filament of a light bulb has a resistance R_{20} at 20°C. Find the resistance R_t of the filament at a temperature t°C. Take the average temperature coefficient of resistance of tungsten to be 0.00554 per degree. Assuming the lamp to be operated on 120 volts, find the initial and final current.

R_{20}	t	R_t	Initial	Final
20.2	2412°	290	6	0.42
13.48	2412	190	9	0.62
10.11	2412	140	12	0.83
5.05	2412	70	24	1.67

10. Neglecting the loss due to radiation, what is the rise in temperature t per second in the following lengths of copper wire, each of mass m grams and resistance r ohms, when subjected to a potential difference of V volts? Take the mechanical equivalent of heat as 4183 joules per calorie and the specific heat of copper as .0939 Calorie per kilogram. Assume the resistance and specific heat to remain a constant as the temperature changes. What would be the effect of allowing for loss due to radiation?

m	r	V	t
9.5	.32	3	16
9.	.16	3	16
18.	.08	3	16
36.	.04	3	16

11. In the potentiometer circuit Figure 389, the slide wire ab is 200 centimeters long and its resistance is .5 ohm per centimeter. The batteries E_1 and E_2 have negligible internal resistance. $E_1 = 2$ volts, $E_2 = 1.6$ volts, $R_1 = 20$ ohms, resistance of the galvanometer $G = 100$ ohms. Find the magnitude and direction of the current in the galvanometer. Find the position of the contact C which gives zero current through the galvanometer. .00027 amp., 8 cm.

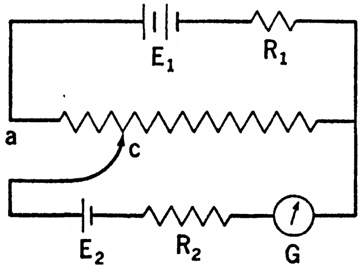


FIG. 389

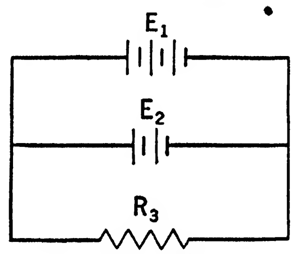


FIG. 390

12. In Figure 390 $E_1 = 6$ volts with an internal resistance of 1 ohm, $E_2 = 4$ volts with an internal resistance of .6 ohm.
    a. If $R_3 = 3$ ohms, find the current in R_3, E_1, and E_2. 1.4, 1.8, −4 amp.
    b. Find the current in R_3 when E_2 is removed. 1.5 amp.
    c. For what value of R_3 will the current in E_2 be zero? 2 ohms.

13. Apply Kirchoff's rules to the network of the accompanying figure to find the current in each branch. The voltage of the battery is *10*, with zero internal resistance; the resistances are as follows: $r_1 = 25$, $r_2 = 45$, $r_3 = 75$, $r_4 = 75$, $r_5 = 125$, $r_6 = 60$. .10, .063, .063, .038, .038, 0.

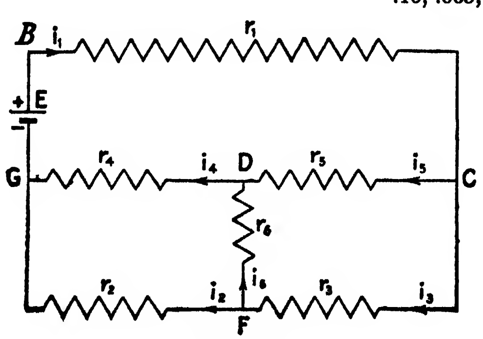


FIG. 391

CHAPTER 40

Capacitance

Condensers

One of the most common electrical devices is the accessory called a *condenser*. Radio apparatus, for example, consists in a surprisingly large proportion of condensers of many sizes both at the sending and the receiving ends. The very act of tuning a radio receiver usually consists of adjusting a condenser. A telephone message invariably passes through a dozen or more condensers. Automobile ignition and battery-charging circuits incorporate condensers. Large condensers are used on commercial electric systems to increase the efficiency of distribution. The variety of uses to which condensers are put make them much more common than many devices that are in fact better known. But in spite of the profusion of condensers, the average user of electrical appliances is not likely to be acquainted with them because they are usually small, are almost always hidden, and seldom require servicing.

Basically, condensers are devices for storing electrical energy, usually in infinitesimal amounts and for tiny intervals of time. There is little occasion to confuse them with storage batteries (page 459), partly because the two are put to very different uses and partly because their construction and action are so different. Structurally condensers are simply two adjacent electrical conductors separated by an insulating medium. Usually the conductors are very close together and of large area; for example, a representative telephone condenser consists of two sheets of tinfoil nearly a square yard in area, separated by a sheet of paraffined paper about a thousandth of an inch thick. The whole is rolled up into a small volume and sealed in a metal case, so that the area is not evident to a casual user.

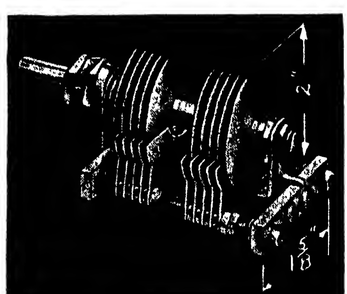


FIG. 392. A VARIABLE CONDENSER OF THE TYPE USED IN RADIO SETS (The National Co. Inc.)

Electrostatic Attraction

Some twenty-six centuries ago a philosopher named Thales, residing in an ancient city of Asia Minor called Miletus, was the first to record the fact that objects which were electrically charged would attract each other. At that time the only known way of generating electricity was by friction on amber, the Greek name of which was *electron*, whence our term *electricity*.[1] It took two thousand years, however, to discover that electricity could be generated by friction in other substances than amber. It was still later that the fact was discovered by Nicolo Cabeo in 1629 (24:194) that repulsion as well as attraction occurred between electrified bodies, that attraction characterized the action between unlike charges (positive and negative, as we term them now) and that repulsion characterized the action between like charges. Still later, Stephen Gray (1696–1736) discovered that electrical charges could travel from one place to another along wires (77:395). Today conduction along wires is almost taken for granted as the most prominent characteristic of electricity, and that electricity can exist at all in a stationary or *static* state is often not fully appreciated. This reverses the historical order of discovery of the principal phenomena of electricity.

How Condensers Work

These basic facts of electricity, namely, the forces between charges and the phenomenon of flow along conductors, account for the ability of a condenser to store electricity. Imagine two metal plates (Fig. 393) brought near together, the lower one connected to the earth. Let a positive charge be placed on the other plate. Under this attraction electrons (being negative) will flow from the earth, which has virtually an inexhaustible supply of electrons, to the lower plate. The greater the area of the plates, the larger the number of positive charges that are accommodated on the top plate and hence the larger the number of electrons that are attracted to the other. The closer together the plates, the easier it is to place more positives on the top plate against the repulsion of the positives already there. This is because the repulsion of the positives on the top plate for additional positives approaching it is more nearly balanced by the attraction of the negatives on the bottom plate. Hence both the areas and the mutual proximity of the two plates conspire to increase the quantity of electricity that can be placed on them. These two conditions go far toward determining the *capacitance* of a condenser.

[1] First used in print in 1646. (See 43:26 and 22, bk. II, chap. IV.)

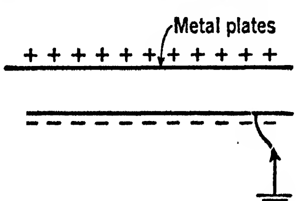


FIG. 393. CHARGING A CONDENSER

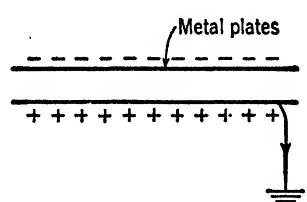


FIG. 394. CHARGING A CONDENSER IN REVERSE DIRECTION

If the upper plate had originally been charged negatively instead of positively, electrons would have been repelled from the lower plate to the earth instead of being drawn to it, thus leaving a corresponding number of unneutralized positives on the lower plate.[1] Thus, the condenser would be charged in the opposite direction to that previously described, but otherwise the state of affairs would be the same as before.

It is usually not necessary in practice to connect one plate of a condenser to the ground as described above. The ground connection is really necessary only when a condenser is to be charged by transfer of a quantity of static electricity. When it is to be charged by a battery or other bipolar source of electrical energy, no ground connection is required. The e.m.f. in any such source pulls electrons away from one plate of the condenser and piles them up on the other plate as indicated in Figure 395. The process continues until the potential difference between the terminals of the condenser becomes equal to the e.m.f. of the source, after which no further current can flow. Practically this seldom requires more than a tiny fraction of a second.

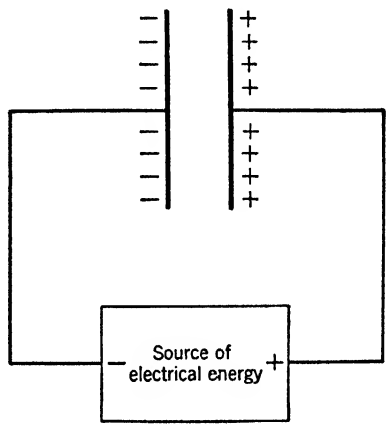


FIG. 395
CHARGING A CONDENSER BY BATTERY

The Evolution of the Condenser

Originally condensers were called "Leyden jars" because of a supposition that their action was first observed by an experimenter in Leyden, Holland. One form of condenser bears that name to this day (Fig. 396). It first consisted of nothing more than a bottle of water with a nail stuck through the cork. Its birth came about through a lucky misapprehension. The gradual loss of electrification that a charged body was observed to experience had been attributed to evaporation. From this arose the idea of putting the charged body into a container to reduce this "evaporation." What more natural than to use a bottle, and what more natural than to put water into the bottle as the object to be charged? A nail through the cork was a convenient way to establish electrical contact without opening the bottle. The amazing behavior of this simple device was first recorded by E. G. von Kleist, Dean of the Cathedral in Kammin, Pomerania, in a letter dated November 4, 1745. Holding the jar in one hand, he presented the nail to an electrical machine, then withdrew the jar and touched the nail with his other hand, whereupon, in his own words (105:81),

> I receive a shock which stuns my arms and shoulders.

[1] Refer in this connection to page 444, where the convention as to direction of flow of an electric current is clarified.

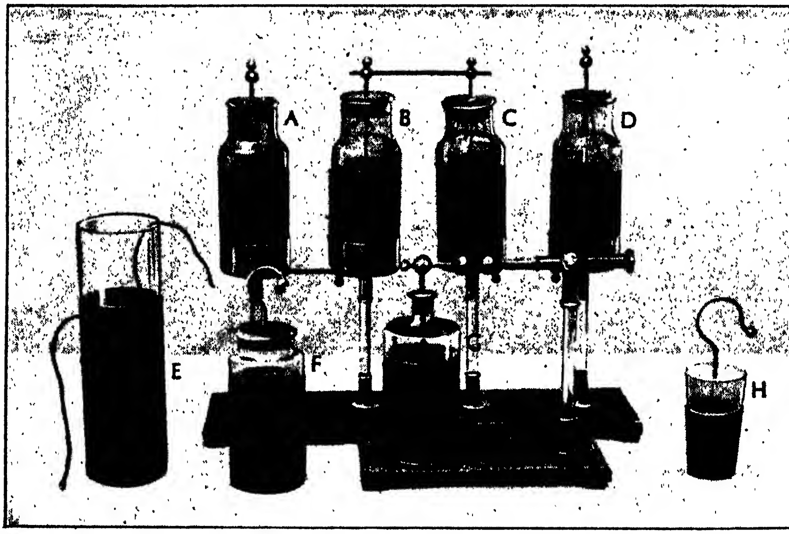


FIG. 396. A GROUP OF LEYDEN JARS
(From Williams: *Foundations of College Physics*. Courtesy of Ginn & Co.)

Taking electrical shocks was nothing new at this time. It had become, in fact, a widespread and favorite form of parlor pastime. But the shocks had always been taken directly from a machine and the victims never before experienced more than a mild degree of discomfort. Here apparently was a device capable of "condensing" a large amount of electricity into a small volume and delivering it with disconcerting violence. Peter van Musschenbroek, Professor at the University of Leyden, immediately made another condenser out of a bowl, the glass of which, be it noted, was very thin, and wrote that when he tried it, he felt himself struck in his arms, shoulders, and breast, so that he lost his breath and was two days recovering from the effects of the blow and the terror. He added, writing to a French scientific friend, that he would not take another such shock for the Kingdom of France. It was more than twenty years before this stunning shock was unequivocally identified with the same sensation produced by contact with an electric eel, a phenomenon that had been known from antiquity, but the idea that they were identical in nature occurred almost immediately and the method of demonstrating it was gradually worked out.

Von Kleist made another observation which was of significance. He said that none of these terrible shocks or bright sparks could be secured unless the jar was held in the hand and concluded:

> The human body therefore must contribute something to it.

This was a natural but erroneous conclusion. It seemed to be associated with electrical phenomena observed in the cat, with its counterpart in the action of human hair when vigorously brushed on a dry day (which just before this time had been recognized by Robert Boyle as electrical), and even more strikingly with the action of the electric eel. These effects were being lumped under the term "animal electricity," and von Kleist's observation was naturally but mistakenly interpreted in those terms. The relevant fact, which was shortly discovered, was that the outside of the jar must be covered with some conducting material. A sheet of metal would have been even more effective than moist human flesh. Similarly, the important point as to the interior was not the nature of the liquid contained in it, but only that the inside as well as the outside of the jar should be covered with a conducting layer. These facts came quickly to be realized and within a very few months the Leyden jar assumed the form characteristic of it to this day, covered inside and out with a thin coating of metal, the condenser delivering its charge through any circuit connecting the inside to the outside coating. Benjamin Franklin was soon to introduce the important modification of substituting a pane of glass, coated on both sides with tin-foil, for the jar. Naturally, the larger the coated jar or pane of glass, the greater the quantity of electricity that could be stored on it with the expenditure of a given amount of energy. Also as Musschenbroek's experience indicated, the thinner the glass the greater the quantity of electricity stored under the same condition.

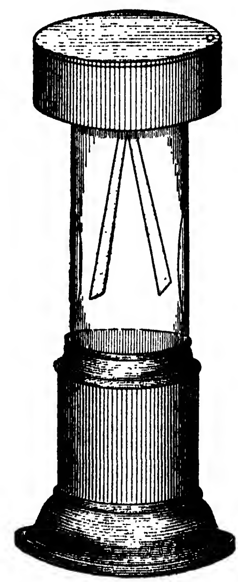


FIG. 397. THE ORIGINAL GOLD-LEAF ELECTROSCOPE (From *Philosophical Transactions*, *77*, 34 [1787].)

Forces Between Charges

The fact that attractions and repulsions exist between electrical charges has already been stated (page 457). Electrical repulsions were used in one of the earliest electrical instruments to be devised, the gold-leaf electroscope (Fig. 397). In this instrument the divergence of two leaves of gold foil bearing charges of like sign was taken as a rough way of comparing electrical potentials of charged bodies to which the electroscope was connected. It also provided a comparison of the quantities of electricity on the same body at two different times, since such quantities were necessarily in proportion to the potential. The electroscope did not, however, give any information on the relation between forces exerted by charged bodies on each other and their charges and separations. This, the most significant of the early measurements in electricity, was made in 1784–85 by Charles A. Coulomb (1736–1806). Coulomb demonstrated that electrostatic forces acted in accordance with an inverse-square law entirely analogous to that which Newton had postulated for gravitation and which

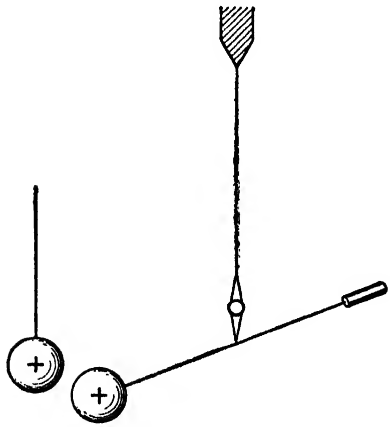


FIG. 398. THE PRINCIPLE OF COULOMB'S APPARATUS FOR MEASURING FORCES BETWEEN CHARGED BODIES

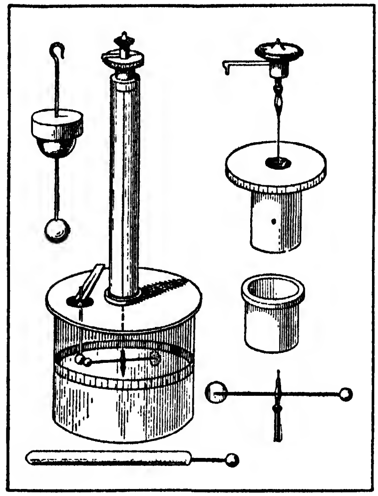


FIG. 399. COULOMB'S APPARATUS FOR MEASURING FORCES BETWEEN CHARGED BODIES
(From *Mémoires de l'Académie Royales des Sciences* [1785], p. 569.)

had just been directly verified by experiment by Cavendish. Coulomb's method was nearly like that of Cavendish, which was described in the section on mechanics (page 112). Electrified bodies were substituted for the gravitational masses of Cavendish's experiment. (See Figs. 398 and 399.) The result for electrical charges Coulomb termed *The Fundamental Law of Electricity* and phrased it substantially as follows (77:411):

> The force between two small spheres charged with electricity is in the inverse ratio of the squares of the distances between the two spheres.

The Idea of Quantity of Electricity

The precision with which Coulomb determined that the inverse-square law was true for electrical charges constituted a major contribution. But even so, a comparison with Cavendish's results will show that they left much to be desired. Cavendish had not only verified the inverse-square law, but he had also demonstrated that gravitational forces were directly proportional to the *products* of the masses, M and m, of the mutually attracting bodies. In addition he determined the constant of proportionality, G. In algebraic form the relation was

$$f = G\frac{Mm}{r^2}. \tag{1}$$

By contrast, Coulomb determined, for electrical forces, only that

$$f \propto \frac{1}{r^2}. \tag{2}$$

Since Coulomb's time this has been rounded out into a form completely analogous to equation (1), namely,

$$f = \frac{1}{4\pi k_0} \frac{Qq}{r^2} \tag{3}$$

where Q and q represent the quantities of electricity on the respective charged spheres and $1/4\,\pi k_0$ corresponds to G as the constant of proportionality. This is now called *Coulomb's law.* Coulomb could undoubtedly have put it into this complete form, including an evaluation of k_0 (now accorded the value $8.854 \cdot 10^{-12}$) if only a unit of electrical quantity had been in existence. But that was not definitively accomplished until eighty years later, at which time a unit of electrical quantity then established was appropriately given the name *coulomb.* This unit has already been defined (page 434).

The Dielectric Constant

One of the respects in which $\frac{1}{4\pi k_0}$ fails to be completely analogous to G is significant. Gravitational attraction seems to be entirely independent of the medium between the two bodies involved. The attraction between the sun and the moon, for example, is entirely unaffected by the intervention of the earth between them at the time of a lunar eclipse. Electrostatic forces, on the other hand, are profoundly affected by the intervening medium. The value of k_0, given above ($8.854 \cdot 10^{-12}$), applies only to a vacuum. For any other medium the constant of proportionality in equation (3) has a different value, and k_0 is replaced by another constant, say k. This constant is different for different substances. The ratio of its value to that of k_0 is known as the *dielectric constant,* K, of the medium in question. That is,

$$K = \frac{k}{k_0}. \tag{4}$$

As the name indicates, the higher the dielectric constant, the better the insulator. Approximate values of K for some common insulators are shown below.

air	1.011
water	81.07
porcelain	5.73
glass	7.
mica	6.
shellac	3.1

The abnormally high value for water is largely without practical significance, for it is next to impossible to remove impurities from water to such a degree as to attain, much less hold, the high tabular value of the dielectric constant. The theoretical significance of the abnormal dielectric constant of water is considerable, however.

Forces on Charged Bodies

The force experienced by a charged body between the plates of a condenser will be required in later chapters. The simplest case is that of a charge of one coulomb between plates whose distance apart s is so small in comparison with the dimensions of the plates that the force f on it is the same everywhere. This is the condition for a so-called *uniform field*. Then the work W necessary to move the unit charge from one plate to the other is

$$W = fs.$$

But by the second definition of the volt (page 457), W is simply the potential difference ΔV between the plates. Hence

$$\Delta V = fs$$

and therefore

$$f = \frac{\Delta V}{s}. \tag{5}$$

The term $\Delta V/s$ is called the electric field strength, and is measured in volts per meter of separation of the plates. Consequently, *the force in newtons on one coulomb in a uniform electric field is numerically equal to the electric field strength in volts per meter.* If the charge experiencing the force is q coulombs instead of one coulomb, then the force is simply q times as great as that on one coulomb, assuming that the charge is not great enough to destroy the condition of a substantially uniform electric field.

Capacitances of Condensers

It is now possible to define the unit of capacitance by which condensers are rated. *A condenser is said to have unit capacitance when a potential difference of one volt between its terminals is required to charge it with one coulomb of electricity.* This unit is the *farad*, so named in honor of Michael Faraday. The farad is too large a unit for most practical purposes and hence one millionth of it, called the *microfarad* (mf) is in common use. For condensers encountered in radio practice another submultiple is usually used, the millionth of a microfarad, the *micromicrofarad* (mmf).

Besides depending on area and separation of its plates, the capacitance of a condenser depends in addition on the dielectric constant of the medium between the plates. When a medium of dielectric constant k is substituted for air (or, more precisely, for a vacuum) of dielectric constant k_0, the capacitance of the condenser is changed in the ratio k/k_0. For the media in common use (glass, mica, paraffined paper, etc.) this always means in practice a considerable increase in capacitance when the new medium is substituted for air between the plates of a condenser.

If two or more condensers are connected in parallel, their group capacitance is the sum of the individual capacitances. Being in parallel the potential differences must all be the same. Hence for the group

$$C = \frac{Q}{V} = \frac{Q_1}{V} + \frac{Q_2}{V} + \ldots . = C_1 + C_2 + \ldots . \tag{6}$$

If two or more condensers are connected in series, their group capacitance is the reciprocal of the sum of the reciprocals of the individual capacitances. Being in series, and initially uncharged, the quantity in each condenser must be the same when a potential difference is applied to the outer terminals of the series. Hence, for the group

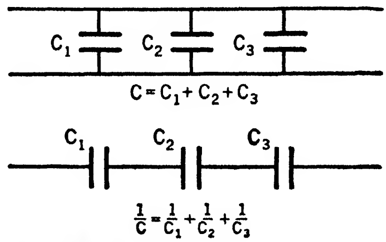


FIG. 400. CONDENSERS IN PARALLEL AND IN SERIES

$$\frac{1}{C} = \frac{V}{Q} = \frac{V_1}{Q} + \frac{V_2}{Q} + = \ldots . \frac{1}{C_1} + \frac{1}{C_2} + \ldots . \tag{7}$$

Electrostatic Induction

The flow of electricity in charging a condenser involves a phenomenon commonly termed *electrostatic induction* [1] which has played a prominent part in the development of electrical theory. Electrostatic induction is simply a general name to cover the redistribution of static electricity on a conductor due to any changes in the charges of neighboring bodies. Gray [2] had examined the phenomenon rather closely, remarking that

> the *Electric Virtue* may be carried from the Tube, without touching the Line of Communication, by only being held near it.

[1] The qualifying term *electrostatic* is introduced to distinguish this effect from *electromagnetic* induction, a totally different phenomenon discovered in 1831 by Faraday and Henry. (See Chapter 42.) It is a misfortune, rather uncharacteristic of scientific terminology, that the same noun should have been selected to denominate two such diverse phenomena. However, so cumbersome is the entire term "electrostatic induction" that only the noun will be used except under circumstances where confusion might result.

[2] *Philosophical Transactions* (abridged), *6*, part II, 26 (1731).

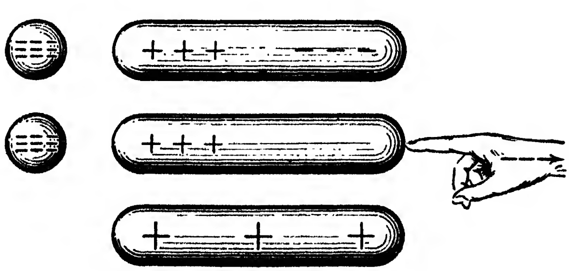


FIG. 401. CHARGING BY ELECTROSTATIC INDUCTION

The nature of this action, though not comprehended at the time, is now quite clear. If a body charged, say, negatively, is brought near the end of an insulated conductor, negative will be repelled to the far end of the conductor, leaving the end near the charged body positive. This was what Gray observed. But much more can be made of it. If the far end be touched, the negative will escape, leaving a surplus of positive on the conductor. Another conductor can be charged in the same way, and as many more as are desired, all without reducing the quantity of electricity on the neighboring charged body. A similar result, *mutates mutandae*, can be secured by a positively charged body. Two generations after Gray identified this effect, Volta used it in his *electrophorus*. This embodied the working principle of all the electrostatic machines such as may be found in any physics laboratory even today. They were successors to frictional machines which were the sole mechanical sources of electricity up to and including the day of Benjamin Franklin.

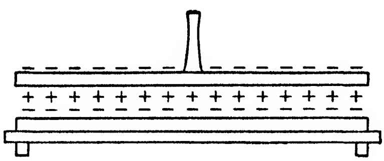


FIG. 402. THE ACTION OF THE ELECTROPHORUS

The Condenser and Lightning

It was the condenser which provided the final link in the chain of evidence that lightning was an electrical phenomenon. This was the most sensational of several major contributions made by Benjamin Franklin to the young science of electricity. Franklin was by no means the first to suspect that lightning was electrical. He was not even the first to capture lightning from the upper atmosphere and put it through its paces in an attempt to furnish conclusive demonstration that it was electrical. But he *was* the first to bottle it in a condenser and to show by this means that it was identical with the electricity similarly bottled from a frictional machine.

In 1746, at the age of forty, Franklin happened to see a Doctor Spencer, from Scotland, perform some electrical experiments at Boston. That was the beginning of an all-absorbing interest in the subject. Within two years, Franklin had become so engrossed in this study that he sold his printing house, newspaper, and famous *Almanack*, with the view of retiring from business at the age of forty-two, to devote all his time to his electrical experiments. His principal interest then was in lightning prevention. In a long letter of 1750, which he submitted through a friend for the consideration of the Royal Society, he described his idea, based on experiments of 1749. He leads up to the proposal in this way (44:*5*:231 ff.):

> Points have a property, by which they *draw on* as well as *throw off* the electrical fluid, at greater distances than blunt bodies can.... Thus a pin held by the head, and the point presented to an electrified body, will draw

off its atmosphere at a foot distance; where, if the head were presented instead of the point no such effect would follow....

If these things are so, may not the knowledge of this power of points be of use to mankind in preserving houses, churches, ships etc. from the stroke of lightning... Would not pointed rods probably draw the electrical fire silently out of a cloud before it came nigh enough to strike, and thereby secure us from that most sudden and terrible mischief?

But Franklin's contribution to lightning prevention, important though it was, he apparently regarded somewhat in the light of a by-product. In his notebook of the same year as his letter to the Royal Society is found the following comparison of lightning and electricity (25:129):

Electrical fluid agrees with lightning in these particulars: (1) Giving light; (2) colour of the light; (3) crooked direction; (4) swift motion; (5) being conducted by metals; (6) crack or noise in exploding; (7) subsisting on water or ice; (8) rending bodies it passes through; (9) destroying animals; (10) melting metals; (11) firing inflammable substances; (12) sulphurous smell.

But there remained one point, whether of similarity or of difference, to be settled experimentally. Would lightning be attracted and drawn off by points like the electric charge in his jars? Two French observers tried it, using pointed iron rods forty and ninety-nine feet high respectively and found that it would.

But Franklin was not satisfied. The lengths of the rods used by the Frenchmen were not great enough to convince him that the electric disturbances at the bottom really originated in the clouds above. Nor had the electricity been caught in a Leyden jar and its identity with frictional electricity established. He puzzled over how to cover these points. There came to him the idea of flying a kite into a thunder cloud. He had one made with a number of metal points projecting from it, communicating with the cord by which the kite was to be flown.

The rest of the story is a part of the education of every school child. At the appropriate time the kite was raised and disappeared into the cloud. As soon as the string had been rendered conducting by becoming wet, its electrified state became evident through the sudden standing out of the loose fibers of the string. From his sheltered position under a shed, holding the string by a dry *silk* cord, knotted to it for the sake of safety, Franklin then performed the crucial operation, no doubt wondering whether it might not be his last act. He presented his knuckle to a key which he had tied to the end of the string. A spark jumped from the key with the characteristic crackling sound which he had heard hundreds of times when taking the discharge of a frictional machine. It was but the work of a moment to lower the key to one of the Leyden jars brought for the purpose, to charge the jar, and to test it in the same way. The great discovery — the most sensational in the annals of electricity and the first major experiment with a condenser — was complete.

Questions for Self-Examination

1. Describe a typical condenser and tell some of the uses of condensers.
2. Describe how it is possible for condensers to retain an electrical charge sufficiently large to be observable.
3. Tell about the evolution of the condenser (Kleist, Musschenbroek, Franklin).
4. State Coulomb's law and tell how it is affected by the dielectric constant of the medium.
5. Define electric field strength and state its unit.
6. Name and define the unit of capacitance.
7. Tell how to calculate the aggregate capacitance of a group of condensers in series and in parallel and show why this should be so.
8. Describe the action of the electrophorus.
9. Recount the sequence of events leading to the final identification of lightning as a discharge of static electricity.

Problems on Chapter 40

1. Coulomb found that when his device (Figs. 398 and 399) was electrically charged, the moving needle carrying the charged spheres turned through 36° in consequence of the repulsion. By twisting the suspending fiber through 126° he brought the needle back to a deflection of 18°. Show how these data indicate an inverse-square law of repulsion.
2. Two small objects carrying positive charges of Q and q microcoulombs respectively are separated a distance r centimeters in air. Find the force f_1 of repulsion in newtons. Find the force f_2 when the objects are immersed in oil of relative dielectric constant *3*.

Q	q	r	f_1	f_2
$3\cdot10^{-2}$	$2\cdot10^{-3}$	5	$2\cdot10^{-4}$	$.7\cdot10^{-4}$
6	4	5	8	2.8
6	4	10	2	.7
9	6	10	5	1.6

3. Two identical conducting spheres, each of mass milligrams, are supported by two very light conducting threads of length l centimeters from two points separated by a distance s centimeters on a horizontal metal bar. The whole system is given an electrostatic charge, the charges on the balls causing them to repel each other so that their supporting threads make an angle θ with the vertical. Find the charge q on each ball in microcoulombs.

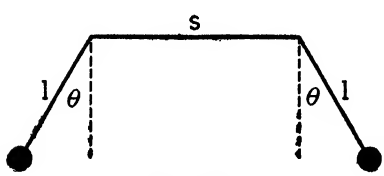


Fig. 403

m	s	l	θ	q
50	1	4	30°	$9\cdot10^{-3}$ microcoulombs
50	2	8	30°	$18\cdot10^{-3}$
50	2	8	10°	$4.7\cdot10^{-5}$
50	2	8	10°	$2.3\cdot10^{-5}$

4. The hydrogen atom consists of a relatively stationary nucleus with a positive charge of $1.600\cdot10^{-19}$ coulombs and an electron of equal but opposite charge ro-

tating about the nucleus in an approximately circular path of radius $5.27 \cdot 10^{-11}$ meter. If the mass of the electron is $9.11 \cdot 10^{-31}$ kgm., what must be the angular speed in revolutions per second in order that the centrifugal force will balance the electrostatic force of attraction? Ans. $7 \cdot 10^{15}$ rev./sec.

5. An electric "dipole" (the electrostatic equivalent of an ideal small magnet) consists of two equal charges of opposite sign separated by a distance so small that the mutual attraction holds the charges tenaciously together. Calculate the field strength in E, in newtons per coulomb both in direction and magnitude, at a distance d centimeters from the center of a dipole made up of two charges q and $-q$ microcoulombs separated by a distance of l centimeters which is very small in comparison with d;
   a. Along the line joining the charges,
   b. Perpendicular to the line joining the charges.

q	l	d	E_a	E_b
1	1	10	$200 \cdot 10^3$	$100 \cdot 10^3$
1	1	100	0.2	0.1
2	1	10	450.	220.
2	2	100	0.9	0.4

6. A condenser of capacity C microfarads is charged to a potential difference of E volts. Its two terminals are then connected to those of another condenser, uncharged, of capacity c microfarads. What is the resulting potential difference e in volts?

C	E	c	e
2	100	1	67.
3	100	2	60.
4	100	3	57.
5	100	4	56.

7. Three condensers of capacity C_1, C_2, and C_3 respectively are connected in series. A potential difference of E volts is applied to the outer terminals of the set. What are the potential differences e_1, e_2, and e_3 across the respective condensers?

C_1	C_2	C_3	E	e_1	e_2	e_3
1	2	3	100	55.	27.	18.
2	3	4	100	46.	31.	23.
3	4	5	100	43.	32.	26.
4	5	6	100	41.	32.	27.

8. What would be the aggregate capacitance of 12 one-microfarad condensers connected in three series groups of four each, the three groups connected in parallel with each other? If each condenser will stand 1000 volts, how many coulombs will it hold? How many would the entire group of condensers, as connected, hold? .75 m.f., .001 coulomb, .003 coulomb.

CHAPTER 41

Magnetism

The Significance of the Study of Magnetism

The attractive force between a magnet and a piece of iron is today a phenomenon so familiar to us that we often fail to realize that we really know very little more than the ancients as to how this attraction occurs. When we observe a piece of soft iron clinging to a steel bar, ordinarily bent into a shape somewhat like a horseshoe, we say that the bar is magnetized and think no more about it. In regarding this attraction thus, we do not measure up to the intellectual acuteness of our ancestors of twenty-five hundred years ago who, when they observed a similar phenomenon, thought a great deal about it.

There is, nevertheless, far more reason why we of this generation should give serious attention to the phenomenon of magnetism than did the Greek philosophers. The entire technology of an electrical age centers in it. Without it we should still be living as humanity did hundreds of years ago. Moreover, in spite of the vaunted prowess of physical science, we have had scant success in accounting for magnetic phenomena. A considerable portion of this subject we comprehend almost as little as did the earliest observers. Yet there is good reason for an opinion that the very pattern of scientific thought received its first impulse from the study of magnetism and it is certainly true that the magnetic compass has had more influence on world affairs than any other single device that could be mentioned.

The first philosopher known to give serious thought to the nature of magnetic forces was Thales of Miletus, the same man who made the first study of frictional electricity. He must have been a most extraordinary man. He lived in the seventh and sixth centuries B.C. All his writings (if he ever set any down) have been lost, and we know him only through his commentators. Yet writers of every age agree in regarding him as the father of philosophy. Thales regarded ability to move about or to cause other objects to move about as *prima facie* evidence of the possession of a soul. Hence he thought, according to Aristotle's account, "that the magnet has a soul because it causes movement to iron." This was a new and radical idea, much more radical than it appears at first sight to us. Until the time of Thales (and indeed for long after), magnetic action had been classified as magic. About such matters, the uninitiated might not even

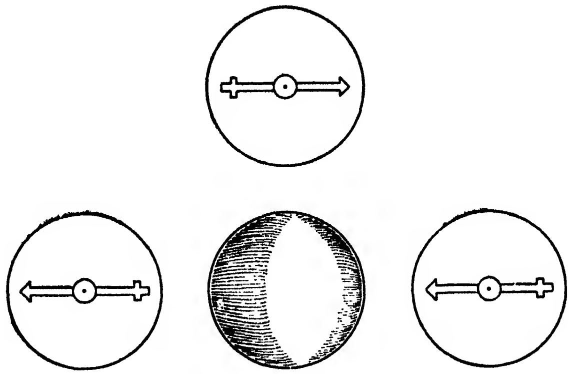


FIG. 404. PEREGRINE'S IDEA OF THE POSITIONS OF COMPASS NEEDLES AROUND A MAGNET
(As represented in Cabeo's *Philosophia Magnetica*, 1629.)

think, for fear of incurring the anger of the gods. But, if the magnet possessed a soul, as did men and animals, that was another matter. It could not be impious to wonder about your own possessions or those of others. Hence, magnetism became one of the first natural phenomena to be subject to investigation. As such it gave an early, and perhaps the first, impetus to scientific thought, which was not to come into full maturity for twenty-three hundred years.

Magnetic Polarity

The magnetic phenomenon that seems most strongly to have impressed the ancient commentators was magnetic attraction. Observations on the correlative phenomenon, magnetic repulsion, seem to have been very infrequent. The Roman poet Lucretius (99–55 B.C.) was probably the first to record this observation (72:269), but its significance was not realized for fourteen hundred years. Today every schoolboy knows that a magnet possesses two ends with characteristics which are, in certain respects, magnetically opposite to each other. These ends are usually called *poles*, a term introduced by William Gilbert in 1600. It is common knowledge now that unlike poles show a mutual attraction as do unlike electrical charges, while like poles show a mutual repulsion. But this distinction between the behavior of different pairs of poles is of fairly recent origin. The first intimation of it was in the *Opus Minus* of Roger Bacon (12:383–84), written near the middle of the thirteenth century. But the first really clear statement of the principle of magnetic polarity was contained in a remarkable letter written in 1269 by Peter Peregrine, a close friend of Bacon's (96:5). He told how to verify it by breaking a piece of lodestone and noting how a new pair of opposite poles appeared at the fractured ends. He also described correctly the orientations of a compass needle in various positions around a spherical piece of lodestone (Fig. 404). It was perhaps natural

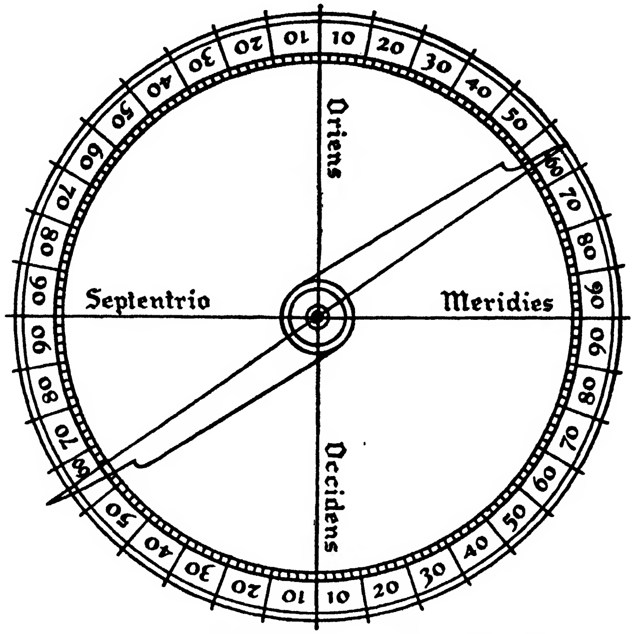


FIG. 405. THE FIRST KNOWN ILLUSTRATION OF A PIVOTED COMPASS (1269)
(From Peregrine's *Epistola de Magnete*, 1269.)

that he should have been the first to do this, since this same letter also contained the first known illustration of a pivoted compass (Fig. 405).

Magnetic Fields

The behavior of a compass needle in the vicinity of magnets indicates that these regions of space are in an abnormal condition. The state of affairs is perhaps most readily visualized by studying the arrangement of iron filings around a magnet (Fig. 406). The pattern of these filings outlines what has come to be called the *magnetic field* of the acting magnet. The concentration of the lines along which filings arrange themselves is a rough measure of the relative strengths of different portions of the magnetic field. Experience shows that the end or "pole" of another magnet, small enough so that it produces negligible disturbances to the field and is hence usable for test purposes, experiences a force along the lines, of magnitude proportional to the concentration of the lines.

To the early workers in magnetism these "lines of force" were very real. They were formalized into the basic principles of magnetism, the strength of a magnetic field being stated in terms of the number of lines per unit of area normal to them. Michael Faraday, in his work on magnetism, went even farther, attributing a corporeal reality to "lines of force" and endowing them with physical properties, notably elasticity. The fact is, of

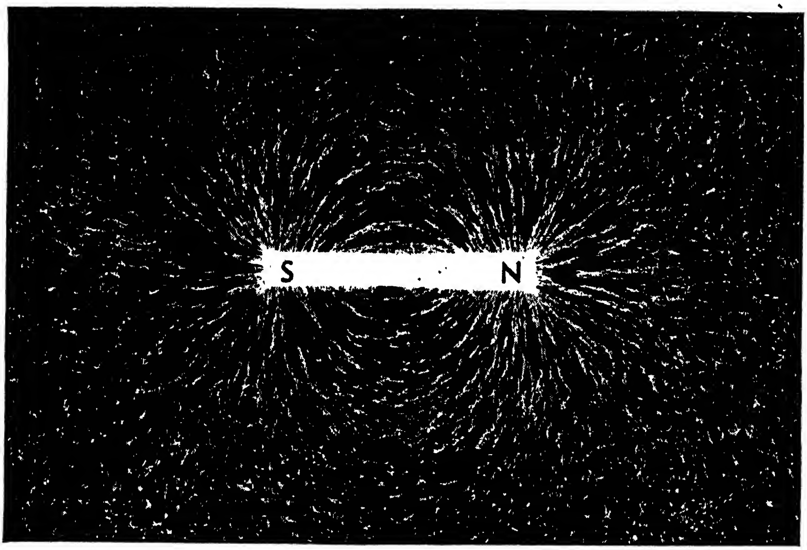


FIG. 406. IRON FILINGS AROUND A MAGNET
(From Williams: *Foundations of College Physics.* Courtesy of Ginn & Co.)

course, that "lines of force," however convenient they may be as aids to visualization, are mere figments of imagination, and as such constitute an inadequate foundation for magnetic theory. It is better to build the central concepts in this, as in any field, out of phenomena that can be observed, operations that can actually be performed in the laboratory.

Magnetic Field Strength

Iron filings, sprinkled on a sheet in the neighborhood of a helically wound wire carrying a current, arrange themselves in a pattern similar to that around a bar magnet, indicating that an electric current may be so disposed as to be the equivalent of a magnet (Fig. 407). A test magnet placed in the field of the helix as described above will experience a similar force. The

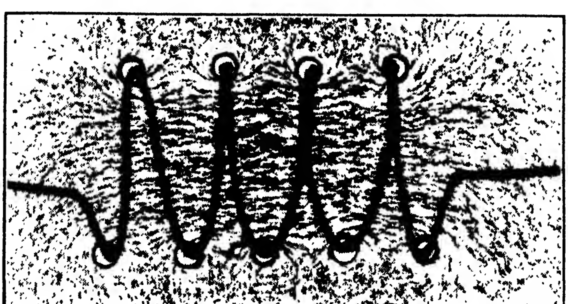


FIG. 407. IRON FILINGS AROUND A SOLENOID
(Sawders.)

magnitude of the force on the end of a given test magnet in either case is taken to be proportional to the strength of the field at the point under scrutiny. But just as the original bar magnet could be replaced by an electric current, so the test magnet may be replaced by another current. The force on a conductor carrying this second current becomes, under prescribed conditions, the actual measure of the strength of the magnetic field produced by the first.

But the force between adjacent currents has already been encountered. It was involved in the definition of the ampere (page 433). From that definition it follows that when a long straight wire carries a current of one ampere, the force on each meter of a parallel wire one meter away is 2×10^{-7} newtons. This is utilized as the basis for the definition of the unit of *magnetic flux density.* The *flux density* one meter away from an infinitely long straight wire carrying one ampere is thus defined as 2×10^{-7} units. Since the force on 1 meter length of wire carrying one ampere in a magnetic field of strength 2×10^{-7} units is 2×10^{-7} newtons, it follows that *when a wire carrying a current of one ampere, perpendicularly to a uniform magnetic field, experiences a force of 1 newton on each meter of its length, the magnetic flux density is 1 unit.* The principal name by which this unit is known is the awkward one of *weber per square meter.* From the foregoing definition of the unit of magnetic flux density the force f in newtons on a wire l meters long perpendicular to a magnetic flux density B webers/m² is

$$f = IlB. \tag{1}$$

From the law of Biot and Savart (page 432) the magnetic flux density B at a distance r meters from a long straight wire carrying a current I amperes is

$$B = 2 \times 10^{-7} \frac{I}{r} \text{ webers/m}^2. \tag{2}$$

As was implied on page 446, however, much stronger magnetic fields can be produced by winding the wire into a coil than are produced by the same current when traversing a straight wire. Though the calculation of the field strengths produced within coils is usually an undertaking of some complexity, it is very simple for two particular forms of coil. One is a long thin coil, the so-called *solenoid*, such as could be produced by winding wire in a single layer on a broomstick. The other is a flat coil, whose windings are substantially all in a single circle, such as could be produced by winding wire around the edge of a thin disk. Inside of a solenoid having n turns of wire per meter of length carrying I amperes, the flux density B at the midpoint of the length is simply

$$B = \mu_0 nI \text{ webers/m}^2. \tag{3}$$

The factor μ_0 has the value $4\pi \times 10^{-7}$ (1.257×10^{-6}). It is termed the *permeability of space.* If the region inside the solenoid is occupied by some material substance, μ_0 must be multiplied by another constant μ characteristic of that substance. Values of μ, termed *relative permeability*, have been measured and tabulated for many substances. (See Appendix.)

If the length of the solenoid is ten times the radius, equation (3) is correct within one half per cent. For "longer" solenoids the accuracy is still higher. Also, the equation applies to points at considerable distances from the midpoint as long as one remains inside of the solenoid at a discreet distance from the ends.

From equation (3) $B/\mu_0 = nI$, the product of the number of turns of wire per meter by the current in amperes. This quantity, which is independent of the medium surrounding the coil, is called the *magnetic field strength* (or *magnetizing force* by engineers), and is represented by the symbol H, whence

$$H = nI. \tag{4}$$

In describing magnetic fields, therefore, two vector quantities B and H are commonly used. If one is given, the other can always be found from the equation

$$H = \frac{B}{\mu_0} \text{ in air,} \tag{5}$$

or

$$H = \frac{B}{\mu\mu_0} \text{ in any medium,} \tag{6}$$

μ being the relative permeability of the medium as described above. As suggested by equation (4) the field strength H in a solenoid is expressed in *ampere turns per meter*. The term applies not merely to the field within a long solenoid, but to any magnetic field strength, no matter how produced.

The second type of coil for which it is easy to calculate the flux density is that of a flat coil. If the coil has a radius r meters, and bears N turns of wire carrying a current I amperes, the field strength at the center is

$$H = \frac{B}{\mu_0} = \frac{NI}{2r} \text{ amp turns/m,}$$

or

$$B = \mu_0 \frac{NI}{2r} \text{ webers/m}^2. \tag{7}$$

Equations (3) and (7) are readily derivable from the law of Biot and Savart (page 432) with the aid of some of the simpler processes of calculus, but will not be derived here.

Magnetic Flux

The fact was pointed out on page 496 that concentrations of lines of iron filings in a magnetic field provided a rough measure of the relative strengths of different portions of that field. In Figure 407, for example, the strength of the magnetic field inside of the solenoid is greater than outside, as indicated by the greater number of lines per unit area (perpendicular to the lines). For some purposes, however, there would be significance to the *total number of lines* within the solenoid. This would be the product of the number per unit area by the area itself. In other words, a new concept is being introduced involving the *product of magnetic flux density by the area.* This is termed the *magnetic flux.* Algebraically, magnetic flux ϕ is related to B and area S (perpendicular to the direction of the field) by the equation

$$\phi = BS. \tag{8}$$

B is, of course, in webers/m^2 and S in m^2. The unit of magnetic flux is termed the *weber.*

Magnetic Moment

The strength of a magnetic field is sometimes defined as the force on a magnetic pole of unit strength. If magnetic poles could be located precisely enough to make this definition at all useful, the next step would be to define the torque on a magnet suspended crosswise of a magnetic field as the product of the force on each end by the distance between the poles. This would then be

torque = flux density × (pole strength) × (distance between poles).

But neither pole strength nor distance between poles is a determinable quantity. Nevertheless, the product is readily determinable. For example, the torque necessary to hold a compass needle at right angles to the magnetic north-south position could be readily measured in spite of the fact that neither the strengths nor the positions of its individual poles could be determined. Consequently, it is customary to focus attention upon the determinable product of these two quantities rather than on the undeterminable quantities themselves. This product (pole strength × distance between poles) is called the *magnetic moment* M of a magnet and is defined by the above equation. That is,

$$L = BM \quad \text{or} \quad M = \frac{L}{B}, \tag{9}$$

where L is the torque in newton-meters required to hold the magnet crosswise of a magnetic field of B webers/m^2. If the magnet is to be displaced from the direction of the field merely by θ degrees instead of by 90°, then equation (9) takes the form

$$M = \frac{L}{B \sin \theta}. \tag{10}$$

Terrestrial Magnetism

That the earth in itself is a huge magnet, in that it possesses a magnetic field, observable in direction and measurable in magnitude, has been realized ever since William Gilbert, sometimes called "the father of magnetism," published his treatise *De Magnete* in 1600 (41A). The realization did not dawn with the beginnings of the use of the magnetic compass for navigation some five centuries earlier,[1] as might be reasonable to suppose. At that time the action of the compass was commonly attributed to mountains or islands of magnetic material, variously located in accordance with the imaginations of different authors (97:*6*, 124 and 355).

That the magnetic compass does not point exactly to the geographic north has been known in a general way almost ever since the compass was devised. The angle between true north and the so-called magnetic north is termed *magnetic declination*. The systematic study of magnetic declination was apparently given its initial momentum through some observations made by Christopher Columbus in 1492 in an early stage of his first voyage to the East Indies (as he thought) by sailing west. In a letter to the King and Queen of Spain, Columbus thus describes his discovery (78:127):

> When I sailed from Spain to the West Indies, I found that as soon as I had passed 100 leagues west of the Azores . . . the needle of the compass, which hitherto had turned toward the northeast, turned a full quarter of the wind to the northwest, and this took place from the time when we reached that line.

By degrees the declination of the compass was established in various parts of the world and charted. The first well-authenticated table was drawn up by Robert Norman in 1581, nearly a century after Columbus' observation of declination (92). Today extensive charts are available indicating the declination in all parts of the world. They consist of lines drawn through all places of equal declination, these being called *isogonic* lines. Such charts have to be redrawn from time to time, to provide for the changes which are continually occurring.

The point that apparently impressed Columbus the most was not the existence of magnetic declination, which he and doubtless others had known about before, but rather the discovery of a region of no declination, that is, a line along which the compass indicated a true geographic north; for obviously in changing from an easterly to a westerly declination, the needle would have to pass through a position of zero declination, that is, point true north. He associated this region of no declination with certain climatic changes, with certain imagined astronomical perturbations, and with the circumstances creating the famous accumulation of sea vegetation at that place, known as the Sargasso Sea, and felt that he had discovered a region for some reason unique on the surface of the earth.

Pope Alexander VI was possibly influenced by similar considerations when he made this line of zero declination the political division between the two portions of the earth in which new territory, when discovered, should be allocated respectively to Portugal and Spain, the two world powers of that era.

[1] H. Winters, *Mariner's Mirror*, *23*, 95 (1937).

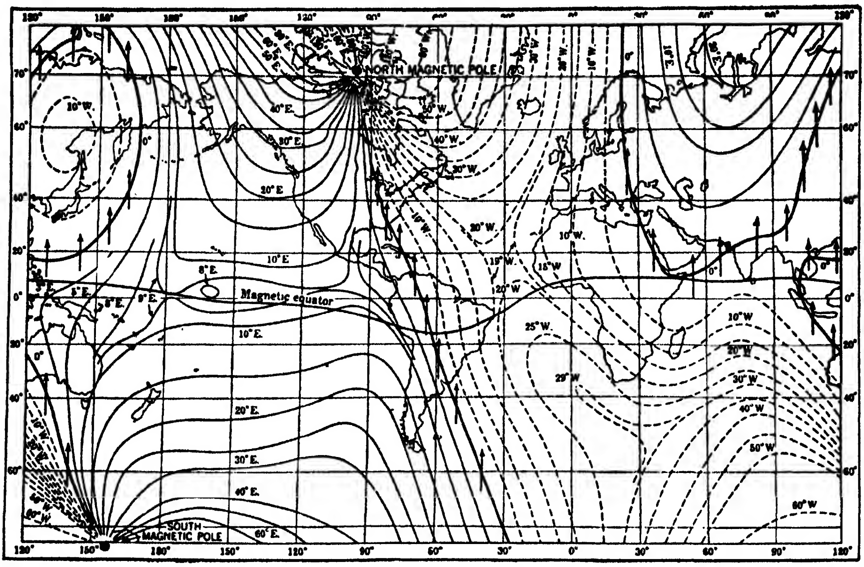


Fig. 408. Isogonic Lines

Vasco da Gama, under Portuguese auspices, following the compass on another great adventure, checkmated Columbus' discovery of America by finding a sea route to the East Indies around the Cape of Good Hope, thus meeting with success, where, in the contemporary view, Columbus had failed. Spain, green with envy, commissioned Magellan to make another attempt to reach the same place by traveling west to bring the East Indies within her own claim. For obviously, though the islands had already been reached by sailing east from the line of zero declination and, hence, had been claimed by Portugal, if they could be attained by traveling west from the same line, the claim of Spain to them under the provisions of Alexander's arrogant decree would be just as good. The really important outcome of this competition was not so much the desired shift in economic advantage of one nation over another, as the new scientific advance which resulted from Magellan's circumnavigation of the globe in 1520–22, from which all nations profited.

Who then shall measure the human achievements stimulated by the study of the earth's magnetism? Vast economic shifts; political rivalries leading to catastrophic wars; unprecedented mass movements in population; the mingling of races with consequent impacts of different cultures and religions; scientific discovery leading to the establishment of new standards of living: In a word, a world which had been palsied for a dozen centuries began to move forward with the might and majesty of an entirely new racial enterprise. And back of it all lay a slender bit of magnetized iron, poised unstably on a pivot, yet always looking to the north.

Magnetic Dip and Field Strength

But discoveries in the realm of magnetism had only begun. In this, as in most other fields, general recognition that the activity was of value to society acted to accelerate the acquirement of knowledge, and successive major steps in advance were, from this time on, to be separated only by decades rather than by centuries as heretofore.

In 1544, Georg Hartmann, a vicar in Nuremberg, observed that a steel needle, balanced to remain horizontal when unmagnetized, was thrown out of balance upon magnetization, the north end dipping toward the earth. He communicated his discovery to Count Albert of Prussia, but the letter failed to come to public notice for nearly three hundred years. The effect was rediscovered in 1576 by Robert Norman, the same man who in 1581 published the first charts of magnetic declination. The magnetic dip of a needle is very great. In fact, the earth's magnetic field in this latitude comes nearer to being vertical than it does horizontal. Norman constructed a needle so that it would rotate in a vertical instead of a horizontal plane, and found the dip in London to be 71° 50′ (93). Like magnetic declination, magnetic *dip* or inclination has now been determined for nearly all portions of the earth, and charted by *isoclinic* lines. Again like magnetic declination, dip changes with the passage of time. The dip in London, which was 71° 50′ in Norman's day, reached a maximum of 74° 42′ one hundred and fifty years later and has since been decreasing.

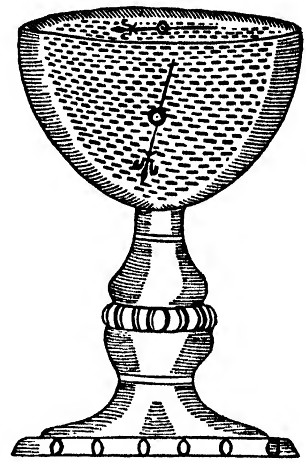


FIG. 409. NORMAN'S FIRST DIPPING NEEDLE
(From *The New Attractive* by Robert Norman, 1590.)

Declination and dip represent two aspects of *direction* of the earth's magnetic field. In some respects, knowledge about the *strength* of the earth's field is of even greater significance. In 1776, just two hundred years after Norman's observation of dip, and at a time when American men of science were concerned with other than scientific matters, Jean Charles Borda, a French mathematician and astronomer, discovered a way to compare the strengths of the earth's magnetic field at different points. His method was comparable to the use of a pendulum to compare the accelerations of gravity (page 127). The period of an oscillating magnetic needle, like that of a swinging pendulum, will be affected by the strength of the earth's field, magnetic in one case and gravitational in the other. In the case of gravitation, the mere swinging of a pendulum at two places gives only the ratio of the strengths of the earth's field at those places, not the absolute values. The same is true of the swinging of a magnetic needle.

Until a unit was established for magnetic field strength, which did not come about until 1833 (77:519), results secured by different observers, or by the same observer using different needles, could not be compared, and world-wide charts of magnetic field strengths could not be made.

It is now known that the strength of the earth's field is nearly three times as great at its point of maximum as at its point of minimum, the corresponding angles of dip being 87° and 8° respectively. Today magnetic charts always go in sets of three, the first two, consisting of isogonic and isoclinic lines respectively, have already been mentioned. The third type of chart lays down a set of *isodynamic* (equal field strength) lines over the map of the world. By consulting all three maps, we can learn the complete magnetic state of almost any point on the surface of the earth.

Like declination and dip, the strength of the earth's field is constantly changing. All three changes are for the most part slow, but ever since 1722 (86:156) the existence of regular daily and even hourly fluctuations have been known. In addition, there are erratic fluctuations, sometimes of major extent, often, though not always, accompanying and evidently produced by sunspots and *aurorae boreales,* this correlation having been observed as early as 1740 (86:139).

The Electromagnet

Aside from the magnetic compass, the commercial electromagnet is undoubtedly the principal channel of application of magnetism. The progenitor of all electromagnets was apparently that of William Sturgeon, made in 1823. It consisted of a single layer of bare wire wound on a varnished iron core. No significant improvement was made for four years. On October 10, 1827, Joseph Henry, the first American to make major contributions to physics after the time of Benjamin Franklin, entered the scene with a modification of the electromagnet, which though small in principle was great in effect. Sturgeon's electromagnet had consisted of a single layer of bare copper wire wound on an insulated iron core. Henry's improvement consisted in insulating the wire instead of the core and winding many layers instead of only one onto the iron core.[1]

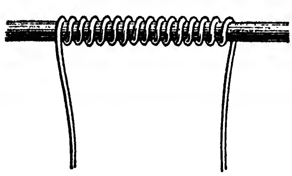


FIG. 410. STURGEON'S BAR ELECTROMAGNET
(From *Transactions of the Royal Society of Arts, 43,* plate 4 [1825].)

This improvement was not as simple as it appears today. Henry had to provide his own insulation, even as Ohm, two or three years before, had had to make his own wire. Some of the Henry relics, still preserved at Princeton University, indicate that the strips of silk cloth which constituted the insulation were provided by cutting up some of his wife's petticoats. If so, it was fortunate for science that modern styles had not yet come in.

[1] *Reviews of Modern Physics, 3,* 472 (1931).

FIG. 411. MODERN INDUSTRIAL ELECTROMAGNET
(Courtesy of General Electric Company.)

On account of this handicap, Henry had to labor interminably at winding insulation on miles of wire for use in his experiments.

The result, however, amply justified the effort. Henry made larger and larger electromagnets until by 1831 he had constructed a magnet which would support 750 pounds. This may be compared with Sturgeon's first which would support only a few ounces. In the course of his successive trials, Henry learned to establish a proper proportion between the voltage of his battery and the resistance of the windings on his electromagnets to secure a maximum effect. This constituted a partial rediscovery of Ohm's law, which, though Ohm had already published it, Henry was not to learn about until 1837, because of the delay in general recognition of Ohm's work.

Joseph Henry and the Beginning of Telegraphy

The further development of the electromagnet to the dimensions of that

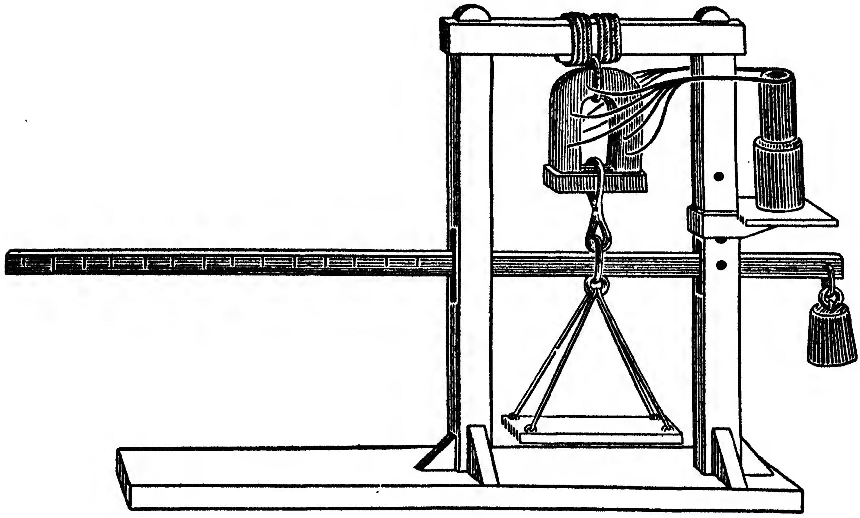


FIG. 412. HENRY'S DRAWING OF ONE OF HIS FIRST ELECTROMAGNETS
(From the *American Journal of Science, 19,* 408 [1831].)

now used in the heavy industries is a matter of improvement in minutiae of design which need be of no present concern. But a somewhat different application which Henry made is worthy of more than passing notice. Henry had not confined his attention to heavy electromagnets, but had also constructed small magnets capable of acting through long lines at a considerable distance from the battery. By 1831, he had made one which actuated a clapper to strike a bell, and demonstrated it to his classes and to visitors through a mile of wire, in that way transmitting signals. These facts are of considerable importance when it is realized that Morse, to whom the invention of the telegraph is commonly attributed, constructed his first model in 1835, applied for a patent in 1837, and was awarded it in 1840, in

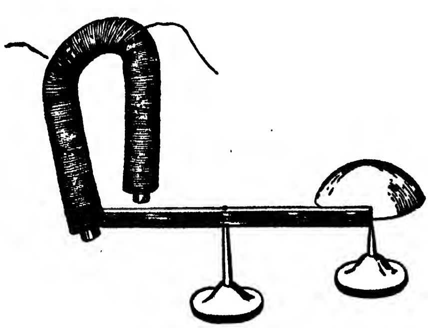


FIG. 413. HENRY'S TELEGRAPH
(His own illustration.)

the meantime having received detailed information about Henry's electromagnets, some of it directly from Henry himself, and incorporated these ideas into his own instruments.[1] Moreover, the corresponding invention of the telegraph in England in 1837, by Wheatstone and Cooke, did not occur until two months after Henry had visited Wheatstone on April 11 and freely given him all the information at his disposal about crucial points in the design of electromagnets for telegraphic purposes, the proper proportioning of resistance to voltage and length of line, as has already been mentioned. Henry never envied these men the fortunes they made, partly through appropriating his ideas, but he did feel bad that neither of them made a word of acknowledgment of the assistance which he had given so freely.

Questions for Self-Examination

1. What is the earliest record of observation of magnetic attraction and why is it significant?
2. Describe the development of the idea of magnetic polarity.
3. What is taken as the measure of the strength of a magnetic field?
4. Name and define the unit of magnetic flux density.
5. Name and define the unit of magnetic field strength.
6. Name and define the unit of magnetic flux.
7. What are some of the principal facts about magnetic declination and dip?
8. Describe the evolution of the electromagnet.

Problems on Chapter 41

1. A wire l meters long carrying I amperes is perpendicular to a magnetic field of flux density B webers/m². What force F in newtons does it experience?

l	I	B	F
1	10	.01	.10
1	15	.005	.075
1	20	.002	.040
1	25	.001	.025

2. A square galvanometer coil of length l centimeters and breadth b centimeters consists of N turns. It is suspended vertically in a horizontal magnetic field of flux density B webers/m² and carries a current of I amperes. When set so that its plane makes an angle α with the direction of the field, what torque L in newton meters does it experience?

l	b	N	α	I	B	L
5	1	100	0	.001	.01	5.0×10^{-7}
5	1	100	5	.001	.01	5.0×10^{-7}
5	1	100	20	.001	.01	4.7×10^{-7}
5	1	100	50	.001	.01	3.2×10^{-7}

3. There are N wires on the armature of an electric motor, each of length l centimeters parallel to the axis of rotation b centimeters from it, carrying a current of I amperes and rotating in a radial magnetic field of flux density B webers/m² at n revolutions per second. What horsepower P does the motor develop?

[1] *Annual Report, Smithsonian Institution, 1857*, pp. 99 ff. and 264 ff.

N	l	b	I	B	n	P
2000	30	10	1	.50	20	5.1
2000	30	10	2	.45	30	14.
2000	30	10	3	.40	40	24.
2000	30	10	4	.30	50	30.

4. What is the flux density B in webers/m² within a long solenoid of length l centimeters, consisting of N turns of wire of radius r centimeters, carrying a current of I amperes?

	l	N	r	I	B
4.	20	200	1	1	.0013
	30	200	1	2	.0017
	40	200	1	3	.0019
	50	200	1	4	.0020

	N	r	I	B
5.	1	20	10.	3.1×10^{-5}
	10	20	2.	6.3×10^{-5}
	25	20	.5	3.9×10^{-5}
	50	20	.2	3.1×10^{-5}

5. What is the flux density B in webers/m² at the center of a circular coil of N turns of radius r centimeters, carrying a current of I amperes?

6. A circular coil of N turns and radius R centimeters carries a current I amperes. Concentric with this, but with its plane at right angles to the plane of this, is another coil of n turns and radius r centimeters carrying a current of i amperes. The second coil is so small that the magnetic field in its neighborhood due to the first may be considered uniform. What is the force moment L in newton meters tending to turn the two coils into the same plane?

N	R	I	n	r	i	L
20	10	1	10	.5	1	$.99 \times 10^{-7}$
20	15	1	10	1.0	1	2.6×10^{-7}
20	20	1	10	1.5	1	4.4×10^{-7}
20	25	1	10	2.0	1	6.3×10^{-7}

CHAPTER 42

Electromagnetic Induction

Henry's First Observation of Electromagnetic Induction

Important as the telegraph is and, still more so, the almost infinite variety of other applications of the electromagnet which Henry developed to the point of practicable use, they are, after all, mere by-products of science. Henry's real contribution to physics was his observation in August, 1830, of a new and fundamental phenomenon, that of induction. He was using one of his electromagnets, the terminals of which were soldered to the plates of the battery. There was no switch in the circuit, and to turn the electromagnet on and off he had to immerse the battery plates in the electrolyte and withdraw them as required. Around the armature (a piece of soft iron laid across the poles of his electromagnet), he had wound a few turns of wire and connected their terminals to a galvanometer. It is to be borne in mind that there was no electrical connection between the wires leading to the galvanometer and those leading to the battery. Let him tell about the subsequent observations in his own words:[1]

> At the instant of immersion, the north end of the needle was deflected 30° to the west, indicating a current of electricity from the helix surrounding the armature. The effect, however, appeared only as a single impulse; for the needle, after a few oscillations, resumed its former undisturbed position in the magnetic meridian although the galvanic action of the battery, and consequently the magnetic power, was still continued. I was, however, much surprised to see the needle suddenly deflected from a state of rest to about 20° to the east, or in a contrary direc-

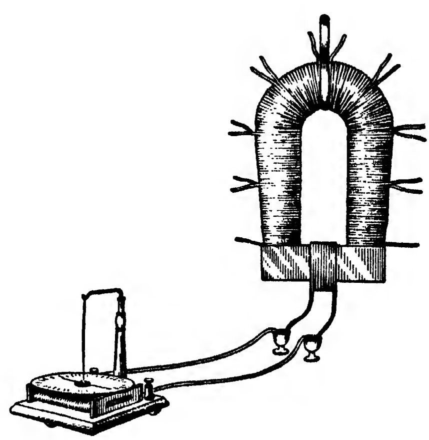


FIG. 414. HENRY'S DISCOVERY OF ELECTROMAGNETIC INDUCTION
(*The Electrical Engineer*, *13*, 53 [1892].)

[1] *American Journal of Science*, *22*, 405 (1832).

tion when the battery was withdrawn from the acid, and again deflected to the west when it was reimmersed. The operation was repeated many times in succession, and uniformly with the same result....

This experiment illustrates most strikingly the reciprocal action of the two principles of electricity and magnetism, if indeed it does not establish their absolute identity. In the first place, magnetism is developed in the soft iron of the galvanic magnet by the action of the current of electricity from the battery, and secondly the armature, rendered magnetic by contact with the poles of the magnet, induces in its turn currents of electricity in the helix which surrounds it; we have thus as it were electricity converted into magnetism and this magnetism converted again into electricity.

The Prior Search for Induction

The key to the whole undertaking is to be found in Henry's last clause, "magnetism converted again into electricity." The phenomenon is now called *electromagnetic induction.* It had been diligently sought by experimenters ever since Oersted in 1820 had demonstrated the reverse effect, that is, the production of magnetic effects by a current of electricity. From that time on, competition had been keen to discover the production of a current of electricity by magnetism. It is somewhat amusing to observe that, such is the influence of wishful thinking, no less a person than Fresnel, in the height of the fever induced by Oersted's discovery, announced that he had decomposed water by current from a coil of wire within which a magnet had been placed motionless. This had emboldened Ampère to remark that he too had noticed something in the way of production of electric currents from a magnet. But within a few weeks both statements were withdrawn by their authors.

The search, however, went on. In 1822, Ampère actually observed the effect of an induced current, but failed to recognize it.[1] In 1824 the same mishap occurred, in a different form, to Arago. His observation (3:4:424) that a pivoted bar magnet would be set in motion by an adjacent rotating copper disk was moreover repeated by others without their recognizing that the torque between the two was produced by the magnetic effects of currents induced in the disk by its motion near the magnet. In 1826, Ampère "muffed" the discovery of induction a second time by failing to see the significance of his successful repetition of Arago's experiment, with a coil of wire carrying a current substituted for Arago's magnet. Among those who made several unsuccessful experiments during this time while searching for electromagnetic induction was Michael Faraday. Faraday, however, was ultimately successful. Indeed, except for Henry's priority of observation, Faraday's work was even more significant than that of Henry, in that he published first, and that, aided by equipment that was far superior to anything Henry could command, his work was more thorough and conclusive.

[1] *Philosophical Magazine* (fifth series), *39*, 534 (1895).

Henry and Faraday

Faraday has been encountered before, both in Faraday's laws of electrolysis (page 435) and as having been the first to demonstrate the rotation of a magnet about a wire and of a wire about a magnet (page 442). He was associated with the Royal Institution in London, where he progressed from a somewhat menial assistant's position to that of the directorship. Between his life and that of Henry there were several curious parallels. Both were of humble and unpromising parentage. Both were apprenticed to tradesmen, Henry to a watchmaker and Faraday to a bookbinder. Both received their inspiration to a scientific career through casual reading during their middle teens of a current book on science, Henry by Gregory's *Lectures on Experimental Philosophy* and Faraday by Marcet's *Conversations on Chemistry*. With both boys, the early death of the father left with each a premature responsibility for the affairs of the family. Henry succeeded in acquiring a better education than Faraday, but Faraday later enjoyed the enormous advantage of continuous association with the most productive scientists of the time. Both men had their attention drawn to electromagnetism and induction and worked assiduously at these fields. Either could easily have made a fortune by exploiting his discoveries commercially, but both resolutely refused to take out any patents, and died without having laid aside even a modest competence. Both men developed a clear prevision of electromagnetic radiation almost half a century before the experiments of Hertz launched the age of radio on its course. And finally, each spent the larger portion of his life as director of a scientific institution, Faraday with the Royal Institution, established at London by an American, and Henry with the Smithsonian Institution, established at Washington, D.C., by an Englishman.

Faraday's Rediscovery of Induction

Faraday must have shared the general "gold rush" of 1820 and 1821 to acquire further information on the relations between magnetism and electricity. But the first actual record that he had embarked on the search for induction occurs in an entry of 1822, a memorandum of things which he expected to undertake. It read: "Convert magnetism into electricity." In 1824 he tried passing a bar magnet through a helix of wire, but could detect no current. In 1825 he carried out some very elaborate experiments of a similar kind, with every variation he could conceive, but still no results. Later observations showed that what he had needed were stronger magnets and more sensitive galvanometers. But at the time he concluded that the effect which he sought did not exist. Apparently he thereupon laid aside this search, not to resume it until 1831 and at last on August 29 of that year found it in a form different from that in which he had looked for it (40:*1*:367). It was at that time, apparently, that he first realized that induced currents were transient effects, existing only during the interval that the actuating magnetic field was changing. Once in possession of this idea, Faraday ran rapidly up the ladder of discovery to a complete comprehension of the phenomena of induction, and within less than a month was able to announce his discoveries. Though this was a full year after Henry's first observation of the same effect, Faraday immediately published his findings and thereby established formal priority.

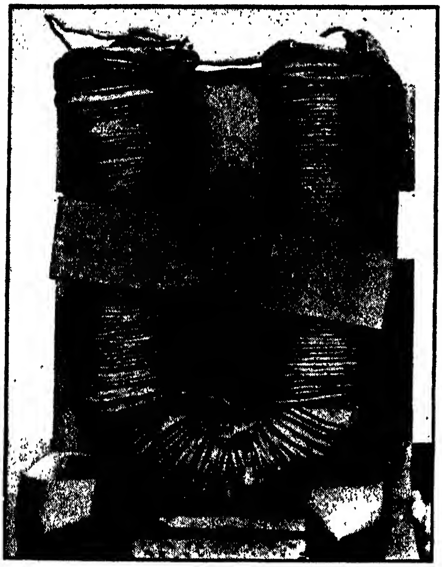


FIG. 415. FARADAY'S "GREAT" ELECTROMAGNET
(The core was a portion of a link of a large anchor chain.)

Faraday's first observation was quite similar to Henry's — the jerk of the needle of a galvanometer connected to one coil when current in an adjacent coil was started or stopped. He varied the experiment by changing the distance between the two coils instead of starting and stopping the current, and secured the same result. He accentuated the effect by using cores of soft iron. Then he thrust a bar magnet into the coil connected to the galvanometer and withdrew it, and also moved the coil, leaving the magnet stationary, in all cases producing the same result (40:*1*:375). Finally he substituted the magnetic field of the earth for that of the magnet and was again successful.

Faraday announced his findings to the Royal Institution in December 1831 and January 1832. He published them in April 1832.[1] Seeing a report of this publication the following June, Henry was stimulated into doing some hasty additional work, having laid aside his researches with the recurrence of his heavy teaching responsibilities the preceding September. He brought his previous experiments to a conclusion and secured publication in July, acknowledging Faraday's priority based on date of publication and drawing a clear distinction between what he (Henry) had done before reading Faraday's article and what he had done afterward. So thorough had been Faraday's work, that there was little for Henry to add, however, except at one point. This he developed in the last paragraph almost as if it were an afterthought. It represented some work, however, which had been done three years before, and which in some respects was fully as significant as the inductive effects which he was describing in the body of his paper. This new contribution will be described on page 514.

[1] *Philosophical Transactions*, *122*, 125 (1832).

The Principle of the Transformer

Out of the principle of induction discovered by Henry and Faraday have grown an almost inconceivable number of inventions and engineering applications. One of them, the *transformer,* bears almost the same relation to induction that the telegraph does to electromagnetism. It is the one device, above all others, that has made the distribution of electric power over distances of more than a mile or two economically possible. A comprehension of this method of solving the problem of electric distribution may be facilitated by referring to a similar problem in the distribution of natural gas. In domestic appliances, gas is used at very low pressure. The cost of the large pipes which would be required to carry gas at low pressure over considerable distances would be prohibitive. But by compressing it, the greater density and higher speed of flow securable makes it possible to pipe natural gas through small pipes for hundreds of miles. Reducing-valves change its pressure at the consuming area to conventional values, whereupon large pipes again become necessary.

An analogous situation exists in the distribution of electric power, with voltage substituted for pressure, current for volume, and size of wires for size of pipes. It is evident from the fact that electric power is the product of voltage and current that a given power rate can be maintained by heavy current at low voltage or by small current at correspondingly high voltage. Application of Ohm's law to the transmission lines for the two cases will show that the power lost in transmission for the first case would be very large unless the size of the wires were so great as to be prohibitive. But with a "step-up" transformer performing the part of the pump at the power-house end of the line and a "step-down" transformer in lieu of the reducing-valve at the consuming end, the conflicting requirements of low voltage and great distance are reconciled in the electrical system somewhat as the corresponding conflict of low pressure and distance is in the gas system.

The Commercial Significance of the Transformer

But it is to be remembered that when induction is produced by the action of the current in one coil on an adjacent coil, as in the transformer, it is only while the current is increasing or decreasing in the former (the "primary") that current is induced in the latter (the "secondary"). Hence, a transformer can function only during *changes* in value of the primary current. It is almost solely for this reason that commercial electric power takes the form of the so-called *alternating current.* As the name implies, the direction of flow is not constant, but is constantly reversing and re-reversing. The American standard *frequency* of such reversals is sixty double reversals per second, termed sixty *cycles.* These reversals are not abrupt, but gradual. The "wave form" most desired for alternating current (though it is seldom completely realized) is the simple sine wave (Fig. 416). Thus at all times, except at the instants represented by the crests and troughs, the current is changing, and hence can induce currents in a secondary coil, which is part of a closed circuit. These induced currents will also alternate, and with the same frequency as the primary currents. If the secondary coil contains a larger number of turns of wire than the primary, the secondary voltage will be the greater of the two and a "step-up" transformer is the result. If the reverse is the case, a "step-down" transformer is the result. The ratio of the voltage is very nearly the ratio of the number of turns on the two windings. Usually the two windings may be used interchangeably, either one as the primary, the other one as the secondary.

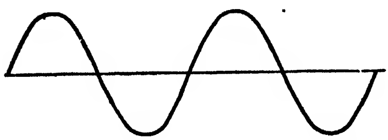


FIG. 416. AN IDEALIZED ALTERNATING CURRENT WAVE FORM

The date when Henry first announced his observation of the interconvertibility of "intensity" and "quantity" was a momentous one. The present-day prevailing use of alternating current, which makes possible the full utilization of his observations, did not occur for nearly seventy-five years. But the existence of great centers of electric power where production in large units makes for economy, and of long cross country power lines for distribution are none the less very directly attributable to that almost casual observation. It would be impossible to estimate the influence which cheap and easily accessible electric power has had on western culture. But the device which typifies it more than any other single thing would be a modern transformer, the product of many minds in succession, beginning with Joseph Henry's conception of the interconvertibility of "intensity" and "quantity" by induction.

Henry's Discovery of Self-Induction

It has been hinted (page 512) that Henry's paper of 1832 described a major phenomenon that Faraday had not observed. This phenomenon has since received the name of *self-induction*, to distinguish it from the inductive effect described above which is somtimes called *mutual* induction. The difference between the two can be made clear by a closer inspection of self-induction.

The appropriateness of the adjective "mutual" to describe the foregoing variety of induction will be evident in the observation that the form in which both Henry and Faraday first identified it was the effect of a change in current in one coil on *another* coil. It becomes natural to inquire, "Will induction still perform its function if the two coils, instead of being entirely separate, are simply different sections of the same coil? If so, how will it make its effect evident?" Henry's observation provided an affirmative answer to the first question and showed one way in which the second might be answered.

He said:

> I may, however, mention one fact which I have not seen noticed in any work, and which appears to me to belong to the same class of phenomena as those before described; it is this; when a small battery is moderately excited by diluted acid, and its poles which should be terminated by cups of mercury, are connected by a copper wire not more than a foot in length, no spark is perceived when the connection is either formed or broken; but if a wire thirty or forty feet long be used instead of the short wire, though no spark will be perceptible when the connection is made, yet when it is broken by drawing one end of the wire from its cup of mercury, a vivid spark is produced.... The effect appears somewhat increased by coiling the wire into a helix.

In a later publication Henry confirmed and extended his observations as follows:[1]

> A wire coiled into a helix gives a more vivid spark than the same wire when uncoiled.
>
> Large copper handles, soldered to the ends of a coil of ninety-six feet, and these both grasped, one by each hand, a shock is felt at the elbows, when the contact is broken in a battery of a single pair with one and a half feet of zinc surface.
>
> The effect produced by an electromagnet, in giving the shock, is due principally to the coiling of the long wire which surrounds the soft iron.

This self-inductive effect Henry termed the "extra current." Though his first published account of it was in 1832, the observations upon which it was based were first made in August, 1829. Always somewhat remiss about publication, Henry did not suffer as seriously from this lapse as he did from delay in publishing his observations on mutual induction. This time his publication was the first to be made, notwithstanding the delay. Faraday did not observe self-induction until 1834, and published it the same year, having somehow missed the significance of the final paragraph in Henry's 1832 paper.

The Contribution of H. F. E. Lenz

The independent and almost simultaneous discoveries by such widely separated workers as Henry and Faraday constitute another example of the oft-repeated observation that scientific developments are as much a product of the time as of the man. In this case, the fact is made more striking by the simultaneous work of a Russian scientist, H. F. E. Lenz (1804–64), in the same field. Lenz was only a step behind the other two men. He knew nothing of the work of Henry and had only a partial knowledge of that of Faraday. Besides duplicating some of the work of each of these men, he formulated a generalization which had eluded both of them and which now goes by the name of Lenz's law.

[1] *Journal of the Franklin Institute, 15,* 169 (1835).

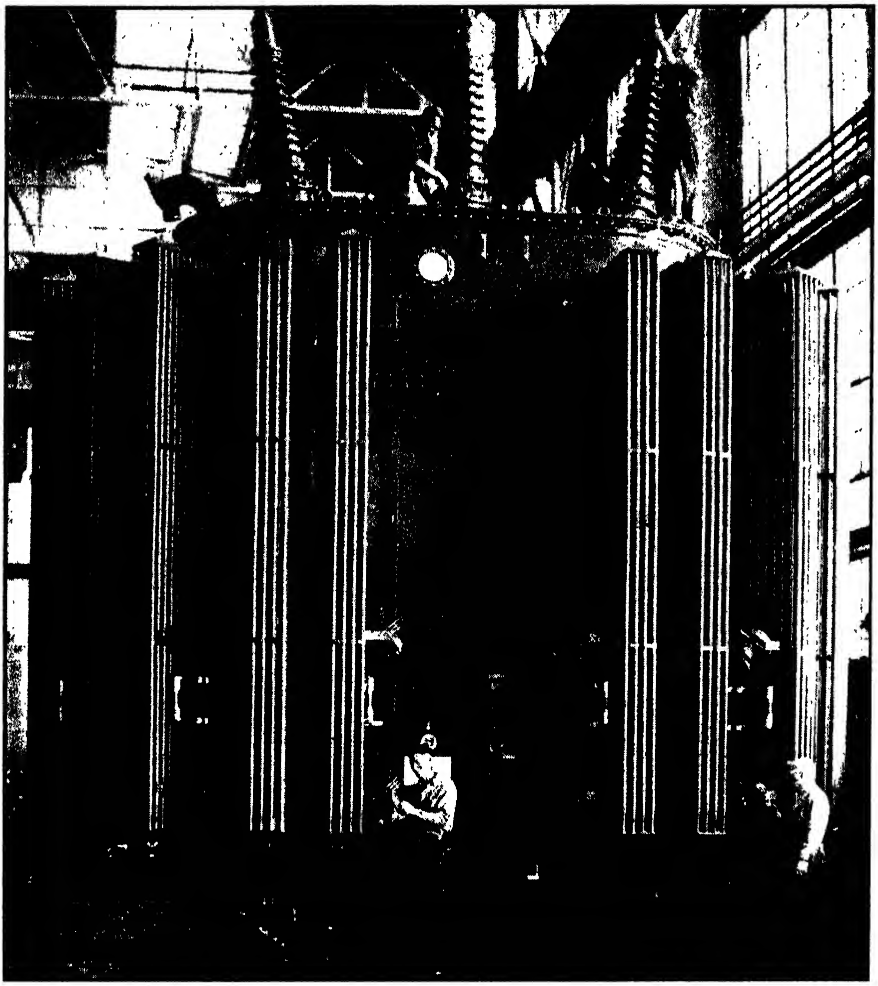


FIG. 417. A MODERN TRANSFORMER
(Courtesy of General Electric Company.)

Lenz's law has to do with the reaction of an induced current on the inducing agency. It may be stated as follows:

> A current brought into action by an induced electromotive force always produces effects which oppose the inducing action.

This generalization is very useful. Except for the inductive opposition to the rise of current, such rise would be instantaneous which is not the case actually, though it is very rapid. Similarly for the stoppage of current, which also requires a finite but very short time. In both cases the induced electromotive force opposes whatever change is occurring, first the rise and then the fall of current respectively. Some of the implications of this fact are developed further in Chapter 43.

This inductive opposition cuts down the mean value of any alternating current which would otherwise flow and which would be determined by Ohm's law, resulting in what appears superficially to be a greater resistance of the coil to alternating current than to direct. This apparent extra resistance, termed *reactance*, will be discussed in a later chapter. A coil designed to utilize it for the purpose of cutting down current flow is termed a *choke coil*. The effectiveness of a choke coil will be greater as the frequency is increased or as any circumstance obtains which increases the inductance, such as the introduction of iron.

When a magnet is thrust into a coil, current induced in the coil will oppose the insertion of the magnet, making it necessary to exert a greater force than would otherwise be required to move it. This additional force is hard to observe in this particular case, because it is small in comparison with the forces involved in overcoming friction and inertia. But it is not always small. It has its counterpart in the power necessary to drive a generator of electricity. When such a generator is anywhere near its maximum output, the added power required because of its inductive action is usually many thousands of times the power necessary to drive it at the same speed when not delivering current. Again, if the direction of the current in the secondary of a transformer were observed it would be found to have a net value opposite to that in the primary.[1] Thus the induced current acts on the surroundings in opposition to the effects of the inducing current. Hence the self-induction of the primary winding is reduced to just the extent that current flows in the secondary. Consequently, increasing the secondary current reduces the choke-coil effect of the primary winding and allows the primary current to increase correspondingly. This produces a result which to the uninitiated is often puzzling. As more current is taken from the secondary of a transformer, the current supplied to the primary increases in substantially the same proportion, though the two are not electrically connected at all. This is one of the many phenomena which, otherwise sometimes difficult to comprehend, is clarified by Lenz's law. Other examples will appear as the subject develops.

The Units of Self-Inductance and Mutual Inductance

The proclivity of inductance for opposing all changes of current reminds one of a corresponding property of mass in a mechanical system. Between the two there is in fact more than a passing resemblance, there is a close parallel. This parallel may be made evident as follows. From Newton's second law of motion,

inertial reaction is proportional to rate of change of velocity

or

$$\frac{\text{inertial reaction}}{\text{rate of change of velocity}} = \text{a constant for a given body termed its } \textit{mass}.$$

[1] It is not opposite at every instant, the conditions varying with the design of the transformer and the degree of its load.

Similarly in an inductive circuit,

induced voltage is proportional to rate of change of current

or

$$\frac{\text{induced voltage}}{\text{rate of change of current}} = \text{a constant for a given circuit termed its } \textit{inductance.}$$

Stating the two cases algebraically,

$$\frac{F}{\Delta V/\Delta t} = M \quad \text{and} \quad \frac{E}{\Delta I/\Delta t} = L, \tag{1}$$

L being the notation commonly assigned to self-inductance. In case the numerical value of E and $\Delta I/\Delta t$ are the same, the value of L will be unity. Hence, the unit of self-inductance is usually defined as

the inductance of a coil in which a rate of change of current of one ampere per second induces an electromotive force of one volt.

This unit has, with entire appropriateness, been named the *henry*. The henry, like the farad, is too large a unit for convenience in rating inductances commonly encountered. The most common subdivision is the *millihenry*, which, as the prefix indicates, is one thousandth of a henry. The same unit, defined in the same way, applies to the inductive influence of a coil on an adjacent but separate coil. In that case it is termed *mutual inductance* instead of self-inductance. It is defined as *the ratio of the induced secondary e.m.f. E* to the time rate of change of current $\Delta I/\Delta t$ in the primary and is expressed, as before, by

$$M = \frac{E}{\Delta I/\Delta t}. \tag{2}$$

If E is in volts, ΔI in amperes and Δt in seconds, M is in henrys.

Induced E.M.F.

Passage of a current through a coil produces a magnetic flux within the coil. Faraday pictured the "lines of force" constituting this flux as forming closed curves, emerging from the magnetic north end of the coil and re-entering the south end. These closed "lines of force" are thus linked with the windings of the coil. When the current changes, the flux changes in proportion. This change of flux is associated with an induced e.m.f., the magnitude of which is proportional to the rate of change of the flux. That is,

$$E \propto \frac{\Delta\phi}{\Delta t},$$

where E is the induced e.m.f. in volts, and $\Delta\phi/\Delta t$ is the rate of change of the flux through the coil. In the M.K.S. system the proportionality factor is unity and hence,

$$E = \frac{\Delta\phi}{\Delta t}. \tag{3}$$

Equation (3) applies only to a coil consisting of a single turn. If there are two turns, the e.m.f. induced by a changing flux will obviously be doubled. Hence for a coil of N turns

$$E = N\frac{\Delta\phi}{\Delta t}. \tag{4}$$

Confining attention for the moment to a coil having a single turn, the rate of change of the flux will be proportional to the rate of change of the current producing it. That is,

$$\frac{\Delta\phi}{\Delta t} \propto \frac{\Delta I}{\Delta t}.$$

The proportionality factor is simply the coefficient of self-induction L of the coil. That is,

$$\frac{\Delta\phi}{\Delta t} = L\frac{\Delta I}{\Delta t}. \tag{5}$$

Combining equations (3) and (5), the e.m.f. induced in a coil by changing the current flowing through it is

$$E_1 = L\frac{\Delta I}{\Delta t}. \tag{6}$$

If a second coil were within the region occupied by the magnetic flux of the first (the case of the transformer), an e.m.f. would also be induced in it by a change of current in the first, the magnitude of the e.m.f. being given by a relation completely analogous to equation (6) except that the self-induction of the one coil is now replaced by the mutual induction between the two. For such a case

$$E_2 = M\frac{\Delta\phi}{\Delta t}. \tag{7}$$

Now equations (4) and (6) both express the magnitude of the induced e.m.f. in a coil, one in terms of changing flux, the other in terms of changing current. Equating the two,

$$N\frac{\Delta\phi}{\Delta t} = L\frac{\Delta I}{\Delta t}. \tag{8}$$

If the flux changes from 0 to ϕ as the current rises from 0 to I, then, at the end

$$N\phi = LI. \tag{9}$$

Equation (9) makes it possible to calculate L for any coil for which the value of ϕ can be determined for a given I. As was shown in Chapter 41, this is possible for simple cases such as the long solenoid and the ring-form coil, the flux in webers being simply the product of the area of the cross-section of the coil and the flux density in webers/m^2.

Questions for Self-Examination

1. Compare the contributions of Henry, Faraday and Lenz to early knowledge of electromagnetic induction.
2. Tell the principle and the main function of the transformer.
3. State Lenz's law and give some examples of its action.
4. Name and define the units of self-induction and of mutual induction.
5. Upon what does induced electromotive force depend?

Problems on Chapter 42

1. A current I amperes through a coil of self-inductance L henrys drops to zero in t seconds. How many volts electromotive force E are induced?

	I	L	t	E		M	E
1.	5	.1	.001	500	2.	.03	150
	4	.15	.001	600		.05	200
	3	.2	.001	600		.07	210
	2	.25	.001	500		.09	180

2. Another coil is placed at such a distance from that of the preceding problem that the mutual inductance between the two is M henrys. What electromotive force E is induced in it in volts?

3. A solenoid having N_1 turns produces a magnetic flux of ϕ webers when I amperes flow through it. A secondary coil of N_2 turns is wound over it. What is the value M of the mutual inductance and the value L of the self-inductance of the primary in millihenrys?

N_1	ϕ	I	N_2	M	L
1000	.0004	2	400	80	200
1200	.0003	3	500	50	120
1500	.0002	4	600	30	75
1800	.0001	5	700	14	36

4. A solenoid of length l centimeters consisting of N turns of radius R centimeters carries a current of I amperes. A secondary coil of n turns is wound upon the middle part of the solenoid. What is the average electromotive force E in volts in the secondary if the primary current falls to zero in t seconds? What is the mutual inductance M in millihenrys? Assume that the total flux through the primary cuts the secondary.

l	N	R	I	n	t	E	M
200	2000	10	5.	1000	.001	200.	39.
100	1500	5	2.	750	.001	22.	11.
50	1000	3	.6	500	.001	2.1	3.6
25	500	1	.5	200	.001	.079	.16

5. What electromotive force E in volts is generated in the *primary* of the preceding problem by the cessation of the current? What is the value of the self-inductance L in millihenrys?

l	N	R	I	t	E	L
200	2000	10	5.	.001	390.	79.
100	1500	5	2.	.001	44.	22.
50	1000	3	.6	.001	4.3	7.1
25	500	1	.5	.001	.20	.39

CHAPTER 43

Electric Transients

Rise and Fall of Currents

When a switch is turned on, thus completing a circuit, a certain time is required for the current to approach its maximum steady value. The time depends principally, though not exclusively, on inductance in the circuit. Ordinarily it is a small fraction of a second, but it may be much longer if the inductance is sufficiently large. The same condition holds when a current is turned off. The current does not fall instantly to zero, but dies away at a rate which is again partly dependent on the inductance in the circuit.

The value of a direct current in a circuit after a steady state has been reached depends upon voltage and resistance in accordance with Ohm's law, as was brought out in Chapter 39. Here, however, is another circumstance to be considered, namely, the temporary or *transient* value of a current prior to the attainment of a steady state and immediately following its cessation. The study of electric transients possesses a value all out of proportion to the time required for them to function. An electric power installation, designed to give "steady state" service, may be seriously damaged by transient electric surges accompanying a too abrupt stoppage of service due to some mishap. Also the sciences of telephony and of radio communication center in the control of transient electric phenomena.

A simple mechanical parallel will help to visualize the nature of electric transients. If one should undertake to push a heavy truck, the "steady state" speed that could be attained would depend entirely on the friction that had to be overcome. But the time necessary to attain that speed would depend primarily on the mass of the truck. The time might be considerable, even if the bearings were substantially frictionless. Also, once under way, the time for the truck to come to a stop after the force ceased

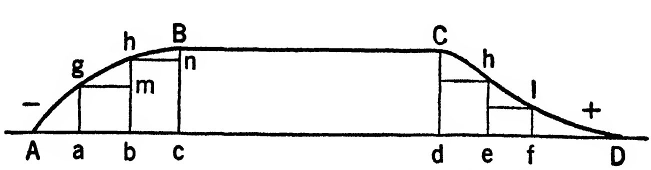


FIG. 418. HENRY'S DIAGRAM OF THE RISE AND FALL OF A CURRENT IN AN INDUCTIVE CIRCUIT

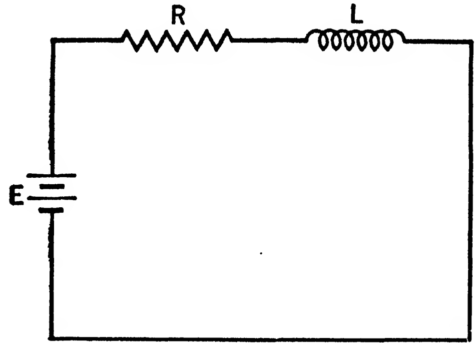


FIG. 419. A CIRCUIT WITH INDUCTANCE AND RESISTANCE

acting would again depend on the mass, now in relation to friction or other forces which act while the truck is in motion. Somewhat comparable statements can be made about the starting and stopping of electric currents.

Joseph Henry appears to have been the first to realize the existence of electric transients. He made his first public statement about them in 1840 in a paper which he read before the American Philosophical Society,[1] in which the diagram of Figure 418 appeared, representing the rise of current at "make" and the dying away of current at "break." He had observed the effect in electric sparks produced in circuits containing inductances studied by a mirror rotating ten times a second. He stated that the time required for a current to rise and to subside seemed to depend on the inductance. He estimated it to be of the order of a ten thousandth of a second for the inductances with which he was dealing. He could hardly have been using any of his heavy electromagnets in these experiments, or the times which he observed would have been much greater.

Circuit Containing Inductance and Resistance

Suppose a battery with an e.m.f. of E volts were connected to a circuit of resistance R ohms and self-inductance L henrys (Fig. 419). The final steady state value I of the current would of course be given by Ohm's law, and after that value was attained, the entire e.m.f. of the battery would be used to overcome the resistance; that is,

$$E = RI. \qquad (1)$$

But prior to the attainment of this steady state, part of the e.m.f. would be diverted to setting the current in motion against the opposition provided by the inductance. Represent the smaller (and variable) value of the current during this interval by i. By equation (1) of the preceding chapter,

[1] *Transactions of the American Philosophical Society*, *8*, 20, 34 (1843).

the voltage e_2 required to produce a change of current Δi in a time Δt through an inductance L is

$$e_2 = L\frac{\Delta i}{\Delta t}. \tag{2}$$

If now the remaining portion of the voltage available for producing the IR drop be called e_1, so that $e_1 + e_2 = E$, then during the transient stage

$$E = Ri + L\frac{\Delta i}{\Delta t}. \tag{3}$$

The solution of this equation for the instantaneous value of the current i at any time during the transient stage lies beyond the mathematical level which students of general physics may be assumed to have attained. Equation (3) is one of the simpler examples of what is termed a *differential equation.* The solution is

$$i = \frac{E}{R}\left(1 - \epsilon^{\frac{-Rt}{L}}\right). \tag{4}$$

The appearance of a new term ϵ (epsilon) need create no difficulty. It is merely a tabular quantity, whose values will be found in the exponential table on page xxiv of the Appendix. To illustrate the use of this table, values of the current through an ordinary electric bell or buzzer of resistance 5 ohms and inductance 1 millihenry will be computed for various intervals following the application of 5 volts by pressing the button. The value R/L is 5000. From the exponential table the following values of i may be taken, denominating by x the values of the fraction $\frac{Rt}{L}$ for the various values of t.

t	x	ϵ^{-x}	i
0.	0.	1.	0.
.00001	.05	.95	.05
.00005	.25	.78	.22
.00010	.50	.61	.39
.00020	1.00	.37	.63
.00030	1.50	.22	.78
.00050	2.50	.08	.92
.00080	4.00	.02	.98

The steady state value of the current, given by Ohm's law, is of course 1 ampere. Hence, within less than $\frac{1}{1000}$ of a second after the button is pressed, the current rises to within 2 per cent of its ultimate maximum. Figure 420 represents the rise of current as calculated above. If now the voltage impressed on the bell is removed, the resistance and inductance of the circuit remaining unchanged, equation (3) is replaced by

$$0 = Ri + L\frac{\Delta i}{\Delta t}, \tag{5}$$

the solution of which may be shown to be

$$i = I\epsilon^{-\frac{Rt}{L}}, \tag{6}$$

in which I is the value of the current at the instant that it begins to subside, namely, the steady state value. The values of i may be calculated

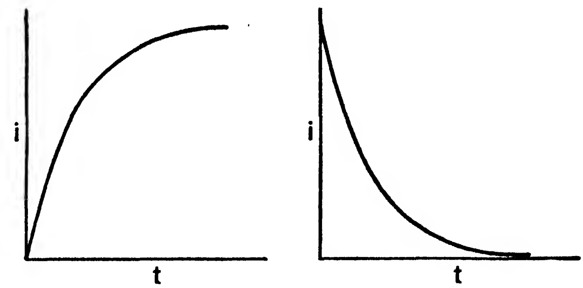


FIG. 420. RISE OF CURRENT IN AN INDUCTIVE CIRCUIT

FIG. 421. FALL OF CURRENT IN AN INDUCTIVE CIRCUIT

for various values of t as before. The results, represented graphically, are indicated in Figure 421. Figures 420 and 421 may be seen to follow the form of rising and subsiding currents indicated by Henry in Figure 418.

In an actual bell system the fall of current would be much more rapid than the preceding calculation indicates. The reason is that in the act of releasing the button a very high resistance (an air gap) is introduced into the circuit which makes the values of x in the calculation much larger than those assumed.

Circuit Containing Capacitance and Resistance

Another simple instance of electric transients occurs in charging and discharging condensers. Suppose a battery of e.m.f. E volts were connected through a resistance R ohms to a condenser of capacitance C farads (Fig. 422). The current-time equation in this case is

$$i = \frac{E}{R}\epsilon^{-\frac{t}{CR}}. \tag{7}$$

If E and R are given the values 5 volts and 5 ohms respectively as before and C is taken as 2 microfarads, the capacitance of a representative tele-

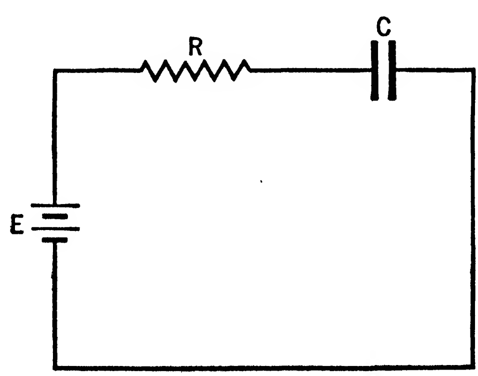


FIG. 422. CIRCUIT WITH CAPACITANCE AND RESISTANCE

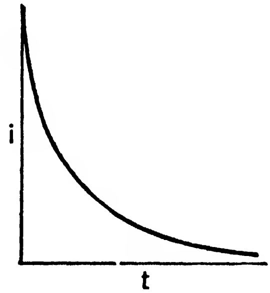


FIG. 423. CHARGING CURVE OF A CONDENSER

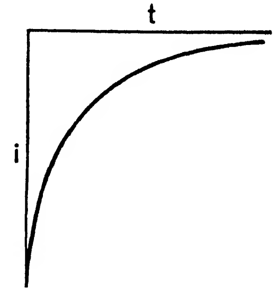


FIG. 424. DISCHARGE CURVE OF A CONDENSER

phone condenser, the exponential table will yield values of the current indicated in Figure 423. The current is a maximum at the beginning before any charge has accumulated on the condenser and hence before any voltage is built up in it to oppose the e.m.f. of the battery. It approaches zero as the voltage of the condenser approaches the e.m.f. of the battery. The discharge curve of the condenser will be the same, except that the currents are taken as negative to indicate flow in the opposite direction. The discharge would be effected by removing the battery and connecting together the terminals previously leading to it.

If, instead of considering *current* involved in the charge and discharge of a condenser, attention be given to the *quantity* of electricity in the condenser, the charge curve is of the same form as Figure 420 and the discharge curve has the same form as Figure 421. Thus the quantity-time curve for a condenser first charged and then immediately discharged would have the same general form as Figure 418, Henry's curve for rise and fall of current in an inductive circuit.

Circuit Containing Inductance and Capacitance

When inductance and capacitance are both present in a circuit, a new phenomenon appears. It takes the form of electric oscillations, comparable to mechanical oscillations under the joint action of a spring (to which a condenser is comparable) and a suspended mass (to which an inductance is comparable). Joseph Henry was apparently the first to identify this phenomenon also. It was an observation deserving to rank with his discovery of self-induction ten years earlier, if it is not, in fact, of even greater importance. It was the seed from which the whole science of radio transmission and reception ultimately grew. Earlier observers had indicated considerable perplexity about the direction of the current when a condenser was discharging. Even though known to be charged in the same direction, a condenser in discharging seemed to produce a current sometimes in one direction, sometimes in the other. The direction of the current was being inferred by observing the magnetic polarity produced in a steel needle when current from the condenser was sent through a coil wound around the needle.

Also some fifty years earlier, Wollaston, the man who first observed the Fraunhofer lines in the solar spectrum, had been similarly perplexed by observing that, when he tried to decompose water by electric discharge from Leyden jars, hydrogen and oxygen appeared at *both* electrodes, instead of hydrogen at one and oxygen at the other as in the case of "galvanic" current.[1] The anomaly, so long unexplained, Henry [2] attributed to a quality of condenser discharges which had never before been recognized. He said:

> The discharge, whatever may be its nature, is not correctly represented by the single transfer of an imponderable fluid from one side of the [Leyden] jar to the other. The phenomena require us to admit *the existence of a principal discharge in one direction, and then several reflex actions backward and forward, each more feeble than the preceding, until the equilibrium is obtained.*

Henry's theory of the nature of condenser discharges was never seriously questioned, though it left much to be desired. It was apparently mere inference, for he presented no experimental evidence not already known. Also, it called for mathematical development, though perhaps this was too much to expect, electrical science still being largely in the descriptive stage. Only eleven years elapsed, however, before the required mathematical treatment was forthcoming. It was furnished by Lord Kelvin in a classical paper entitled "On Transient Electric Currents." [3]

Kelvin's paper showed that a condenser discharge would be oscillatory only if the circuit contained an inductance. Thus the very coil, which Henry had used to magnetize needles in studying condenser discharges, had itself rendered the discharge oscillatory.

Kelvin also found what determined the number of oscillations per second which the system would produce (the *frequency*). He discovered it to depend primarily upon the inductance and capacitance in the circuit according to the relation

$$f = \frac{1}{2\,\pi\,\sqrt{LC}}. \tag{8}$$

One of Kelvin's comments is worthy of special notice, in view of subsequent occurrences. He expressed the opinion that an oscillatory discharge

> might be produced artificially from a Leyden phial or other conductor [sparking] across a very small space of air, and through a conductor of very large electrodynamic capacity [4] and small resistance. Should it be im-

[1] Quoted by Helmholtz, in *Philosophical Magazine* (4), *5*, 401 (1853).

[2] *Proceedings, American Philosophical Society, 2*, 193 (1842).

[3] *Philosophical Magazine* (4), *5*, 393 (1853).

[4] "Electrodynamic capacity" was the current term for *inductance*, the latter term not having yet come into use.

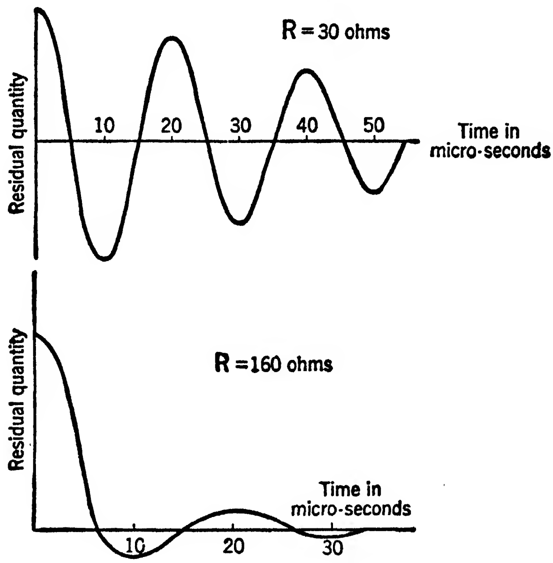


FIG. 425. TWO INSTANCES OF ELECTRICAL OSCILLATION IN CIRCUITS OF DIFFERENT RESISTANCE

possible on account of the too great rapidity of the successive flashes, for the unaided eye to distinguish them, Wheatstone's method of a revolving mirror might be employed and might show the spark as several points or short lines of light separated by dark intervals.

This experiment suggested by Kelvin was employed six years later by Fedderson [1] in a series of brilliant experiments. He identified in his revolving mirror the oscillatory nature of the spark and the circumstances under which oscillations degenerated into a single discharging surge; and he measured the frequency when oscillations were occurring.

Circuit Containing Inductance, Capacitance, and Resistance

If a resistance be added to the inductance and capacitance, the oscillations do not necessarily cease unless the resistance be made too large. In the complete absence of resistance or other agent causing progressive loss of energy from the oscillating circuit, electric oscillations once set up would, of course, continue undiminished indefinitely as would an oscillating mass on the end of a spring in the absence of friction or air resistance. Practically such loss of energy can never be avoided, and resistance is usually the principal occasion for it. Consequently, even though oscillations may continue in the presence of resistance, their amplitude falls off steadily, or is *damped.* To continue the above simile, if a mass hung on the end of a

[1] *Poggendorf's Annalen, 103,* 69 (1859).

spring were immersed in heavy oil, its oscillations would be damped more rapidly than in a light liquid or in air. If the oil were extremely viscous — as molasses is — no oscillations would occur at all. The corresponding electrical case was identified in Kelvin's theory, the *critical damping resistance* being the minimum resistance required to inhibit all oscillation. Two representative cases of electrical oscillation are represented graphically in Figure 425. Types of "oscillation" that result if the resistance is made equal to and greater than the critical damping value are shown in Figure 426.

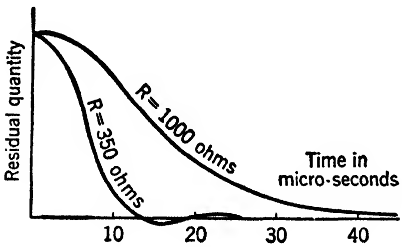


Fig. 426. Two Instances of Condenser Discharge in Circuits Having Greater than the Critical Damping Resistance

It was recognized very early, however, that resistance was not the only agency that acted to dissipate the energy of an oscillating circuit. Energy was also dissipated from such a circuit by radiation into surrounding space, much as a candle sends out energy in the form of light as well as by the heat it gives its surroundings. In 1864, J. Clerk Maxwell [1] developed the consequences of this concept in what was probably the most momentous mathematical discovery in the history of electrical science. This paper laid the foundation for the whole science of radio. Unfortunately it was too far ahead of its time to be comprehended. It was not until twenty years later that Heinrich Hertz, of Karlsruhe, Germany, produced and was able to receive at a distance, the electromagnetic waves, which Maxwell had predicted. The series of engineering developments from that point on, beginning with Branley, Fleming, and Marconi and leading to radio as we know it now, is common knowledge.

The series of events which Henry initiated in 1842 furnishes an excellent illustration of the interplay of experiment and mathematical theory in physics. Beginning with Henry's experiments, the major successive developments were alternately experimental and theoretical, one after the other, until the foundation was laid for perhaps the greatest engineering accomplishment of all time. Henry, Kelvin, Fedderson, Maxwell, Hertz; experiment, theory, experiment, theory, experiment. Not a single one of those stages could have been realized without the sequence that went before it. Even Henry's work was an outgrowth of his own earlier investigations and those of others. As a means of securing an insight into the scientific method, one can do no better than to consider this example. The technological sequels, which form the subject matter of the next four chapters, are little more than top foliage of this primary process, the taproot of our material civilization.

[1] *Philosophical Transactions*, *155*, 419 (1864).

Questions for Self-Examination

1. Describe the behavior of currents when turned on and off. What circumstances determine how fast they rise and fall?
2. Does Ohm's law apply while a current is changing? Explain.
3. Describe what happens to current and quantity of electricity in a condenser as a circuit containing resistance and capacitance is turned on and off.
4. A circuit contains a condenser, an inductance, and a resistance connected in series. The condenser is charged. What are the various possibilities as to what may ensue?
5. What happens in the circuit of the preceding question when inductance and capacitance are made larger in turn?

Problems on Chapter 43

1. A condenser of capacity 1 microfarad is charged to a potential difference of 100 volts. Its two terminals are then connected by a resistance of 10^6 ohms. The voltage across the condenser is given as a function of the time by the equation $e = E\epsilon^{-\frac{t}{CR}}$ where ϵ is the base of natural logarithms. Find the voltage e across the plates of the condenser for the following values of t in seconds. (See exponential table in Appendix.)

t	0	.2	.4	.7	1.	2.	4.
v	100	82	67	50	37	13	2

2. Repeat problem 1 for $R = 10^5$ ohms.

t	0	.1	.2	.3	.4	.5
v	100.	37.	13.	5.	1.8	.7

3. Plot values of v as ordinates against values of t as abscissas for problems 1 and 2.
4. A condenser of capacity C microfarads is charged to a potential difference of E volts. Its two terminals are then connected by a resistance of R ohms. What quantity q in coulombs remains in the condenser after the lapse of t seconds? Use the equation $q = CE\epsilon^{-\frac{t}{CR}}$.

	C	E	R	t	q		C	E	R	t	q
4.	1	100	10^6	0	$1.00\cdot10^{-4}$	5.	1	100	10^6	.25	$.22\cdot10^{-4}$
	1	100	10^6	.25	$.78\cdot10^{-4}$		1	100	10^6	.75	$.53\cdot10^{-4}$
	1	100	10^6	.75	$.47\cdot10^{-4}$		1	100	10^6	1.5	$.78\cdot10^{-4}$
	1	100	10^6	1.5	$.22\cdot10^{-4}$		1	100	10^6	3.0	$.95\cdot10^{-4}$

5. A condenser of capacity C microfarads, initially uncharged, is suddenly connected to a potential difference of E volts through a resistance of R ohms. What is the quantity q in coulombs in the condenser at a time t seconds after the connection is made? Use the equation $q = CE\,(1 - \epsilon^{-\frac{t}{CR}})$.
6. A condenser of capacity C microfarads is charged to a certain potential difference and immediately discharged through an undamped ballistic galvanometer, producing a deflection D. Its two terminals are then connected to the core and sheath respectively of a cable under test, and charged to the same potential difference as before. After t seconds it is discharged through the same galvanometer, pro-

ducing a deflection *d*. What is the insulation resistance *R* of the cable? Use the equation $Q = CE\epsilon^{-\frac{t}{CR}}$ where *Q* represents the quantity in coulombs remaining in a condenser of capacity *C* (in farads) originally charged to a potential difference of *E* volts, after having discharged through a resistance of *R* ohms for *t* seconds.

	C	*D*	*t*	*d*	*R*		*L*	*R*	*I*	*t*	*i*
6.	1	20	30	3	$16.\cdot 10^6$	7.	1	20	1	0	1.00
	5	20	30	4	$3.7\cdot 10^6$		1	20	1	.01	.82
	10	20	30	7	$2.9\cdot 10^6$		1	20	1	.03	.55
	20	20	30	10	$2.2\cdot 10^6$		1	20	1	.10	.14

7. A coil of inductance *L* henrys and resistance *R* ohms carries a current of *I* amperes. The potential difference which is setting up this current is suddenly removed. What is the current *i* in amperes at a time *t* seconds after the removal?

8. A coil of inductance *L* henrys and resistance *R* ohms is suddenly connected to a potential difference of *E* volts. What is the current *i* in amperes after the lapse of a time *t* seconds following the establishing of the connection?

	L	*R*	*E*	*t*	*i*		*C*	*L*	*N*
8.	1	20	20	.01	.18	9.	5	.1	230
	1	20	20	.03	.45		.5	.01	2,300
	1	20	20	.10	.86		.05	.001	23,000
	1	20	20	.25	.99		.0005	.0001	710,000

9. An oscillating circuit of negligible resistance contains *C* microfarad of capacity and *L* henrys of inductance. What is the natural frequency *N* of the circuit in double oscillations per second?

CHAPTER 44

Dynamos

Electrical Power as a Servant of Man

Comparisons are sometimes made between Greek culture and modern culture, in which a parallel is drawn between slave labor and modern machinery. It is occasionally said that the average American has at his disposal the mechanical equivalent of forty or fifty or sixty slaves, the number depending on the inclinations of the statistician making the statement. The limitations on the personal services that can be performed by mechanical and electric appliances remove much of the validity of this comparison, but, strained though the metaphor is, there is in it real material for thought.

The one circumstance above all others that lends a certain amount of validity to the parallel is the extended use of electric appliances, especially those involving the use of electric motors. All such devices, whether lighting, heating, or power-producing, consume electric energy. The consumption is usually far greater than batteries can supply, a condition which has occasioned the development of power-driven machines to produce the required electric energy; these machines are called *generators*. Electric motors and generators are identical in principle, the former transforming electrical into mechanical power and the latter performing the reverse process, mechanical power into electrical. Though in practice there are usually certain minor differences in construction between motors and generators, the two are customarily treated together. Two generations ago when motors and generators were evolving into something like their present forms, the generic name for them all was *dynamo-electric machines*, later contracted into *dynamos*.

The Birth of the Dynamo

The feature which all dynamos have in common, regardless of whether they are motors or generators, and regardless of type, is continuous relative rotation of conductors (usually coils of wire) and the magnetic fields. Thus when Michael Faraday in 1821 caused a magnet to rotate around a wire carrying a current as well as the same operation vice versa (40:*1*:50) (Fig. 357) he thereby devised an electric motor. It was rudimentary but it was, by a surprisingly wide margin, the first. Similarly when ten years later he rotated a copper disk between the poles of an electromagnet (Fig. 427) and observed the resulting current under various conditions he thereby devised the first rudimentary generator (40:*1*:381).

But Faraday, as has already been noted, took no interest in developing his discoveries into utilitarian forms. He was a scientist, not an inventor. He once remarked *à propos* of these very discoveries (123:248):

> I have rather, however, been desirous of discovering new facts and new relations dependent on magneto-electric induction than of exalting the force of those already obtained; being assured that the latter would find their full development hereafter.

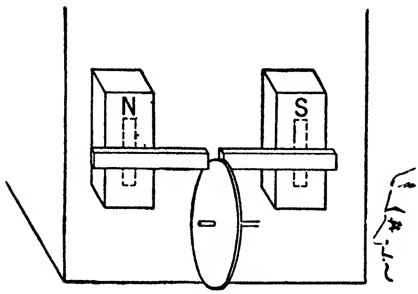


FIG. 427. FARADAY'S SKETCH OF HIS DISK-TYPE GENERATOR OF 1831

(From *Faraday's Diary*, published by the Royal Institution of Great Britain. G. Bell & Sons, Ltd., publishers.)

Faraday's confidence that applications "would find their full development hereafter" was so fully justified that the expected developments have constituted the larger part of the age of electricity. The developments were especially prompt in the case of the electric motor, which sprang almost full fledged into something like its modern form in 1834. The generator, curiously enough, in spite of its basic identity with the motor, and in spite of the fact that the motor could not realize its full potentiality while relying upon the expensive and inconvenient battery as the source of electric power, did not reach a form that was measurably complete, even in principle, until nearly twenty years after the perfection of the motor.

Faraday's rotating magnet and rotating wire of 1821 were mere toys in size and not even that in power, but they were immensely significant as the first embodiment of the transformation of electrical into mechanical energy. They were followed in 1823 by the stellate wheel of Peter Barlow, in reality simply Faraday's disk generator working as a motor. Nothing further occurred until 1831 when in quick succession three more toy motors were devised by an American, an Italian, and an Englishman respectively. But in 1834 two models were made almost simultaneously in the United States and in Russia which contained for the first time all the elements which were to make the electric motor an industrial tool instead of a mere toy. Moreover, both inventors realized their objective of constructing really powerful motors, as standards of power went in those days.

Thomas Davenport and the Electric Motor

In December of 1833, an enterprising young blacksmith of Forestdale, Vermont, named Thomas Davenport (1802–51), heard of a remarkable magnet used to extract iron from pulverized ore in the iron works at Crown

Point, New York, twenty miles away. Among its other reputed capabilities was that of suspending a hundred-and-fifty-pound anvil. The "magnet" was, of course, an electromagnet. It had been furnished by Joseph Henry, who was destined to give Davenport some direct aid in developing the electric motor. Davenport went to Crown Point to see this extraordinary device, intending also to purchase some iron for his shop while there. When the electromagnet was exhibited, he asked permission to make some experiments with it, a request which was denied. Thereupon he offered to buy the magnet outright, and a price being agreed upon, he laid down the money with which he had expected to purchase iron. This act was prophetic of his later experiences, for he was destined to reduce his family to destitution in his struggles to perfect and market the electric motor. Let him tell in his own words of his first experience with the new device:[1]

> As soon as I became the possessor of the magnet, I immersed the cups in the solution and then severed one of the conductors, so as to break the circuit of galvanism. Of course, I found the magnetism wholly destroyed, but on connecting the wires together with my fingers, the magnet became again fully charged. However rapidly the connection and separation of the conductors were made, I found the charging and discharging of the magnet to correspond, and I observed that the magnet produced a hundredfold more power than was required to make and break the connection.
>
> Like a flash of lightning the thought occurred to me, that here was an available power which was within the reach of man. If three pounds of iron and copper would suspend in the air 150 pounds, what would 300 pounds suspend? "In a few years," I said to some gentlemen present, "steam-boats will be propelled by this power."

Davenport's prediction was destined to be fulfilled sooner than he expected, but not by him. He set to work immediately to devise a motor, and within a few months was successful. He says:

> In July, 1834, I succeeded in moving a wheel about seven inches in diameter at the rate of thirty revolutions per minute. It had four electromagnets, two of which were upon the wheel and two were stationary and placed near the periphery of the revolving wheel. The north poles of the revolving magnets attracted the south poles of the stationary ones with sufficient force to move the wheel upon which the magnets revolved, until the poles of both the stationary and revolving magnets became parallel with each other. At this point, the conducting wires from the battery changed their position by the motion of the shaft; the polarity of the stationary magnets was reversed, and being now north poles, repelled the poles of the revolving magnets that they before attracted, thus producing a constant revolution of the wheel.

Unfortunately, no model or diagram of this motor has survived, but a later model is shown in Figure 429. The subsequent experiences of Davenport were the disheartening ones that have beset so many inventors who

[1] *Electrical Engineer, 11,* 4 (1891).

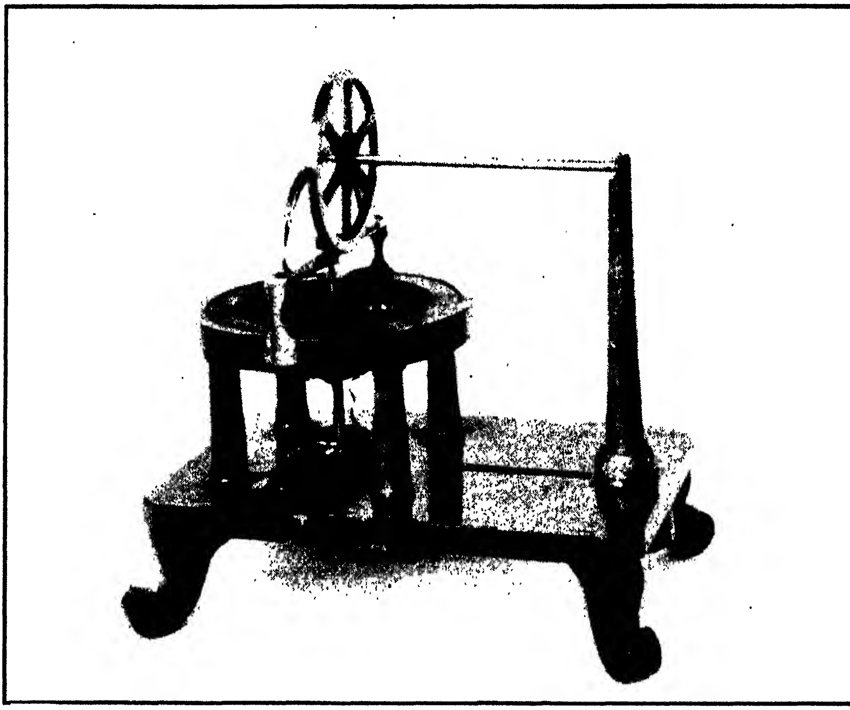


FIG. 428. THE MOTOR OF DAVENPORT'S FIRST PATENT, FEBRUARY 25, 1837
(United States National Museum.)

lacked the financial resources to develop and market their own inventions. His first patent application and the accompanying model were destroyed in the disastrous Patent Office fire of 1836. A new application was prepared and filed and was granted in 1837. The accompanying model, now in the Smithsonian Institution, is illustrated in Figure 428. In that same year he constructed and used motors of a third of a horsepower. In 1839 one of his motors weighing one hundred pounds, horsepower unknown, was driving a large printing press in New York City. But he and his supporters were being victimized by unscrupulous promoters and ultimately his financial liabilities overtook his engineering resources, compelling him to retire from the inventive field.

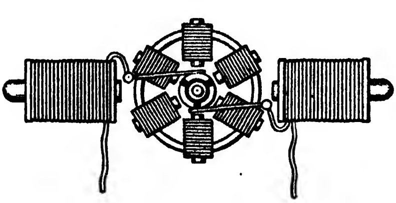


FIG. 429. THE DAVENPORT MOTOR OF 1835

Jacobi and the Electric Motor

During the interval covering Davenport's activities a Russian physicist named Moritz Hermann von Jacobi (1801–74), a brother of the famous mathematician Karl G. J. Jacobi, was independently pursuing almost exactly the same path. So nearly simultaneously were these efforts timed that it is quite impossible to establish priority of either over the other in any conclusive or significant degree. Jacobi's work grew out of Faraday's discoveries in somewhat the same way that Davenport's grew out of Henry's. It was in December of 1834 that Jacobi reported to the French Academy of Sciences his experiments with electric motors which he stated had begun in the preceding May. He measured the power of his first model at about $\frac{1}{50}$th horsepower. But in 1838, with another motor almost identical except in size, he succeeded in driving a boat along the river Neva in St. Petersburg (now Leningrad), the first fulfillment of Davenport's prophecy of five years before.

Many years elapsed before the motors of Davenport and of Jacobi were improved upon. Slow development was partly a consequence of the expense of operation, the only source of power being batteries. But it was also partly due to the fact that both men had so unerringly identified the basic principles of design of electric motors that there was little left for the next generation to do except to make improvements in details.

The Anatomy of the Direct-Current Motor

Several features contributed to the basic completeness of these first effective electric motors. One was the use of rotatory motion instead of the oscillatory motion produced in several of the early "toy" models, the latter being suggested no doubt by the reciprocating steam engine. Another was the use of electromagnets for both the fixed and moving parts, which was peculiar to these models and made it possible to produce magnetic forces which were far greater than when one set consisted of either permanent magnets or merely soft iron armatures. A third feature was the introduction of the *commutator* to change periodically the polarity of one of the two sets of electromagnets, thus taking advantage of magnetic repulsions as well as attractions in a way which none of the numerous prior motors had done. Both Davenport and Jacobi gave particular emphasis to this feature in their descriptions.

With both the field magnets and the *armature* (as the rotating part of the direct-current motor came to be called) supplied with current from the same source, the question of whether to connect them in series or in parallel with each other arose. The early motors were all series-wound, apparently for no very good reason. In later years, beginning about 1880, the shunt-wound motor (the parallel-connected type) came into use. Today both are common. The series type is used on street-cars and wherever heavy torque is frequently required, as in starting and in hill-climbing. Shunt-wound motors are however somewhat the more common, as their speed is more readily controlled and is steadier under fluctuating loads.

The Phenomenon of Counter-Electromotive Force

When a direct-current motor is in operation, it is functioning not only as a motor but as a generator also. The motion of the wires of the armature across the magnetic field induces an electromotive force which, by Lenz's law, will oppose the current which is being forced through the armature from the outside source of electrical energy. This internal opposing voltage is termed the *counter-electromotive force.* Any diminution in speed of the motor, such as would result when a street-car encounters a hill, reduces the counter-electromotive force, diminishing the opposition to the flow of current through the armature, thus permitting a greater current to flow. If the field windings are in series with the armature, the increased current produces a stronger magnetic field, thus increasing still further the torque which the motor exerts. Thus the effect which the automobile driver secures by shifting gears is automatically secured in the series-wound motor of a street-car through the agency of counter-electromotive force.

Another peculiarity of the counter-electromotive force is taken advantage of to facilitate control of the speed in shunt-wound motors. Contrary to one's natural preconception, increase of speed is secured by decreasing the strength of the magnetic field, and vice versa. This peculiar circumstance comes about through the fact that weakening the field decreases the counter-electromotive force, thus permitting more current to flow through the armature. It is simpler to interpose a resistance in the field circuit of a shunt-wound motor than in the main circuit, because the currents are far smaller in the former. So the shunt-wound motor lends itself more readily to speed control than does the series-wound, albeit in what seems at first to be a rather back-handed manner.

Jacobi comprehended fully the central rôle played by counter-electromotive force in the operation of motors, calling it the "electromotive force of reaction." The importance of the concept can scarcely be overemphasized, but it is remarkable that Jacobi should have caught the significance of it as early as he did. There is no evidence that Davenport was equally discerning.

To Jacobi also is due the somewhat less impressive first recognition that motors and generators are in principle structurally identical. The lateness of this discovery (1850) seems curious today, especially in view of the identity of Barlow's wheel of 1823 with Faraday's disk of 1831. But, as will be seen, the early designers of generators struggled through the discoveries of commutator, armature windings, and shunt and series field connections for generators entirely independently of the corresponding previous discoveries in connection with motors. Only when the process was nearly completed did it dawn on anyone that motors and generators were completely interconvertible.

The Industrial Function of the Electric Motor

It had been the idea of Davenport and Jacobi that the electric motor would compete with the steam engine as a "prime mover," that is, as a device for the direct conversion of heat or chemical energy into work. They envisioned their motors as consuming zinc and other chemicals making up batteries in the same way that steam plants consumed wood and coal, and expected their method of producing commercial power to be the more economical of the two. This expectation was not borne out. The result was that commercial development of the electric motor had to await the discovery of an economical source of supply of electricity. This ultimately came in the form of the electric generator, a device which itself required to be driven by a prime mover. Thus combinations of motors and generators, together with the circuits connecting them, became merely very convenient and versatile channels for distributing the power produced by steam engines and water wheels. Instead of competing with the steam engine, the electric motor facilitated the easy application of the engine's power and thus tremendously stimulated the growth of the machine age. This is perhaps illustrative of the basic misconception which underlies much of the apprehension with which "technological unemployment" is often regarded. While machines often do supplant particular labor groups, they have almost invariably created new demands which have multiplied the general market for labor by hundreds and thousands-fold.

The Development of the Generator

In a certain sense, the first electric generator consisted of the coil and bar magnet with which Faraday in 1831 produced a surge of induced electricity by mechanical energy (page 512). Like his first electric motor, however, it was far too rudimentary to entitle him to any credit except for the bare discovery of the principle. Even his disk generator (Fig. 427), the reverse of Barlow's wheel of eight years earlier, and a wire rectangle in which he later produced currents by rotation in merely the earth's magnetic field fell short of being real generators, largely because Faraday characteristically declined to interest himself in designing a commercial machine. Anticipatory of a variety of machine which had a considerable vogue for a time was the "magneto" of Hippolyte Pixii of France, made in 1832, and that of Joseph Saxton, exhibited in 1833 to the two-year-old British Association for the Advancement of Science. In each case the magnetic field was furnished by permanent magnets. In Pixii's machine the magnets rotated and the armature was fixed, whereas in Saxton's the opposite was the case.

One difficulty with all the early machines was that they produced alternating current which, unlike the situation today, was almost useless. A device for converting this alternating to direct current, now termed the *commutator*, was first constructed in 1834 by William Sturgeon, who called it by the descriptive though cumbersome name of "unio-directive discharger."[1] This "commutated" the current as illustrated in Figure 430, by periodically re-reversing the current by means of a rotating contact in such a way as to maintain the successive surges in one direction. In 1841 the surges, still objectionable, even though rectified, were smoothed out by a multipolar armature devised by Wheatstone of "bridge" and telegraph fame. This produced *continuous* direct current, substantially as from a battery, but generators were still very inefficient. It was not until 1845 that the superiority of electromagnets over permanent magnets for the production of the required field became generally recognized in spite of the fact that electromagnets had displaced permanent magnets in motors more than ten years before. Notwithstanding this, the famous "Alliance" generator, using 240 permanent magnets, was in common use as a source of electric current in French lighthouses in the late sixties.

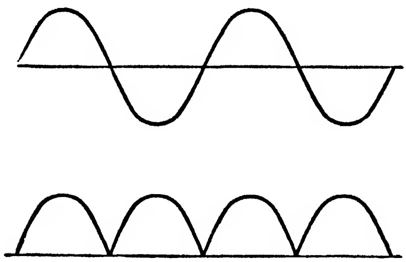


FIG. 430. ALTERNATING AND COMMUTATED CURRENT

Types of Field Winding

The advent of electromagnets brought a new problem to designers of electric generators, that of the source of current to energize the field mag-

[1] *Philosophical Magazine*, 7, 230 (1835).

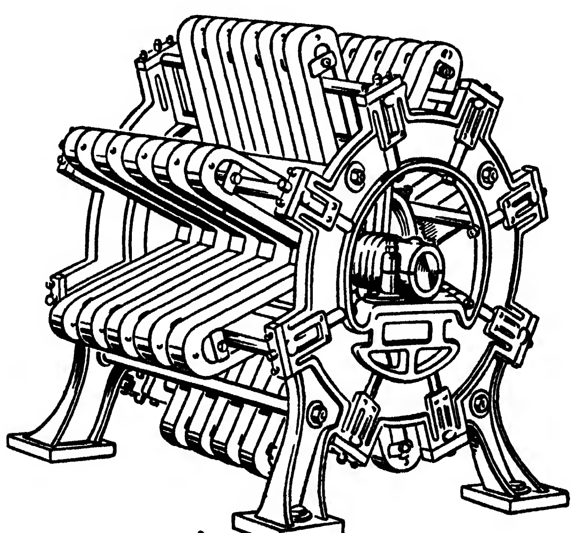


FIG. 431. THE ALLIANCE GENERATOR OF THE LATE SIXTIES

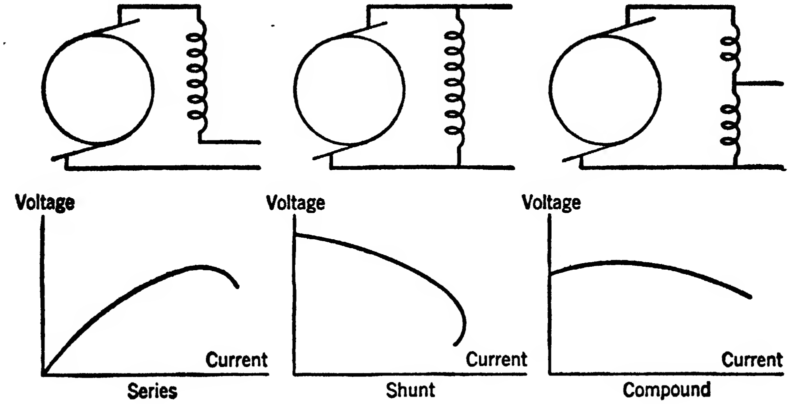


FIG. 432. TYPES OF FIELD WINDING AND THEIR PERFORMANCE

nets. At first batteries were used. The possibility of utilizing a part of the output of the generator to energize its own magnets evidently did not occur to anyone until 1851 (122:*1*:10), and nearly twenty years more elapsed before self-excitation became a common feature in the design of direct-current generators.[1] Self-excitation, in turn, raised a problem: how to connect the field coils to the armature. The simplest way, in series, had the grave disadvantage that unless considerable current were flowing, the magnetization of the field would be weak and the voltage too low. The immediate alternative, connecting the field coil directly across the terminals of the machine, had the opposite disadvantage that the voltage would be a maximum when no current was being delivered away from the machine and would become steadily less as the external current increased, in consequence of the operation of Lenz's law. Thus, though the series- and shunt-wound motor seemed each to have its place, neither the series- nor shunt-wound generator was acceptable as a sufficiently constant source of electric power.

The problem was finally solved, though not completely until after 1880, by combining shunt and series windings on the field coils in such proportions that the fall in voltage characteristic of the shunt winding was measurably compensated by a corresponding rise due to the increasing current in the series winding, thus keeping the net voltage substantially uniform over the full range of operation. Such a generator was said to possess *compound-wound* field magnets.

Induced E.M.F.

The working principle of all generators consists of moving a conductor across a magnetic field. The induction of electromotive forces by relative motion of conductors and magnetic fields has already been encountered

[1] *Transactions of the American Institute of Electric Engineers, 10,* 166 (1893).

in Chapter 42, though the relative motion was produced in a different way. In this case also, however, the e.m.f. induced is proportional to the rate at which a magnetic flux is cut by a conductor. Algebraically

$$E = \frac{\Delta\phi}{\Delta t}, \tag{1}$$

which is simply a repetition of equation (3) of Chapter 42, E being the induced e.m.f. in volts and $\Delta\phi/\Delta t$ the rate at which a magnetic flux is cut by a moving conductor.

A particularly simple case of induced e.m.f. is illustrated in Figure 433. A wire CD perpendicular to a magnetic field of flux density B webers/m^2 is moved in the direction indicated by the arrow. If the width of the field spanned by the wire is l meters and the wire moves v meters per second, the area of the magnetic field across which it sweeps in a second is lv and the rate at which magnetic flux is cut by the wire in consequence is

$$\frac{\Delta\phi}{\Delta t} = Blv. \tag{2}$$

Therefore, the e.m.f. in volts induced in the moving wire during this interval is, by substitution of equation (2) in equation (1),

$$E = Blv. \tag{3}$$

This e.m.f. may be communicated to other points through the horizontal wires on which CD slides and, provided there is adequate electrical contact, a current can flow, the magnitude of which will be determined in accordance with Ohm's law by the total resistance in the circuit.

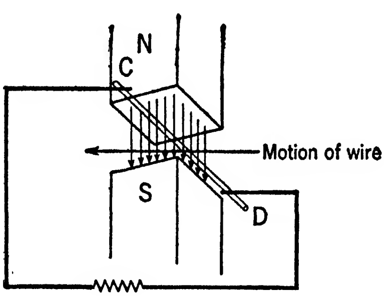


FIG. 433. INDUCTION BY MOTION OF A STRAIGHT WIRE ACROSS A MAGNETIC FIELD

Equation (3) is the basic equation of the generator. Though no actual generator is ever like the one indicated by Figure 433, all generators embody that principle, and equation (3) may be made to describe their operation by proper interpretation in light of the details of design of the particular generator to which it is applied.

Alternating Current

Up to this time the entire objective in the design of generators had been to produce steady direct currents with greater power and with less expense than could be expected from batteries. No sooner was this problem solved than another appeared which seemed to render of little avail all previous progress in generator design. It arose out of the fact, not previously realized, that efficient distribution of electric power over wide areas called for alternating current instead of direct. The reason for this has already been pointed out (pages 513 ff.). It now became necessary to develop alternating-current generators to the degree of perfection already reached by direct current. The subject is too technical to justify the presentation here of any save a few elementary points.

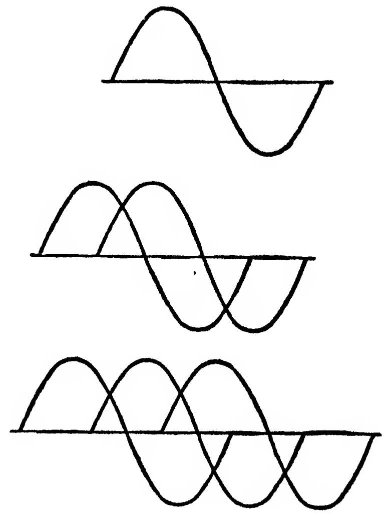


FIG. 434. SINGLE-PHASE, TWO-PHASE, AND THREE-PHASE CURRENTS

The simple generator, prior to Sturgeon's introduction of the commutator, generated alternating current. But more was involved than simply perfecting and enlarging this device. A simple alternating current was not satisfactory in the operation of A.C. motors, roughly for the same reason that a single-cylinder engine is not satisfactory for an automobile. Two separate alternating currents, out of step by 90°, or better three currents mutually out of step by 120° were better. Such double- and triple-barreled sources of alternating currents are termed *two-phase* and *three-phase* currents respectively. Practically, two-phase alternating current is seldom encountered any more, its distribution being less efficient than that of three-phase currents for reasons which will not be developed here.

But since two-phase circuits are simpler than three-phase circuits, the former can be used to illustrate the principles common to both. In this way the armature of a two-phase generator is represented in Figure 435.

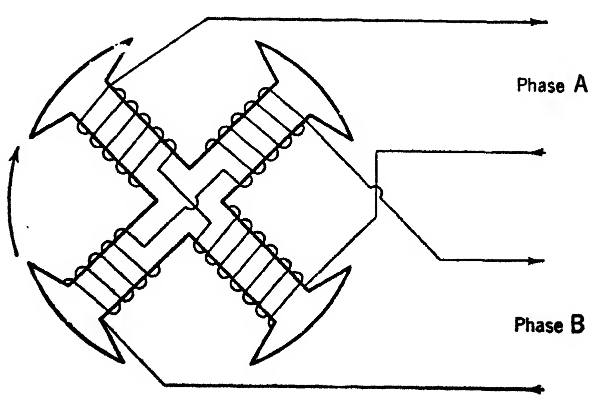


FIG. 435. THE ARMATURE OF A TWO-PHASE GENERATOR

It is supposed to be rotating in a magnetic field produced by a pair of electromagnets not shown.[1] When the current has its maximum value in phase A, it will be zero in phase B and vice versa.

A separate source of direct current is necessary to energize the field magnets of an A.C. generator, since it is not possible, as in D.C. generators, to utilize the output of a machine to energize its own magnets. No better way has been devised than to provide a separate D.C. generator for each A.C. machine. The separate D.C. generator is termed an *exciter* and is usually built into the A.C. machine as an integral part of its structure.

The Relative Merits of Alternating and Direct Current

All of the great power plants of the world now generate alternating current exclusively, and most of the small ones. Business sections of the large cities are usually supplied with direct current, their requirements being dominated by the greater versatility of D.C. motors, especially for elevator service. Most of the large office and store buildings have their own plants, so that the problem of distribution does not present itself. For similar reasons passenger ships and naval vessels usually use direct current. But both on land and at sea, the wide use of alternating current is encroaching on the domain of direct current. It is interesting to observe, however, the appearance of a possibility that high-potential direct current may come into use to reduce the inductive line-losses characteristic of A.C. distribution. But the prominent part that has been played by electric power in the last fifty years in revolutionizing the prevailing modes of living is in very large measure due to the adaptability of alternating current to economical transmission over large areas. Without alternating current we should still be living much as people did in 1890 as far as availability of commercial electric power is concerned.

Synchronous Alternating-Current Motors

This chapter began with an account of the development of the direct-current motor. It will end with a brief description of the three principal types of alternating-current motor. The first is simply the alternating-current generator reversed, termed the *synchronous* motor, a descriptive name, as will presently appear. Figure 435 will serve to illustrate the single-phase and polyphase synchronous motor as well as the generator. With its field magnets energized by direct current from some external source and the armature in motion, supplied by alternating current, it will be evident that if the armature passes a field-pole before a reversal of currents takes place, the continuing attraction will retard it. It is perhaps not so evident that if it fails quite to reach the field-pole before a

[1] In actual practice the field magnets of an A.C. generator almost always rotate inside a stationary armature, instead of vice versa as here indicated. The essential principle of any generator is *relative* motion of conductors and magnetic fields so that the question of which shall be the moving and which the fixed element is secondary, to be decided by engineering expediency.

reversal occurs, the inductive effect of the reversal will hurry it along. Thus the armature is forced into step with the frequency of the alternating current supplying it, regardless of how lightly or how heavily it is loaded. It is this characteristic which gives it the name "synchronous." Such a motor is not self-starting, and requires an external source of D.C. power for its field magnets, hence it has its limitations. It is very efficient however, and the absolute constancy of its speed is for many purposes a great advantage. Hence it is used in large factories where direct current is available in addition to alternating and where means for starting the motor are not difficult to effect. Sometimes the exciter, acting as a D.C. motor, is used to bring the synchronous motor up to speed and then, by changing connections, is made to resume its normal function of a D.C. generator furnishing current to the field magnets of the synchronous motor now running on alternating current. Tiny synchronous motors, which may or may not have self-starting attachments and which do not require a supply of direct current, constitute the motive power of electric clocks. If such clocks are to keep time, the central-station A.C. generators supplying the current must "keep time" with the degree of precision required of the clocks. Special devices, controlled by standard clocks and radio time signals, bring this condition about.

Induction Motors

The second type of A.C. motor is the ***induction*** motor, a name as descriptive as was "synchronous" for the preceding. It is in principle a polyphase motor, though with a certain diminution of efficiency it can be designed to perform on single-phase currents. It depends for its action upon the existence of a rotating magnetic field. How a polyphase alternating current may be made to produce a rotating field will be described shortly. For the present assume such a field to exist, rotating as shown in Figure 437, and consider the effect on the copper bars of a "squirrel-cage" ***rotor*** of the type indicated in Figure 436 (the moving element in an induction motor). When this rotor is placed in the rotating field, induction comes into play in consequence of the motion of the field across the copper bars. The bars being welded to a copper ring at each end, the resistance is low and the induced current correspondingly high. But a current experiences a force in a magnetic field. The forces on the bars on opposite sides of the rotor are oppositely directed and the rotor will consequently turn about its axis. A double application of the right-hand screw rule (page 442) will show that the torque will cause the coil to rotate in the direction that the field is rotating. This, in brief, is the principle of the induction motor, so called because the current in the rotor is induced, there being no electrical connection between that element and any outside source of current.

The production of a rotating magnetic field by two-phase alternating current depends basically on the fact that two simultaneous harmonic

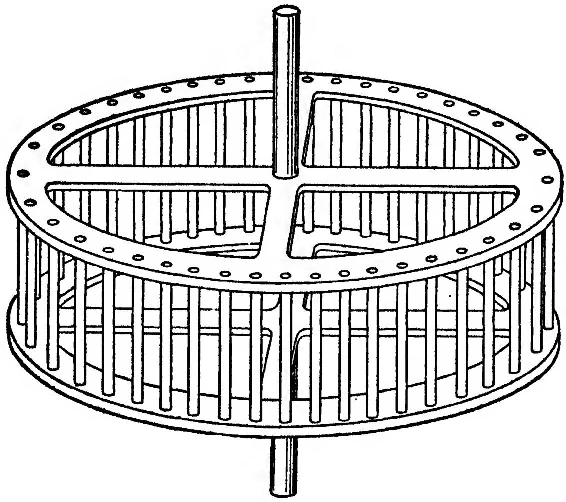


FIG. 436. "SQUIRREL-CAGE" ROTOR

motions of the same amplitude and period, one quarter cycle out of phase with each other, compounded at right angles, will produce circular motion. If phase *A* actuates one pair of magnets in Figure 437 and phase *B* the other, this condition will be realized. Suppose that at the instant when the *A* magnets are traversed by the maximum current and the *B* magnets carry zero current, a compass needle between the magnets points up and to the left as shown in (*a*). An eighth of a cycle later the current is equal in both pairs of magnets (decreasing in the *A* and increasing in the *B* magnets), and under their joint effect the needle assumes a vertical position (*b*). After another eighth of a cycle the current in the *A* magnets has fallen to zero, and the needle, now under the sole influence of the current in the *B* magnets, now at its maximum value, points up and to the right (*c*). The process continues until at the end of one complete cycle the compass needle will have returned to its initial position. So the needle rotates continuously; in other words the two-phase alternating current produces a rotating magnet field. With three electromagnets (or any multiple of three) the same effect may be produced by three-phase alternating current. The production of a rotating field by a single-phase alternating current involves some engineering tricks-of-the-trade which, while simple, need scarcely be described here.

The Universal *Motor*

The third and last of the basic A.C. motors is called the *universal* motor, because it may be run on either alternating or direct current. It is a modified D.C. motor whose operation on alternating current depends on the fact that reversal of current in *both* armature and field does not reverse the

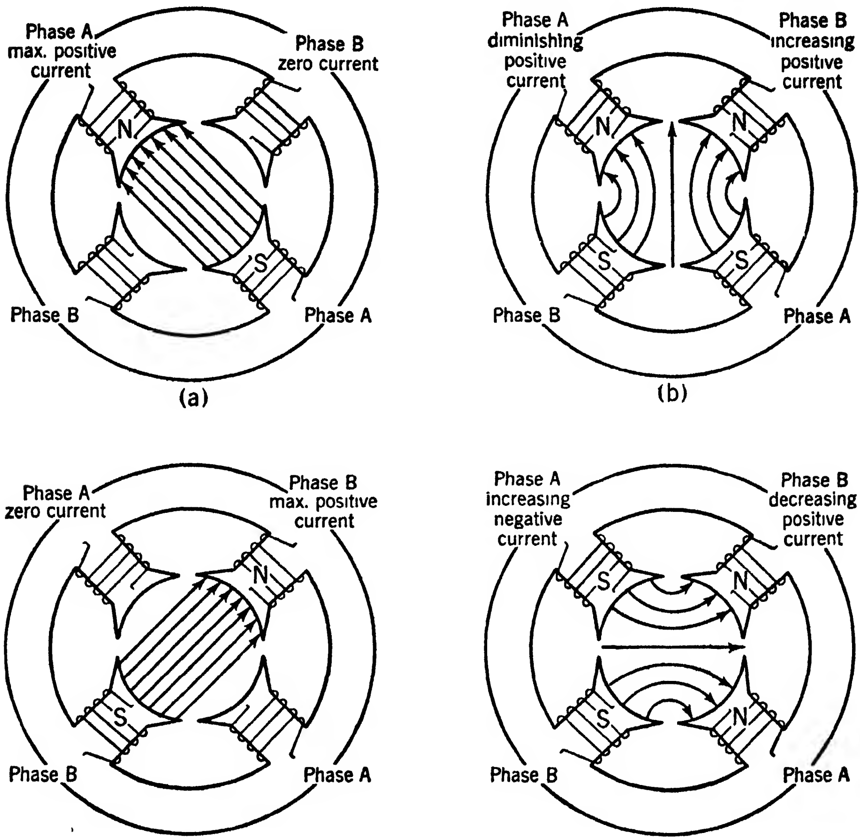


FIG. 437. PRODUCTION OF A ROTATING MAGNETIC FIELD BY A TWO-PHASE ALTERNATING CURRENT

direction of rotation. The universal motor is inefficient and by reason of sparking at the commutator produces so much radio interference that it is rapidly going out of use.

Until recently, alternating-current motors could not compare with direct-current motors in efficiency, ease of control, and general versatility of operation. It is this circumstance that has led to the retention of direct current in the business and industrial sections of large cities. But improvements in A.C. motor design have changed this condition, so that the time is now in sight when alternating current will replace direct current completely for power purposes.

Questions for Self-Examination

1. State the principle of the dynamo and cite the earliest examples.
2. Sketch the main points in the early development of the D.C. electric motor.
3. Why does the current flowing into a D.C. motor increase as increasing load diminishes the speed?
4. Tell how counter-electromotive force adapts the operation of series-wound motors to fluctuating loads and facilitates speed control of shunt-wound motors.
5. Outline the development of the D.C. generator.
6. Draw wiring diagrams of series-, shunt- and compound-wound generators and describe the effect of the respective types of winding.
7. Upon what factors does the voltage of a generator depend?
8. Write a short exposition on two-phase (or three-phase) alternating currents and show how such currents can be made to produce a rotating magnetic field.
9. Describe the operation of one type of A.C. motor.

Problems on Chapter 44

1. Two parallel guide wires are l meters apart. A cross-wire in contact with them has a speed of v meters per second. (See Fig. 427.) A magnetic field perpendicular to the plane of the cross-wires has a flux density B webers/m². The circuit to which the guide wires connect has a resistance R ohms. How many amperes I flow in the circuit? What is the backward "drag" F on the wire in newtons?

l	v	B	R	I	F
.30	.2	.1	.01	.60	.0180
.25	.3	.09	.02	.34	.0076
.2	.4	.08	.03	.21	.0034
.15	.5	.07	.04	.13	.0014

2. A copper disk of radius r centimeters rotates with its axis parallel to a magnetic field of flux density B webers/m². (See the Faraday disk, page 532.) Its angular velocity is n revolutions per second. What is the potential difference E between the center and the edge in volts?

r	B	n	E
10	.5	5	.016
10	.3	10	.019
10	.2	20	.025
10	.1	30	.019

3. The windings on the armature of a certain generator are parallel to the shaft, r centimeters from it, and l centimeters long. The magnetic field in which they move is radial and has a flux density B webers/m². The armature turns n revolutions per second. What e.m.f. E in volts is developed if the windings are all connected in parallel?

r	l	B	n	E
100	120	3.	5	110
80	100	2.5	10	130
60	80	2.	20	120
40	50	1.5	30	57

CHAPTER 45

Alternating Currents

The Nature of Alternating Current

Some of the attributes of alternating current have been mentioned as an incidental feature of the treatment of the subjects of induction and dynamos. The all-but-universal prevalence of alternating current for illumination and other applications of electric power indicates the advisability of a systematic treatment of a few of its rudiments.

Until now considerable use has been made of an analogy between water systems and corresponding systems of electric distribution. The utility of the analogy becomes less when alternating current is encountered. A water system in which the flow was alternately in one direction and then the other would seem at first thought to be quite fantastic. The picture of water alternately emerging from a faucet and then re-entering is somewhat misleading, however. We do not run electricity into a pail and carry it away, even in direct-current practice. If the principal use made of a water system were the heat developed by the friction incident to the flow — which is somewhat the condition in the electrical parallel — the direction of the flow would be of slight importance, and such a water system might have an analogy value for present purposes.

Effective Values of Current

Naturally, some of the concepts which have been formed to deal with direct current will require modification when applied to alternating current. For example, with the current fluctuating rapidly from a maximum in one direction through zero to a maximum in the opposite direction with indefinite repetition, what is the significance of the steady deflection that may be seen on an A.C. ammeter connected into the circuit? The instrument can scarcely indicate an average value in the ordinary sense of the term, for the linear average of a fluctuating quantity that consists of regularly repeated equal and opposite surges is zero. The difficulty, largely a conceptual one, is surmounted simply by defining the *effective* alternating current of, say, one ampere, as that alternating current which produces the same heating effect as one ampere of direct current. The instantaneous values of the alternating current will successively take all values, positive and negative, between zero and a maximum considerably

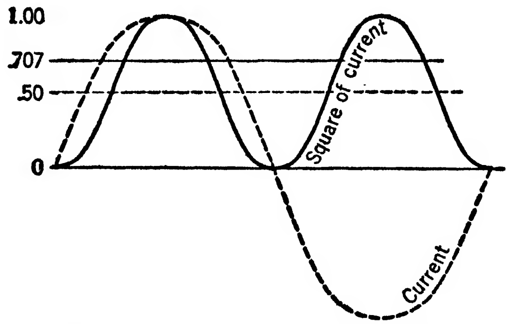


FIG. 438. DEDUCTION OF EFFECTIVE VALUE OF AN ALTERNATING CURRENT

in excess of the effective value. Since heat is evolved at a rate proportional to the square of the current, the effective value of an alternating current really involves striking a linear average of the squares of all the instantaneous values of the current. The square root of this is the effective value of the alternating current, sometimes given the awkward name of the *root mean square* value.

The process of arriving at the effective value of an alternating current having an instantaneous maximum value of one ampere is illustrated in Figure 438. The dotted curve represents the alternating current. The full curved line represents the squares of the instantaneous values outlined by the dotted curve. The horizontal dotted line represents the linear average of these squares, and the horizontal full line represents the effective value. The effective value, .707 ampere (or $\frac{1}{\sqrt{2}}$), is seen to be produced by an alternating current whose maximum values, positive and negative, are one ampere.[1] The same ratio which has here been noted for the current curve applies to the voltage curve. Thus the effective voltage of 110 — nominally furnished to commercial alternating-current accessories — really reaches maximum values of $110 \times \sqrt{2}$ or 156 volts. It is partly for this reason that alternating current is more uncomfortable to "take" than direct current of the same nominal voltage, and that the problem of insulation is more exacting.

On page 449 the utility of the dynamometer-type instrument as an A.C. meter was pointed out. It may now be observed that this utility is even more pronounced than it was possible to specify at that point, because it is the only kind of instrument which may be calibrated by direct current and then used with complete accuracy to measure alternating currents. This is true because, as was also mentioned on page 450, the deflection of a dynamometer-type instrument is proportional to the square of the current. Hence, the "average" deflection which an alternating current produces

[1] This is true only if the alternating-current curve follows a simple harmonic pattern. This condition is usually realized to a first approximation and is assumed to hold throughout this discussion of alternating current.

in such an instrument has the same proportion of the maximum instantaneous values — $\frac{1}{\sqrt{2}}$ — already noted as the defining condition of effective values. Though there are other types of A.C. instruments, they must all be standardized by comparison with dynamometer-type instruments.

Phase Displacement in an Inductance and in a Resistance

The obstruction which an inductance provides to any change of current in the circuit containing the inductance has already been commented upon (pages 516 ff.). The form which this obstruction takes in the case of an alternating voltage may most easily be understood by invoking once more the parallel between inductance and mass. Imagine a well-lubricated heavy cylinder oscillating in a cylindrical guide under an harmonic force (Fig. 439). In the resulting harmonic motion of the cylinder there will be found the usual quarter-period phase displacement between velocity and acceleration and, hence, between the velocity and the applied force. At the instant that the cylinder is at the extreme left, its velocity, of course, is zero, and its acceleration is directed to the right with a maximum value. Hence, the applied force possesses its maximum positive value. A quarter period later the cylinder is passing the mid-point, the velocity is at its maximum positive value, and the applied force has fallen to zero. After another quarter period the cylinder is at the extreme right, its velocity is zero, and the applied force possesses a maximum negative value.

In the corresponding electrical case let the current be compared to the velocity and the applied voltage to the force. The inductance in the circuit may be considered the analogue of the mass (see page 517). It will then be evident that the corresponding electrical events can be represented as in Figure 440. It would be correct and it is indeed common to say for this case that the current lags behind the voltage by a quarter of a period.

Contrast this with the case of a circuit which contains only resistance, in which the maximum instantaneous value of the current would coincide with the corresponding value of the voltage (Fig. 441). The mechanical

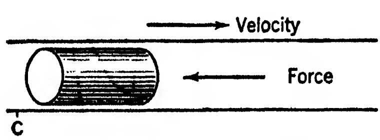


FIG. 439. MECHANICAL ANALOGUE TO INDUCTANCE

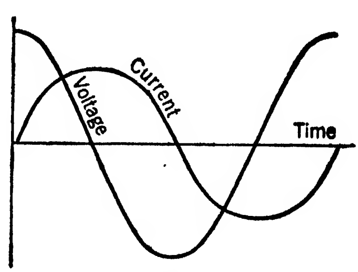


FIG. 440. QUARTER-PERIOD PHASE DISPLACEMENT BETWEEN ALTERNATING VOLTAGE AND RESULTING CURRENT IN A CIRCUIT CONTAINING ONLY INDUCTANCE

analogy to this would be realized by substituting for the arrangement of Figure 439 a loose weightless piston working in a cylinder of oil — the conventional dashpot used in door checks, automobile shock-absorbers, and similar mechanical devices. For this case an alternating force would produce an oscillation of the piston strictly in phase with it, the instants of high and low speed coinciding with those of greater and less force.

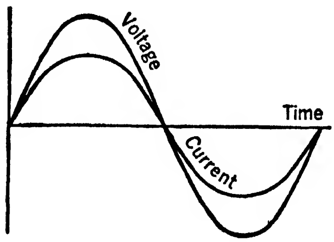


FIG. 441. ZERO PHASE DISPLACEMENT BETWEEN ALTERNATING VOLTAGE AND RESULTING CURRENT IN A CIRCUIT CONTAINING ONLY RESISTANCE

Power Consumption in A.C. Circuits of Pure Resistance or Pure Inductance

In a direct current circuit carrying I amperes under a potential difference of E volts, the rate of consumption of electrical energy in watts is given by the product EI (page 452). This is also true for alternating current, both as to instantaneous and effective values, provided that the circuit contains *only resistance*. In Figure 442 this case is represented graphically. The two full lines represent respectively the voltage and current, the dotted line being the product of the two, plotted point by point. The product, EI, is positive at all times, the effective value being one-half the maximum instantaneous value.

The case of pure resistance is somewhat unusual, however, since electrical circuits which do not contain inductance and capacitance are rare. Most systems distributing commercial electrical power involve primarily inductance in addition to resistance.

The voltage, current, and power curves are shown in Figure 443 for the

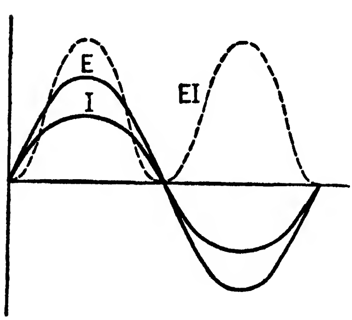


FIG. 442. INSTANTANEOUS POWER CONSUMPTION FOR ALTERNATING CURRENT WITH RESISTANCE ALONE

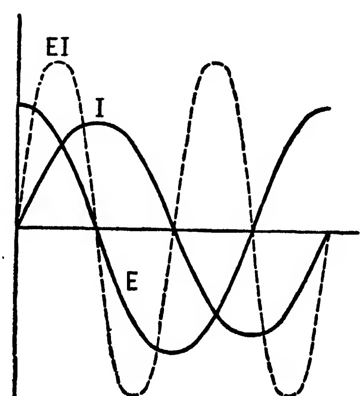


FIG. 443. INSTANTANEOUS POWER CONSUMPTION FOR ALTERNATING CURRENT WITH INDUCTANCE ALONE

case of pure inductance, the resistance being zero. The power curve is positive wherever the voltage and current curves have the same sign and is negative wherever they have opposite signs. Negative values mean merely that the system is returning energy to the source instead of receiving it from the source. The symmetry of the positive and negative loops indicates that as much energy returns to the source as comes from it. This electrical case is comparable to the mechanical case of the pendulum swinging without friction or other resistance.

This sometimes produces a rather paradoxical situation. If an alternating-current generator in a power station could be connected to an electrical system consisting of pure inductance, its instruments might indicate that it was producing the maximum current that it was capable of generating, at its full rated voltage, and yet be delivering no power whatever. A wattmeter connected to the circuit would, in fact, show a zero reading. This is the real basis for the fact, presented on page 517, that the primary winding of an unloaded transformer consumes no measurable electrical energy. It consists of almost pure inductance. Such current as flows lags behind the voltage by a quarter period and is what the electrical engineer terms "wattless current." For this reason, too, A.C. generators are not rated in watts or kilowatts as are D.C. generators, but in volt-amperes or kilovolt-amperes. The actual power that A.C. generators can deliver depends upon the character of the circuit as well as upon the size and design of the machine. One of the problems of the electrical engineer is to reduce the "wattless current" to its smallest possible value.

Inductive Reactance and Impedance

The observation has already been made (page 517) that inductance provides an obstruction to the flow of alternating current, entirely aside from that furnished by resistance. The magnitude of this obstruction, termed the inductive *reactance*, is proportional to the magnitude of the inductance, as would seem natural. It is also proportional to the number of alternations per second (the *frequency*). Call X the reactance measured in ohms like resistance, L the inductance measured in henrys, and f the frequency measured in cycles per second. Then the relation is

$$X_L = 2\,\pi f L. \qquad (1)$$

Thus the inductive reactance depends upon both inductance and frequency; it increases as either is increased and *vice versa*. A form of what might be termed Ohm's law for alternating currents then applies, namely,

$$E = IX, \qquad (2)$$

assuming the resistance of the circuit to be zero or negligible.

If the resistance is appreciable, however, equation (2) cannot be used. For the reactance X another term must be substituted which includes the resistance in the circuit. Though resistance and reactance are both measured in ohms, the total obstruction which they jointly present to the flow of an alternating current is not the sum of the two, but is instead the square root of the sum of their squares. The total obstruction is termed the *impedance*, Z (also measured in ohms). It thus has the value

$$Z = \sqrt{R^2 + X_L^2}. \tag{3}$$

The corresponding form of the A.C. version of Ohm's law is then

$$E = IZ. \tag{4}$$

Power Consumption in A.C. Circuits of Resistance and Inductance

As might be expected, A.C. power consumption in circuits containing both resistance and inductive reactance is an intermediate case between the two foregoing cases. It is, indeed, the practical common case, for resistance is seldom so small as to be entirely negligible; hence, the case of pure inductance, though a convenient pedagogical artifice, is mostly imaginary.

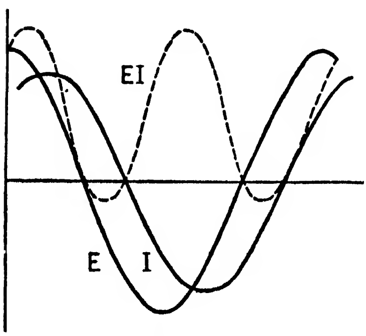


FIG. 444. INSTANTANEOUS POWER CONSUMPTION FOR ALTERNATING CURRENT WITH INDUCTANCE AND RESISTANCE

For this case the current curve lags behind the voltage curve, but not by a full quarter period (Figure 444). The power curve is neither entirely positive nor equally divided between positive and negative, but intermediate between the two. The actual power delivered would be less than the product of voltage and current. The ratio of the power to this product is termed the *power factor*. That is, the power factor is the fraction

$$\frac{I^2R}{EI} = \frac{IR}{E} = \frac{IR}{IZ} = \frac{R}{Z}. \tag{5}$$

In equation (3) above, the three terms which it contains may, by virtue of the Pythagorean theorem, be represented in a right triangle as shown in Figure 445. The power factor is then by equation (5) simply the cosine of the *phase angle* θ. That is,

$$P.F. = \frac{R}{Z} = \cos\theta. \tag{6}$$

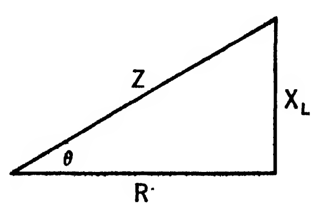


FIG. 445. ELEMENTS OF AN INDUCTIVE CIRCUIT CARRYING AN ALTERNATING CURRENT

Thus, the power factor may be computed for any circuit in which resistance and impedance are known.

Effects of Capacitance in an A.C. Circuit

But the constitution of electric circuits is not limited to resistance and inductance. Besides these two, the typical circuit contains capacitance, which may be provided explicitly by the presence of condensers as in the telephone and radio circuits, or may be implicitly present as a by-product of the character of the circuit. Examples of the latter case will be found in Chapter 46. In either event, the presence of capacitance profoundly modifies the behavior of an alternating current.

Inserting a condenser in a D.C. circuit simply stops the flow of the current. It would be quite natural to imagine that it would have the same effect on an alternating current. Surely no current can traverse the insulating medium which constitutes a part of the condenser. Yet a very simple experiment will demonstrate that a condenser, not only does not stop an alternating current, but, under certain conditions, may increase the reading of an ammeter placed in the circuit.

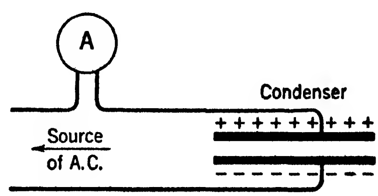


FIG. 446. TRANSMISSION OF ALTERNATING CURRENT BY A CONDENSER

It is easy to see why a condenser appears to "conduct" an alternating current. Imagine a condenser and an A.C. ammeter connected to a source of alternating current (Figure 446). During the interval that the voltage acts in one direction the condenser is charged in that direction, necessitating a flow of current to do it. Had the source been direct instead of alternating, the current would have ceased as soon as the accumulated charge on the condenser brought its voltage to an equality with that of the source. But with alternating current the voltage soon reverses, the condenser discharges, and then is charged in the opposite direction. Thus the quantity of electricity which is alternately poured into one side and the other of the con-

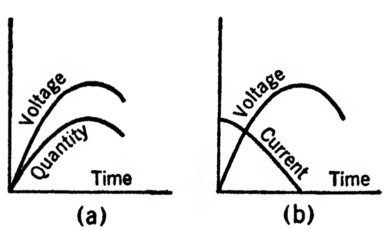


FIG. 447. WHAT HAPPENS IN A CONDENSER AS AN ALTERNATING CHARGING VOLTAGE RISES

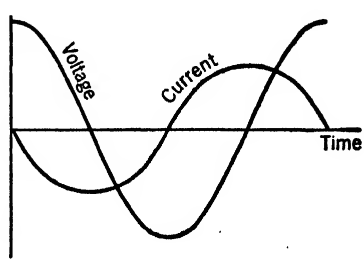


FIG. 448. QUARTER-PERIOD PHASE DISPLACEMENT BETWEEN ALTERNATING VOLTAGE AND RESULTING CURRENT IN A CIRCUIT CONTAINING ONLY CAPACITANCE

denser under the alternating voltage being applied constitutes an alternating current, even though there is no conduction occurring across the insulating medium between the two plates of the condenser.

It is now possible to examine in greater detail the current response to an alternating voltage when only a condenser is involved. Suppose such an alternating voltage to be applied to a condenser. Consider the rising portion of the voltage curve. As the voltage increases, an increasing quantity of electricity accumulates in the condenser. The rate of such accumulation, that is, the current flowing into the condenser, is proportional to the rate of change of the voltage. When the voltage approaches its maximum, its rate of change approaches zero, and the current correspondingly approaches zero. The condenser is now "full" — or at least it is as "full" as the maximum instantaneous voltage is capable of filling it — the charging current has ceased; and the diminution of voltage now due will be accompanied by a reversal of the current, beginning the process of discharging the condenser. The sequence of events up to this point is represented in Figure 447. Part (*a*) represents the quantity of electricity accumulating in the condenser as a function of time; part (*b*) the diminishing charging current. (Compare the two in light of the definition of current as rate of change of quantity.)

It is evident from part (*b*) of Figure 447 that in this case also, as in the case of pure inductance, there is a quarter-period displacement between the voltage and the current. But there is an important difference in the two cases in that *the displacement is in the opposite direction.* Contrast part (*b*) with the right-hand quarter of Figure 440, the case of an inductance. There the voltage is also rising, but the current is due to reach its maximum a quarter period *later* than the voltage maximum. Here the current maximum occurs a quarter period *earlier* than the voltage maximum. If the two curves are extended to cover a complete period as in Figure 448, the comparison between the two cases can be made to better advantage. The quarter-period displacement between voltage and current for the case of inductance has been termed a lag of the current behind the voltage. In the corresponding but opposite displacement for the case of capacitance the current is commonly and quite appropriately said to be ahead of or to *lead* the voltage by a quarter of a period.

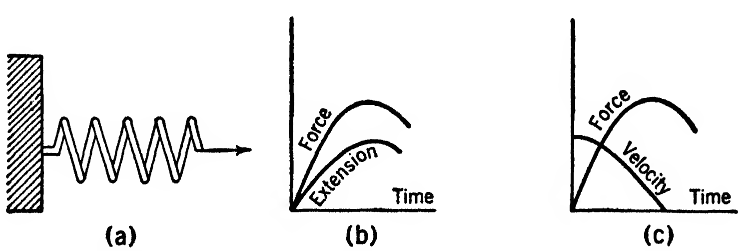


FIG. 449. THE EFFECT OF AN HARMONIC FORCE ON A SPRING

The analogous mechanical phenomenon is observable whenever an alternating force is applied to a weightless spring. Beginning at an instant of zero distortion of a spring which is being alternately extended and compressed by an harmonic force, imagine the spring to be extended by an increasing force to the right. In Figure 449, part (*b*) represents the extension of the spring as a function of time; part (*c*) the diminishing speed of the end of the motion. Comparison of these diagrams with those of Figure 447 will show the formal identity of capacitance and elasticity. The parallel is occasionally useful, just as the corresponding one between inductance and mass has been found to be.

Circuit Containing both Inductance and Capacitance

Comparison of Figure 448 with Figure 440 will show that the current responses to the same alternating voltage by a capacitance and an inductance are a full half-period out of phase. At every instant, the currents in the two cases are moving in exactly opposite directions. This fact produces some curious effects in A.C. circuits containing both inductance and capacitance. If the two are connected in series, for example, the total current response is the *difference* (in more general terms, the algebraic sum) of the responses in the two portions respectively. By proportioning the inductance and capacitance properly, the two current responses may be made equal and the total response zero. The circuit then acts in some respects as though neither were present, the current being determined solely by Ohm's law as in the case of direct current.

The possibility of thus "neutralizing" inductance by capacitance and *vice versa* has many applications. Electrical engineers reduce "wattless current," by using large condensers ("capacitors") to neutralize the surplus of inductance normally characterizing commercial systems. Telegraph and telephone engineers, on the contrary, use inductances to neutralize the surplus of capacitance normally characterizing communication systems, especially cables. This will be enlarged upon in Chapter 46.

Circuit Containing Resistance, Inductance, and Capacitance

It is now possible to deal with the general problem of A.C. circuits containing all three elements, resistance, inductance, and capacitance, in series. The case of these elements connected in parallel will not be treated. The total reactance of a series circuit will, of course, be the difference of the inductive reactance and the capacitative reactance. The latter has the value

$$X_C = \frac{1}{2\pi f C}, \tag{7}$$

corresponding to equation (1) for inductive reactance X_L. This equation shows that capacitative reactance depends upon both capacitance and frequency, but, unlike inductive reactance, decreases as either of the factors upon which it depends is increased.

The total reactance X is then

$$X = X_L - X_C. \tag{8}$$

Equation (2) then applies to the case of inductance and capacitance with zero resistance, as does equation (4) to the corresponding case where resistance is included. Equation (3), however, has to be modified to

$$Z = \sqrt{R^2 + X^2} \tag{9}$$

to allow for the effect of capacitative reactance.

Electrical Resonance

The case of equality of inductive and capacitative reactances is of particular interest. If $X_L = X_C$, then from equations (1) and (7)

$$2\,\pi f L = \frac{1}{2\,\pi f C}. \tag{10}$$

Inspection of this equation will show that if any one of the three elements — inductance, capacitance, or the frequency of the impressed alternating voltage — should be changed, the equality of the reactances which had been previously established would no longer obtain. Since the impedance of the circuit possesses its minimum value when the total reactance is zero — equation (9) — the current will be a maximum for zero reactance and less than this maximum if the reactance is rendered greater than zero. Figure 450 shows the effect on the current of changing the frequency of the impressed alternating voltage. Since for zero reactance the magnitude of the current depends solely upon the resistance, this magnitude will decrease at the so-called point of *resonance* with any increase in the resistance. This is illustrated in the figure.[1] Effects of fluctuating frequency are especially pronounced at high frequencies such as those involved in radio. Canadian and foreign broadcasting stations often lack the steadiness in frequency of carrier wave which American stations are required to maintain. Diminutions in signal strength from such stations sometimes illustrate the diminution in the alternating-current response of the receiving set with change in the frequency of the incoming signal. In such cases, the original intensity is restored by "retuning" the local set, which is usually effected by altering the capacitance to produce resonance to the new frequency.

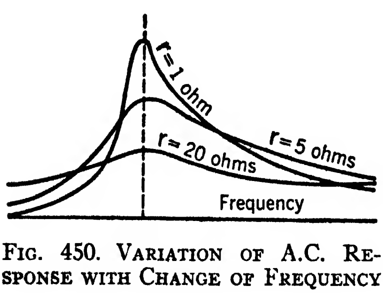


FIG. 450. VARIATION OF A.C. RESPONSE WITH CHANGE OF FREQUENCY

[1] The three curves of Figure 450 are each plotted to a different scale of ordinates. Otherwise they would not intersect each other.

If equation (10) be solved for f, the result is

$$f = \frac{1}{2\,\pi\sqrt{LC}}. \tag{11}$$

This is the same as equation (8) of Chapter 43. But that expressed the natural frequency of vibration of an oscillating circuit, whereas this is the condition that a circuit should resonate to an alternating voltage, the inductive and capacitative reactance being equal to each other. This may be recognized as another parallel between mechanical and electrical oscillations, for it is one of the best-known phenomena of mechanics that resonance occurs when the natural frequency of a vibrating body is the same as that of the impressed vibration.

Questions for Self-Examination

1. Why do alternating-current appliances have to be so designed as to withstand higher voltages than those indicated by voltmeters in the circuit? How much higher in normal operation?
2. Why does an alternating current through a pure inductance lag behind the voltage by a quarter cycle? What does this do to power consumption?
3. What does inductive reactance depend on? In what units is it expressed? How does it combine with resistance?
4. State and explain the "alternating-current" form of Ohm's law.
5. What is "power factor" and on what does it depend?
6. How can a condenser "conduct" alternating current?
7. Why does an alternating current through a condenser lead the voltage by a quarter cycle?
8. When an alternating-current circuit contains both inductance and capacitance, what relation do they bear to each other in their joint effect? Why? Give an example.
9. What happens in a circuit containing resistance, inductance, and capacitance when alternating current of steadily increasing frequency is impressed upon it?

Problems on Chapter 45

1. A power line is distributing 1200 kilowatts of power at 2400 volts. What is the current if the power factor is .8? How much power is lost in the transmission line if it has a resistance of $\frac{1}{2}$ ohm? 625 amps.; 200 kw.
2. If the voltage were stepped up by means of a transformer to 60,000 volts and distributed at that voltage, what would be the current and loss of power? 25 amps.; 310 watts.
3. An inductance of .2 henry and a capacitance of 35 microfarads are connected in series. To what frequency will the combination resonate? 71 cycles per sec.
4. An inductance of .2 henry is connected in series with a variable capacitance. To what value should the latter be adjusted to cause the combination to resonate to a frequency of 100 cycles per second? 13 m.f.
5. To the combination of problem 4 a series resistance of 10 ohms is added and the

whole connected to 100 volts at 100 cycles. What is the current when the capacitance is adjusted to 5, 12.67, and 20 microfarads in turn?

.52, 10, and 2.1 amperes.

6. A coil has an inductance of L henrys. What is its reactance X in ohms for a frequency of f cycles per second?

	L	f	X
6.	1.	25	150
	1.	60	370
	1.	500	3,100
	0.03	60	0
	5.	60	1,900
	85.	60	32,000

	L	R	Z	I
7.	.03	10	15	7.7
	.25	100	140	.82
	1.	400	550	.21
	5.	2000	2700	.043

7. An inductance of L henrys is connected in series with a resistance of R ohms. Calculate the impedance Z of the combination for a frequency of 60 cycles, and the effective current I which flows if the combination is connected to an alternating voltage of 115 volts 60 cycles.

8. An alternating current i may be expressed as a function of the maximum value I of the current, the frequency f, and the time t by the equation

$$i = I \sin 2\pi ft.$$

If $I = 10$ amperes and $f = 60$ cycles/second, find the value of i in amperes for each $\frac{1}{720}$ of a second from $t = 0$ to $t = \frac{1}{60}$ second.

$720t$	0	1	2	3	4	5	6	7	8	9	10	11	12
i	0	+5	...	+10	..	+5	0	−5	......	......	......	......	0

9. Plot the values of i obtained in problem 8 as ordinates against the corresponding values of t as abscissas.

10. An alternating voltage leads the current of problem 3 by a phase angle of $\frac{\pi}{4}$ (90°). The expression for it is $e = E \sin\left(2\pi ft + \frac{\pi}{4}\right)$. Given that E is 10 volts and f is 60 cycles, find the values of e in volts for each $\frac{1}{720}$ of a second from $t = 0$ to $t = \frac{1}{60}$ second.

$720t$	0	1	2	3	4	5	6	7	8	9	10	11	12
e	+7.1	+9.7	......	......	......	−2.6	......	......	......	......	......	+2.6	+7.1

11. Plot the values of e in problem 10 against the corresponding values of t on the same graph used in problem 9.

12. Given a condenser of capacity C microfarads, find its reactance X in ohms for a frequency of f cycles per second.

	C	f	X
12.	1.	25	6,400.
	1.	60	2,600.
	1.	500	320.
	0.03	60	84,000.
	0.03	500	10,500.
	0.03	5000	1,050.
	0.03	$5 \cdot 10^5$	10.5
	0.001	60	260,000.
	0.001	$5 \cdot 10^5$	320.
	5.	60	520.
	5.	5000	6.4
	100.	60	26.

	C	f	R	Z	I
13.	100.	60	100	103.	1.1
	5.	60	100	530.	0.22
	1.	60	100	2600.	0.044
	1.	500	100	330.	0.34
	1.	5000	100	105.	1.1
	0.03	5000	100	1050.	0.11
	0.03	$5 \cdot 10^5$	100	100.5	1.1
	0.001	$5 \cdot 10^5$	100	330.	0.34

13. A condenser of capacity C microfarads is connected in series with a resistance R ohms and connected to 115 volts A.C. (effective), f cycles per second. Find the impedance Z in ohms for the combination and the effective current I in amperes that flows.

14. A coil of inductance L henrys, a condenser of capacity C microfarads, and a resistance R ohms are connected in series. An effective voltage of 115 volts at 60 cycles is applied. Find the inductive reactance, X_L, the capacitive reactance, X_C, the total reactance, $X = X_L - X_C$, the impedance, Z, all in ohms, and the effective current, I, in amperes.

L	C	R	X_L	X_C	X	Z	I
1.	10	100	380	270	110	150	0.77
1.	5	100	380	530	− 150	180	0.64
1.4	5	100	527	530	− 3	100	1.1
0.7	10	100	264	265	− 1	100	1.1

15. An inductance of .2 henry, a capacitance of 35 microfarads, and a resistance of 10 ohms are connected in series. 100 volts A.C. is applied at a frequency of f cycles per second. Calculate the impedance and current for each of the following frequencies, 0, 20, 40, 50, 55, 60, 70, 80, and 100. Plot the current as a function of the frequency. (Compare Fig. 450.) At exactly what frequency does resonance occur?

f	0	20	40	50	55	60	65	70	80	100
I	0	.5	1.5	3.3	6.	10.	6.5	4.	2.2	1.2

16. Repeat problem 15 for a resistance of 100 ohms, and plot the current against frequency on the same graph.

f	0	20	40	50	55	60	65	70	80	100
I	0	.45	.76	.96	.99	1.	.99	.97	.91	.78

17. An electrical appliance is found to use I amperes when connected to V volts. A wattmeter indicates that W watts of power are used. What is the power factor?

I	V	W	*P.F.*
1	115	100	0.9
2	115	190	0.8
3	230	600	0.9
4	230	750	0.8

CHAPTER 46

The Telegraph and Telephone

Essential Elements in Electrical Communication

> Underlying all the electrical arts of communication — and these include not only telegraphy and telephony but also telephotography, television and sound-pictures — are a few basic principles. To appreciate those fundamentals is to possess a pass-key which opens to ready understanding the systems of communication which at first thought appear to have no essential in common. Whether the transmission medium is wire or wireless, whether the signals which convey the information are audible or visual, there is a unity to all the communication arts. However different their devices and contrivances may seem to be, all of them are merely means for performing one or another of six typical operations. Four of these operations are essential to any system of communication and the other two to all except the simplest systems.[1]

The four operations thus stated to be common to all devices for electrical communication are: (1) *generation* of some effect which can be transmitted; (2) *modulation* (or "molding") of the current which has been generated in accordance with the signals which are to be transmitted; (3) *transmission* to the point of destination; and (4) *detection*, the conversion of the modulated current into intelligible signals. The other two operations are: (5) *amplification*, which becomes necessary where great distance enfeebles the current; and (6) some process of *selection*, which is required when the transmitting medium is carrying more than one message at the same time.

The first four operations are evident in the electric telegraph of Henry. The battery generated a current. His switch, which evolved into the telegraph key of Morse and of Wheatstone, modulated this current. Wires transmitted the current, and a bell — later a recording device and still later a telegraph "sounder" — detected the modulated current. It will be illuminating to describe other agencies of communication, notably the telephone and the radio, from the standpoint of the elements common to all electrical devices for conveying intelligence.

Obstacles to Early Development

Electrical communication plays such a prominent part in modern life

[1] John Mills, *Bell Telephone Quarterly*, *14*, 13 (1935).

that it is hard to visualize the circumstances of a social order which felt no need for it and was in some ways hostile to it.

It was, to be sure, scarcely surprising that even a well-informed man should have said as early as 1837 that "the electric telegraph, if successful, would be an unmixed evil to society," but that it should have occurred to no American railroad official for seven years after Morse's patent was granted that the telegraph might be useful in dispatching trains is almost unbelievable. Perhaps it was natural for a southern Kentucky community to destroy a near-by telegraph line in 1849 on the basis that

> it robbed the air of its electricity, the rains are hendered, and ther' ain't been a good crop sence the wire was put up.

But that the United States War Department as late as 1885 should have declared officially that provision for electrical communication was not needed by the army, would pass the bounds of credibility were it not actually on record (50:487).

But, though the early lack of appreciation of the value of electrical communication was unedifying, some of the manifestations of its actual appreciation are even more so. It is a great misfortune that this inventive field has been the subject of more frequent and bitter dispute than any other. The three-cornered dispute between Morse, Wheatstone, and Henry set a precedent which was to be paralleled and even exceeded in the development of the telephone and radio. The victories in these disputes seem, regrettably, to have gone more often to the sides possessing the heaviest economic artillery than to those who could really present the best cases. While these disputes will not be of primary concern here, such evaluation of them as circumstances require will be made more from the evidence itself than from popular impressions of what that evidence indicates.

Early Telegraphy

The three major improvements in telegraphy which gave the art its present place were submarine-cable telegraphy, multiplex telegraphy, and the printing telegraph, now called the teletype. The last two were almost entirely engineering problems, that is, the adaptation of well-known scientific principles to particular requirements. The first, however, was almost as much on the scientific as on the engineering frontier. It commanded the attention of Lord Kelvin and of Michael Faraday, both being physicists of the highest standing. A considerable portion of the early difficulties experienced with transatlantic telegraphy arose in fact from the disregard by "practical" engineers of the procedures specified by these men of science.

There was, to be sure, no lack of purely engineering problems to be solved. The construction of submarine cables was made possible in the first place by the invention of gutta-percha in 1847. This, with all its shortcomings, was the only material then known that would perform satisfactorily the function of an insulating substance under such exacting conditions. The problems involved in laying such cables were also primarily within the engineering realm, and were of the first magnitude. Though a few relatively short cables had been laid during the fifties — across the English channel, between islands in the Mediterranean, and between coast cities in the United States — engineers of the sixties were for the most part without experience in cable manufacture or cable laying; and their acquirement of such experience in connection with the Atlantic cable was accompanied by several disastrous and very expensive failures. Four major unsuccessful attempts ate up millions of investors' dollars and almost discredited the whole idea of a trans-oceanic cable.

The engineering problems of laying the Atlantic cable, serious though they were, were rendered less troublesome than they might have been by the discovery in 1852–53 of a submarine plateau between Newfoundland and Ireland, the two natural termini of the main section of such a cable. Ordinarily the ocean-bottom has localities of rugged "scenery": rocky hills, deep gorges, swift submarine currents. But between Newfoundland and Ireland was found a smooth plain, such an ideal bed for a cable that it was immediately named *Telegraph Plateau.* It was deep enough to place the cable beyond the reach of anchors, icebergs, and drifts, yet shallow enough to render feasible the laying of cable. Microscopic shells, brought to the surface unbroken and unabraded, indicated the absence of destructive currents. Except for this purely fortuitous circumstance, the history of Atlantic cable-laying would have been even more checkered than it was.

But besides the engineering problems, major scientific problems were involved in the operation of a cable as long as that across the Atlantic. The necessity for sufficiently delicate instruments spurred sensitive galvanometers into being long before they would otherwise have been designed, as has already been observed (page 447). But the main difficulty with the operation of the cable was the apparent sluggishness with which signals seemed to be transmitted. It was a difficulty which had been largely unforeseen and which has been ameliorated only within the last twenty years.

The reason for this peculiarity was the fact that a cable was so long in comparison with the usual line, and its wires were so close together that a condenser of very large capacitance was produced. Transmission was slow because it was necessary to "fill" this condenser before a signal could emerge at the receiving end. Even upon emergence, the signal was so distorted by its experience that it was likely to be unrecognizable. Hence, even when conditions were at their best, the transmission of a message by transatlantic cable was found to require much more time than by the short lines in use before then.

No way of even ameliorating this handicap was devised from 1858 when it was first encountered until 1924. In that year Oliver Buckley of the Bell

Telephone Laboratories introduced into the sheath of a new transatlantic cable a layer of tape of a newly devised alloy called permalloy. The peculiar magnetic properties of this alloy (high permeability, whence the prefix "perm") had the effect of increasing the inductance of the cable. The consequent lag in the phase angle of the current constituting the signal partly neutralized the lead produced by the high capacity of the cable (see page 555), and made it possible to send messages five or six times as rapidly. There are still serious strictures on the utility of long submarine cables, however. To this day they are totally unusable for telephonic purposes.

The Telephone

In the telephone may be found a second illustration of the basic principles of electric communication. In this case a *microphone* or "transmitter" replaces the telegraph key as the modulator of the electric current. The current, instead of consisting of a succession of long and short signals according to a code, is molded to a wave-form similar to that of the sound waves incident on the transmitter. This fluctuating current is reconverted into sound by the *receiver* at the listening end of the line. The whole process is indicated in its simplest form in Figure 452.

The type of microphone commonly used in telephone practice effects

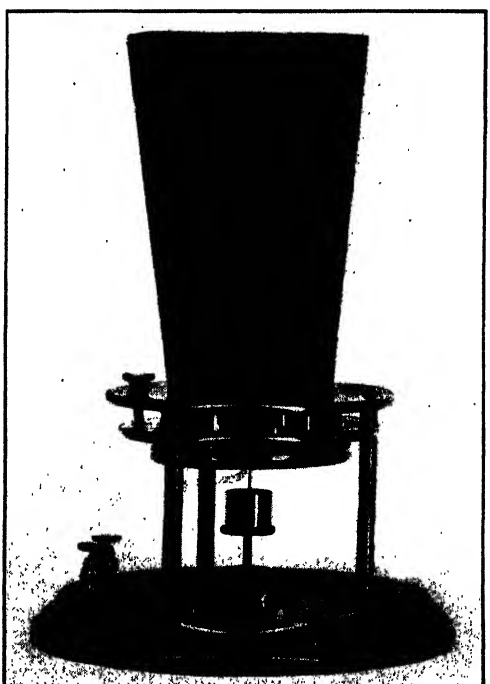


FIG. 451. THE TRANSMITTER USED BY BELL ON MARCH 10, 1876

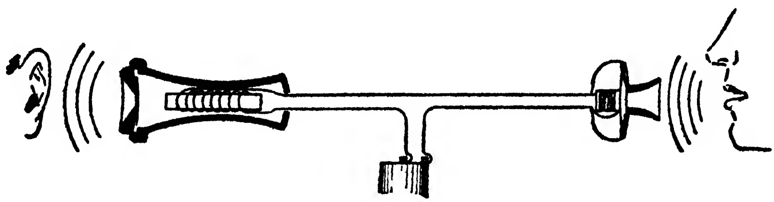


FIG. 452. A SIMPLE TELEPHONE CIRCUIT

changes in the resistance of a telephone circuit under the influence of incident sound waves, such that the resulting fluctuating current will be a passably faithful replica of the sound. The oscillation of a diaphragm alternately compresses and releases carbon granules contained in a connected pellet somewhat as idealized in Figure 453.

The receiver is essentially an electromagnet acting on a diaphragm. The incoming electric current, modulated according to the original sound pattern, sets the diaphragm in motion in a sequence of oscillations approximately similar to those executed by the diaphragm of the microphone at the other end of the line. The corresponding sound waves, upon reaching the ear of the listener, complete the function which the telephone is designed to perform.

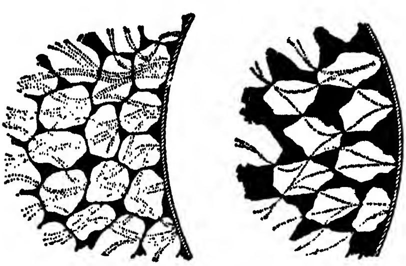


FIG. 453. ACTION OF CARBON-GRANULE MICROPHONE

(Drawing reproduced by permission of John Mills, Bell Telephone Laboratories.)

The reader will recognize a certain parallel between the telephone receiver and the electric motor. Like the motor, the receiver responds with mechanical motion to an electric current, modulated for the purpose. This fact leads one to speculate whether the reciprocal relation between motor and generator applies to this case; that is, whether, upon imparting motion to the diaphragm of a receiver, a current will be generated in the winding of the electromagnet. The discovery that this was the case was made in 1875 by Alexander Graham Bell, and receivers thus used "in reverse" were at one time the only kind of transmitter generally known.

The first well-authenticated electric telephone appears to have been devised in 1861 by Philipp Reis, a teacher of physics in Friedrichsdorf, Germany. His microphone consisted of a single loose contact between metals, and his receiver had no diaphragm other than the sounding-board to which the electromagnet was attached. But these circumstances merely reduced the efficiency of operation without affecting the principle. His telephone operated well enough to carry out Reis's main objective, which was to transmit speech. That he accomplished this has been amply confirmed by the unequivocal testimony of several reputable men of science who actually used Reis's original instruments, both under his direction and independently.[1] Though there have been many other claimants to the distinction of having devised the first successful telephone instruments, Reis presents the clearest title of them all.

The first words understood over a telephone in the United States were spoken by Alexander Graham Bell on March 10, 1876. The transmitter described and illustrated in his own patent, granted three days before, had been of the receiver type, mentioned above. But on the famous tenth of March, Bell used, not that transmitter, but one which was a precursor of the microphone type. This latter transmitter had been very completely described and illustrated in a document submitted to the patent office nearly a month before, by a competitor of Bell named Elisha Gray, about which Bell subsequently acknowledged having received information prior to March 10. Bell sedulously refrained from mentioning his now famous exploit of 1876 for more than four years, and in the meantime his company developed the receiver type of transmitter exclusively. All this indicates unmistakably what Bell thought during that time about where the credit belonged for the transmitter with which the first American telephone conversation was conducted.[2]

Later Telephone Development

The invention and perfection of the carbon-granule microphone spanned the eighties. It was first devised by an English clergyman named Hunnings, the principal improvements being by Thomas Edison and Anthony White.[3] Since that time neither the transmitter nor the receiver used in ordinary telephony has been modified in any respect except minor details. As applied to public address systems and to radio telephony, however, major modifications have been made and will be described in appropriate connections.

The lack of improvements in modulation and detection of telephone messages during the last fifty years has been simply an outgrowth of the fact that by 1890 these two operations of telephony had developed to a point fifty or more years in advance of the intervening operation, transmission. It is in this operation that all the significant developments have occurred since 1890. They have followed two main lines. The first is the

[1] Though Reis's success and even his intention to transmit speech have been vigorously denied and though the denial has been supported by court decisions in the United States, both the denial and the decisions seem to have been based on incomplete and *ex parte* testimony. The reader is referred to S. P. Thompson's *Phillip Reis, Inventor of the Telephone* (London, Spon, 1883), especially the Preface, pp. 36–38, and pp. 112 ff.

[2] For a more extended account of the issue between Bell and Gray the reader is referred to *The American Physics Teacher*, *5*, 243 (1937).

[3] For a detailed history of the development of the microphone see *The Bell Telephone Quarterly*, *10*, 164 (1931). This account, though well documented, is unfortunately seriously in error at several important points.

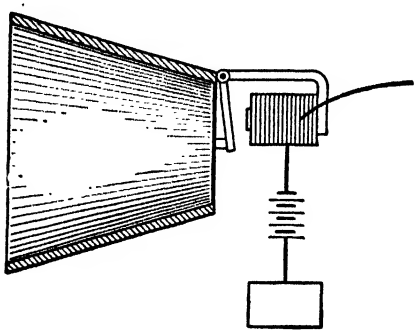


FIG. 454. THE TRANSMITTER ILLUSTRATED IN BELL'S PATENT GRANTED MARCH 7, 1876

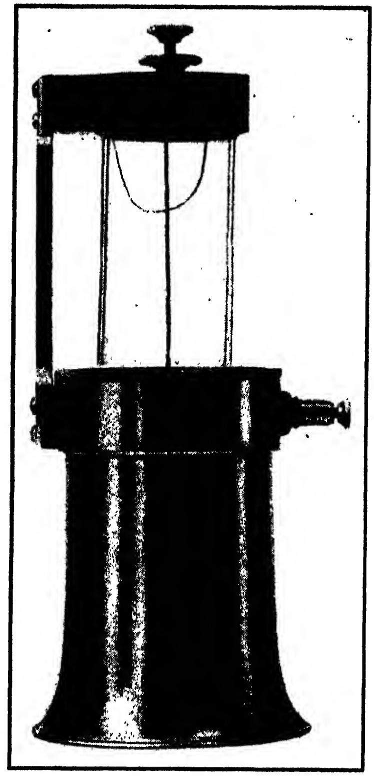


FIG. 455. THE TRANSMITTER ILLUSTRATED IN ELISHA GRAY'S CAVEAT OF FEBRUARY 14, 1876

improvement in central office equipment and organization, facilitating the control of ever-increasing numbers of calls with ever-decreasing delay and culminating in the machine-switching equipment — the prevalence of which is now indicated by the use of dial telephones in all large communities. The second is the extension of the distances over which telephone conversations could be conducted. The first has been primarily a matter of refinement of technological minutiae which need not concern us. The second, however, centers in two major scientific developments which are worthy of note: the development of the so-called *loading coil* and that of the electron-tube amplifier or *repeater*, as it is called in telephone practice.

The Extension of Telephonic Distance

The circumstance that limited the range of early telephone conversation was the same as that which had slowed up communication over submarine cables, namely, the electrostatic capacitance of the lines. The difficulty was met in the case of the cables by retarding the speed at which telegraphic messages were clicked off. No such solution was possible for the telephone problem. This was partly because there was a definite lower limit to the speed of intelligible enunciation, and partly because the frequency of telephone signals (speech sounds) was in the range of hundreds or even a few thousands per second, whereas that of telegraph messages never exceeded eight or ten characters per second.

The problem was solved in 1899 in somewhat the same way that a partial solution of the cable problem was effected many years later, namely, by neutralizing the effect of the capacitance of the cables by judicious introduction of inductances at intervals along the line. It was necessary that the distances between these inductances should not be greater than a certain fraction of the mean wave-length of the telephone messages, and that the magnitude of the inductances should be properly adjusted to produce a phase lag that should equal the phase lead which it was desired to neutralize. This difficult problem was solved by Michael I. Pupin (1858–1935), a professor in Columbia University, who in 1874 had landed in this country, a Serbian immigrant with just five cents in his pocket and not a word of English on his tongue. The effect of the cable's capacitance had been to cause voice-waves of different frequencies to travel with different speeds, thus distorting the emerging speech signals beyond the point of intelligibility. The "loading coils" which corrected this difficulty may now be seen placed in collections of iron boxes along transcontinental telephone cable lines at intervals of approximately a mile and a quarter. Pupin's solution for overland telephone cables was much more complete than Buckley's later solution of the submarine telegraph cable problem, notwithstanding the fact that the performance of telephone cables is immeasurably more exacting than that of telegraph cables. The impossibility of interpolating properly designed inductance coils at intervals along a submarine cable precludes the utilization of Pupin's discovery in that field. In 1902 public telephone service between New York and Chicago, made possible by the loading coil, was inaugurated with great ceremony. This was to mark the virtual maximum of telephone distance until the advent of the repeater in 1915, which made it possible to renew the energy of the voice currents at appropriate intervals along transcontinental lines.

The repeater is simply a special form — for telephone use — of a device more commonly known as an amplifier. By means of it, weak electrical impulses can so control a strong local source of electric output as to mold it into the same forms. Thus a telephone current, which is weak to the point of utter inaudibility, upon passing through a repeater station emerges with as great energy as at the source, or even greater. This amplification may be repeated as many times along a line as is necessary to reach a required distance.

The heart of the amplifier is a device which, in its usual forms, presents somewhat the appearance of a peculiarly shaped electric light globe. It is variously known as a thermionic vacuum tube, a three-electrode vacuum tube, or a triode. It originated in an observation by Edison in 1883 in connection with his study of electric illuminants. To investigate the cause of an asymmetrical blackening of the inside surfaces of his early lamps, he had sealed an additional electrode into the side of one of them. He observed that a small current could be made to flow across the evacuated space between the incandescent filament and this second electrode if the latter were given a positive polarity, but not otherwise. Since no possible use could be imagined for this peculiar phenomenon, Edison merely recorded it, but

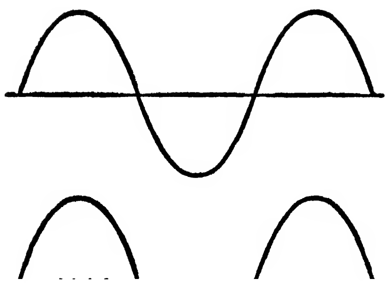


FIG. 456. A RECTIFIED ALTERNATING CURRENT WAVE FORM

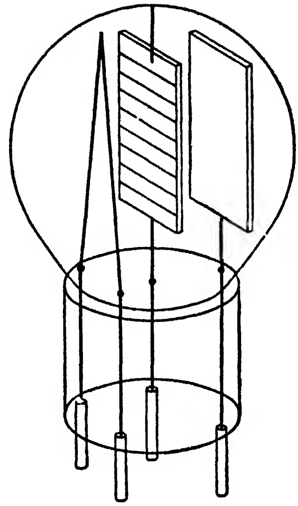


FIG. 457. THE INTERIOR ARRANGEMENT OF A THREE-ELECTRODE TUB

gave it no further attention. In 1889 an English scientist showed that the effect was due to a flow of electrons, released from the filament by its high temperature, to the second electrode when attracted by a positive charge. This accounted for the absence of a contrary flow when the polarity of the second electrode was reversed. A German investigator promptly utilized the phenomenon to "rectify" alternating current. A common modern example of this process is to be found in the "Tungar" rectifier used to charge storage batteries from A.C. power lines.

The next two stages in the development of the vacuum tube — and, except for minor improvements, the final stages — were made by two early workers in the field of wireless telegraphy. It occurred in 1904 to James A. Fleming, an Englishman who was one of Marconi's associates, that the rectifying action of the vacuum tube might solve the most troublesome problem of detection of wireless waves. His subsequent development of what is known in England to this day as the "Fleming valve" gave the vacuum tube its first real start to fame. In 1906, Lee Deforest, an American radio experimenter and promoter, not satisfied with the performance of the Edison-Fleming tube, modified it by introducing the element which, above all others, rendered the vacuum tube adaptable to the purposes it is now serving. He tried the experiment of placing between the filament and the second electrode (now called the *plate*) a tiny *grid* with bars of fine wire, through which he hoped to secure an extra measure of control over the cloud of electrons occupying that space when the tube was in operation. This step proved to be a stroke of genius, the results of which must have been a surprise even to Deforest himself.

There has probably never been another single invention that has influenced technology so profoundly. The fact that it facilitated long-distance telephony is one of its lesser outgrowths. Most of the modern attributes of radio communication are traceable to the addition of a third electrode to the vacuum tube, and the broader fields of electrical engineering are finding new uses for the three-electrode tube almost every day. It is indeed quite impossible to overestimate the ultimate importance of Deforest's contribution to vacuum tube design.

The principle of the action of the third electrode is simple. Thrust into the electron cloud which always exists between the plate and the incandescent filament, it is in a position to exercise, through its electrical potential, an intimate control over the motions of these electrons. A certain negative potential will stop the flow of electrons from the filament to the positively charged plate; a smaller negative potential will permit a limited flow; a positive potential,[1] if not too great, will cause the flow to be greater than it would have been in the absence of a grid. Thus the flow of electrons from filament to plate or, in the usual terminology, the *plate current* is intimately controlled by the *grid potential.* With the proper attention to design and operation, changes in plate current may be made proportional to changes in grid potential. Small changes in potential may thus, through the agency of a three-electrode tube, control a local source of electrical energy in such a way as to produce in it relatively large but proportional surges. This is the principle of operation of the tube used as an amplifier. If one such amplifier is insufficient, its output may constitute the input of a second tube. Though there are practical limits to the number of such stages of amplification, set by a sort of electrical analogy to the economic "law of diminishing returns," six or seven stages are common, giving amplification ratios up to many millionfold. This is the action of the telephone "repeater," more familiar examples being the action of public address systems and ordinary radio receiving sets.

Refinements of Telephony

Under certain circumstances several telephone messages may be sent simultaneously over the same wire. In the case of the so-called "coaxial cable" between New York and Philadelphia, as many as two hundred and forty two-way messages may be transmitted at one time. In such cases the technique is similar to that of the generation and reception of radio messages. The telephone transmitters each modulate an alternating current of different frequency, instead of modulating direct currents as in the case of the ordinary telephone. The frequencies of the alternating currents are so great as to be beyond the audible range; hence, the ear hears only the modulated waves and is entirely unaware of the *carrier wave.* Each of the receiving instruments is tuned to the frequency of the corresponding transmitter. The signals travel along a wire, instead of through space.

[1] As usually arranged, a triode seldom utilizes a positive potential on its grid.

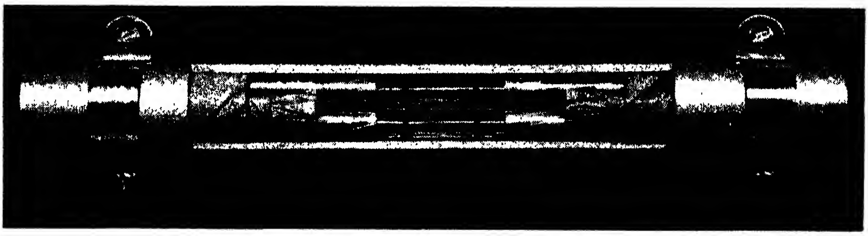


FIG. 458. A SECTION OF THE FIRST "COAXIAL CABLE" CONSISTING OF A PAIR OF WIRES EACH IN THE AXIS OF A COPPER CYLINDER
(Courtesy of Bell Telephone Company.)

The sending of photographs by wire, which has in recent years become one of the routine features of news dispatch, differs only in detail from sending telegrams or conversations by wire. The chief difference is that the electrical current is modulated by a light beam instead of by a sound wave at the sending end and is reconverted into a correspondingly fluctuating beam by the receiver. A microscopic view of a picture sent by wire shows it to consist of lines either varying in density or varying in width, to produce the required effects. These lines are produced by a process of *scanning* at the sending end. A tiny pencil of light moves across the transparency to be transmitted in successive, closely spaced parallel lines. The fluctuations in intensity of the portion of the beam that penetrates the transparency go over the line in the form of fluctuations of current. These actuate a *light valve* at the receiving end, controlling the intensity of a

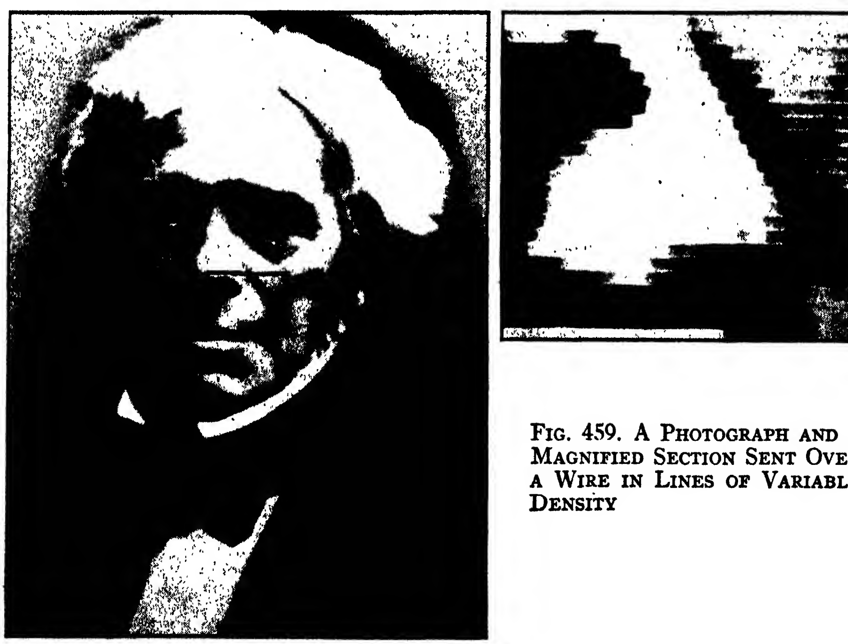


FIG. 459. A PHOTOGRAPH AND A MAGNIFIED SECTION SENT OVER A WIRE IN LINES OF VARIABLE DENSITY

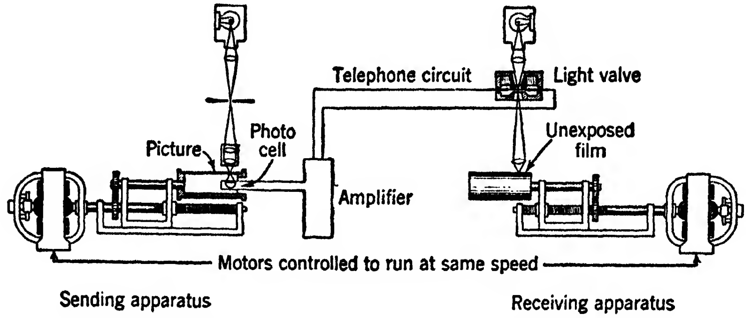


FIG. 460. A TELEPHOTOGRAPHIC SYSTEM IN OUTLINE
(Drawing reproduced by permission of Bell Telephone Company.)

pencil of light to duplicate the variations at the transmitting station. The details of how these steps are effected are not of present concern. The important point is to realize that in principle the operation of sending pictures by wire is the same as that of sending conversations by wire. The process of (1) generation, (3) transmission, (5) amplification, and (6) selection (see page 560) are identical in the two cases. Only (2) modulation and (4) detection are different because the original and final forms are light patterns instead of sound patterns. The process of broadcast television differs from that of telephotography somewhat as moving picture photography differs from ordinary still photography. Instead of there being nearly twenty minutes to send a picture, as in telephotography, a televised picture must be completed in about one-sixteenth of a second. During that brief interval the scanning, transmission, and restoration at the receiving end must be completed, for to avoid flicker, sixteen such

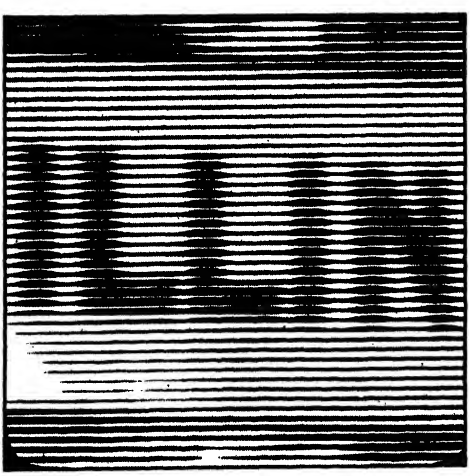


FIG. 461. A PORTION OF A TELEPHOTOGRAPH PRODUCED BY LINES OF VARIABLE WIDTH

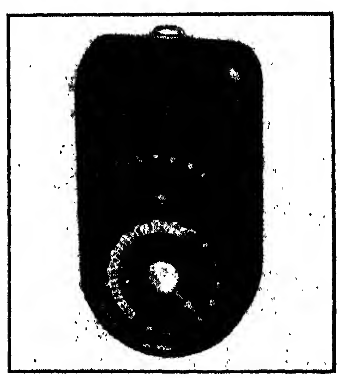


FIG. 462. A COMMON TYPE OF EXPOSURE METER
(Courtesy of Weston Electrical Company.)

pictures must succeed each other every second. The basic process is the same in both cases, however, though the mechanism involved in so enormous a speeding-up is necessarily different.

The Photoelectric Cell

Playing a central rôle in telephotography as well as in television is a device which will merit special attention by way of concluding this chapter. It is known as the *photoelectric cell.* It is the principal agent through which electric currents may be modulated by light patterns. The pencil of light which penetrates the transparency at the sending end of the line strikes a photoelectric cell which, with the aid of its accessories, acts to convert the fluctuating intensities of light into corresponding fluctuations of current. It depends on a phenomenon, first observed by Hallwachs in 1888, that under appropriate conditions electrons are liberated from metal surfaces by light, especially by ultra-violet light. The shortness of the interval between the incidence of the light and the release of electrons was astonishing, being only a few billionths of a second. Amplification of the small potentials thus created made it possible to use the photoelectric cell to effect modulation of an electric current by light. In recent years, synthetic surfaces have been constructed which are enormously more effective than metals.[1] A different type of cell, which generates currents instead of merely emitting electrons when exposed to light, is becoming familiar to photographers as the working nucleus of the "exposure meter." The current generated in such cells is caused to actuate a delicate but rugged galvanometer, so that the whole compact assembly can be used to make a rapid determination of the intensity of light. These cells are sometimes termed "photovoltaic," a very descriptive name, to distinguish them from the older type termed "photoelectric."

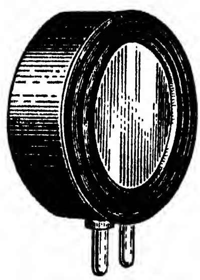


FIG. 463. A COMMON TYPE OF PHOTOELECTRIC CELL (Courtesy of Weston Electrical Company.)

Questions for Self-Examination

1. What is the principal scientific problem involved in operating long telegraph cables?
2. How does the simple modern telephone differ from that invented by Alexander Graham Bell?
3. Describe the two main stages in the development of long-distance telephony.
4. Tell how a "vacuum tube" acts as a rectifier.
5. Outline the principle of sending pictures by wire.

[1] Notably that devised by M. J. Kelley of the Bell Telephone Laboratories. See *Bell Laboratories Record*, October, 1933.

CHAPTER 47

Radio Communication

The Versatility of Radio

> Radio is the miracle of the ages. Aladdin's Lamp, the Magic Carpet, the Seven League Boots of fable and every vision that mankind has ever entertained, since the world began, of laying hold upon the attributes of the Almighty, pale into insignificance beside the accomplished fact of radio. By its magic the human voice may be projected around the earth in less time than it takes to pronounce the word "radio."
>
> The story of wireless broadcasting has ramifications as ancient as civilization because since time began men have struggled with the very riddle to which this latest triumph of human ingenuity brings solution. Blind gropings of generations of experimenters have contributed to progress along the pathway to the pinnacle from which ... science suddenly glimpsed the secret of radio broadcasting. (4:3.)

Radio is playing an ever-increasing part in human life on land, on sea, and in the air. It makes possible a great deal more than the mere communication of news, entertainment, and messages from place to place. It plays an important part in navigation. The radio beacon and the radio compass permit the pilot and mariner to locate positions in fog and storm. The radio printer delivers his weather maps daily. Without radio, travel by air would not be, as it is fast becoming, one of the safest methods of transportation. Radio instruments indicate the absolute altitude of the plane, guide the airliner from port to port, and facilitate landing, making "blind" flying possible. The pictures in our daily newspapers are transmitted by wireless. Even the geophysicist does prospecting by means of radio instruments.

In the opening paragraphs of the chapter on the telegraph and telephone six typical operations involved in various methods of communication were enumerated. Each of these six operations plays an important part in radio communication. They are important enough to warrant restatement here.

These six operations are: (1) *generation* of some effect which can be transmitted, this effect being called in radio the *carrier wave*, or carrier; (2) *modulation* of the carrier by the signal to be transmitted; (3) *transmission* to the receiver; (4) *selection* of the modulated carrier at the receiver, this process being commonly called *tuning;* (5) *detection*, or conversion of the

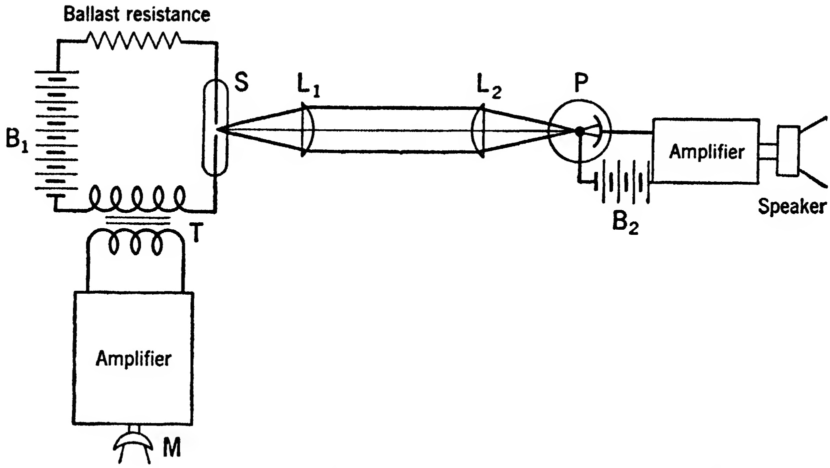


FIG. 464. THE TRANSMISSION OF SOUND OVER A LIGHT BEAM

modulated carrier into intelligible signals; and (6) *amplification* of weak electric currents.

The Transmission of Sound over a Light Beam

Light beams have occasionally been used as carriers of sound. Though such an experiment is more of a "stunt" than a scientific or utilitarian measure, it furnishes an apt illustration of some of the principles of radio. The process is shown in Figure 464. The *carrier* is a beam of light generated in S, a special light source operating on principles similar to the familiar neon lamp. In fact a small neon lamp may be used. This type of lamp has its light produced by an electrical discharge in a rarefied gas and differs from the ordinary incandescent, or filament type, lamp in that the intensity of the light emitted can be made to change very rapidly.

The battery B_1 furnishes the voltage to operate the light source S. As long as the current through this lamp remains constant the intensity of the light emitted remains constant. If, however, a fluctuating current, produced when sound strikes the microphone M, is sent through the primary of the transformer T, the fluctuating voltage produced in the secondary of this transformer is superimposed on the voltage already applied by the battery B_1 to the lamp S. This causes the current through the lamp, and hence the light radiated from it, to fluctuate in intensity in accordance with the sound vibrations striking the microphone. The light is then said to be modulated by the sound current originating in the microphone M. The light radiated from the source falls on the lens L_1 which converts it into a nearly parallel beam. The apparatus described thus far may be called the transmitter.

The modulated beam of light is directed at the lens L_2, which focuses it on the photoelectric cell P. The photoelectric cell, as explained in the preceding chapter, has the property of conducting electricity in proportion to the amount of light incident upon it. Thus, when the light beam is constant in intensity, the current through the photoelectric cell is constant, and when the light beam is modulated, the current through the cell fluctuates in accordance with the modulation. The photoelectric cell serves as the ***detector***, or converter of the modulated carrier into a current, the fluctuations of which are the same in frequency and relative intensity as those produced by the microphone at the transmitter. This fluctuating current through the cell is very feeble and must be amplified before being applied to a speaker to be converted into sound. The receiving apparatus, beginning with the second lens and ending with the speaker, is commonly referred to simply as the receiver.

Thus has been described the generation of the carrier at the light source; the modulation of the carrier by a voice current; the propagation of the carrier through space; the detection of the modulated carrier by a photoelectric cell; and amplification both in the transmitter and receiver.

If modulated beams of light from two similar transmitters were to fall simultaneously on the receiver, interference would result. This confusion could be eliminated by making the light radiated from transmitter *A* red and that from *B* blue. Then either transmitter could be selected at will, *A*

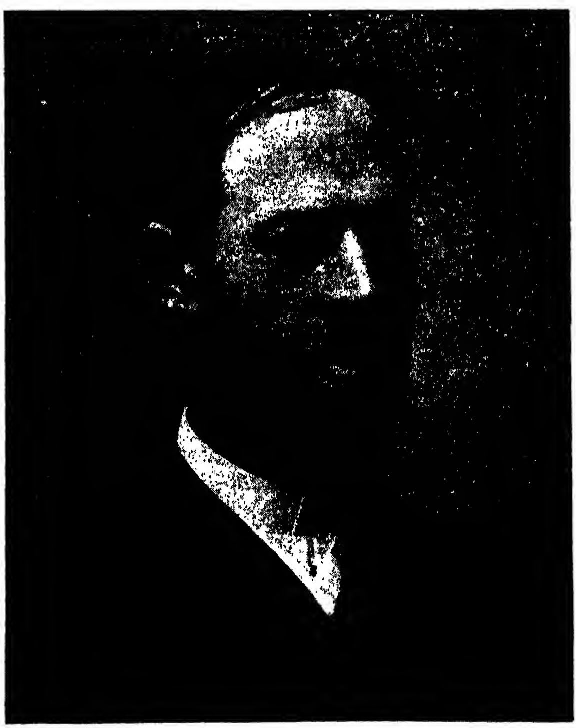


FIG. 465. HEINRICH HERTZ (1857–1894)

by placing a red transmission filter in front of the photoelectric cell, and *B* by a blue filter. This process could be termed *selection.*

The transmission of sound by a modulated light beam involves the same fundamental operations as form the basis of radio communication. Indeed, it will presently be shown that in both cases the sound is carried by modulated electromagnetic waves.

The Discovery of Electric Oscillations

Wireless communication was made possible, not through the efforts and discoveries of any one man, but by the accumulated results of the works of many. As stated in Chapter 43, Joseph Henry in 1842 discovered the existence of the electrical oscillations produced by the spark discharge of a condenser. Though the nature and properties of these oscillations were not comprehended until many years later, Henry surmised that there was some similarity between their behavior and that of light. He made the comparison as follows (70:33):

> It would appear that a single spark is sufficient to disturb perceptibly the electricity of space throughout a cube of 400 feet capacity, and ... it may be further inferred that the diffusion of motion in this case is almost comparable with that of a spark from flint and steel in the case of light.

Based on the vast amount of data that had been accumulated and the theories that had been formulated on electricity and magnetism, especially on the experimental researches of Faraday on electricity and the work of Henry, Lodge, and others on electrical oscillations, James Clerk Maxwell wrote in 1856 his famous theoretical paper, *A Dynamical Theory of the Electromagnetic Field.* In this paper he propounded the theory that if electrical waves could ever be generated, they would travel through space with the speed of light and that light was essentially an electromagnetic phenomenon. Furthermore, he pointed out that light and electrical waves should differ only in wave-length or frequency; and any difference in their behaviors should be the result of the difference in wave-lengths. We know now that gamma rays, X-rays, ultra-violet light, visible light, infrared rays, heat radiation, and radio waves are all electromagnetic in nature.

Maxwell's theoretical and highly mathematical developments leading to his electromagnetic theory were not fully appreciated until some years later. One outstanding scientist (Fitzgerald) even went so far as to publish a paper *On the Impossibility of Originating Wave Disturbances in the Ether by Means of Electric Forces.*

Hertzian Waves

The electromagnetic waves predicted by Maxwell on purely theoretical grounds and on indirect evidence were not demonstrated until the later years of his life. In 1888 a young German physicist, Heinrich Hertz (1857–94), not only produced and demonstrated these waves but also showed that they had many of the properties of light. They were subject to the laws of reflection, refraction, and interference. Subsequently it was shown that their speed was identical with the speed of light.

Hertz's experiment is shown diagrammatically in Figure 466. The left spark gap was connected to the terminals of the secondary of an induction coil. The passage of a spark produced an oscillatory discharge in space, of form similar to that shown in Figure 425, creating an electromagnetic wave which traveled out into space with the speed of light. A few feet away was placed a rectangular conductor with a small spark gap. A sliding rod *EF*, while making good contact with the other two wires, could be moved along, thus changing the electrical properties of the circuit. Hertz found that when the sliding rod was properly adjusted a spark discharge produced a small but definite spark at the second gap, thus proving that energy in the form of an electromagnetic wave did pass from the "transmitter" to the "receiver."

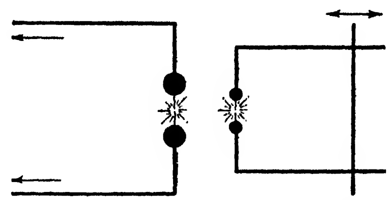


FIG. 466. THE GENERATION OF HERTZIAN WAVES

Experimentation was carried on further by Hertz and others, notably Lodge in England, Popoff in Russia, and Branly in France. Strange as it may seem, none of these men realized that they had within their grasp a means of wireless communication, a goal long sought by such men as Henry, Faraday, Morse, Bell, Trowbridge, Edison, and Preece. Sir William Crookes, in 1892, was the first to predict the use of electromagnetic waves for telegraphic communication. He even indicated the possibility of tuning to special wave-lengths.

The Birth of Wireless Communication

In the summer of 1894 a young Italian, Guglielmo Marconi, twenty years of age, chanced to read an article describing the work of Hertz, who had died earlier in the year. It was there that the young experimenter got his idea of using the radiated Hertzian waves for communication. "It seemed to me," said Marconi some time later, "that if the radiation could be increased, developed, and controlled it would be possible to signal across space for considerable distances. My chief trouble was that the idea was so elementary, so simple in its logic, that it seemed difficult for me to believe that no one else had thought to put it into practice. I argued, there must be more mature scientists who had followed the same line of thought and arrived at almost similar conclusions. From the first the idea was so real to me that I did not realize that to others the theory might appear quite fantastic."

Marconi had a vision and fulfilled it. Within two years he succeeded in sending code signals by electromagnetic waves over a sufficiently great distance to warrant taking out a patent and presenting his invention to the public. Almost immediately the cry went up from many quarters that Marconi was not the true inventor of wireless. It was claimed that all he did was to take advantage of the work done by others, merely acquiring and adapting their theories and their devices. Some commentators were more generous. One competent writer said (70:28):

> Let me say at once, to avoid misunderstanding, that without the energy, ability, and enterprise of Signor Marconi, what is now called the "wireless" would not have been established commercially, would not have covered the earth with its radio stations, and would not have taken the hold it has upon the public imagination.

Marconi remained a man of visions and dreams until the day of his death (1937), but he was neither a visionary nor a dreamer. His keen perception of the latent possibilities in the field of wireless revealed to him years before their consummation many of the developments and practices which are now considered commonplace by the scientific world.

The Reception of Radio Broadcasts

When the unmodulated carrier sent out by a radio broadcasting station falls on the antenna, or "pick-up" wire, of a receiving set, there is generated in the receiver a high frequency alternating current of form shown in Figure 467. For transmission within the so-called broadcast band the frequency of the carrier will be somewhere between 550 and 1600 kilocycles per second.

In contrast to these high frequencies, the frequencies of the "sound" currents generated by the microphone of the transmitter are usually confined to between 50 and 10,000 cycles. These are said to constitute the

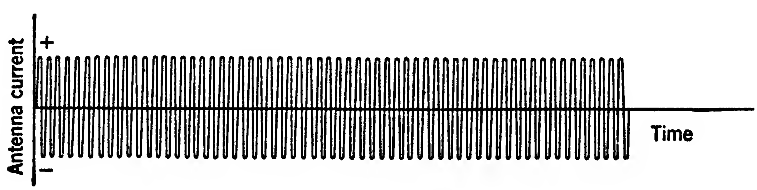


FIG. 467. UNMODULATED CARRIER CURRENT

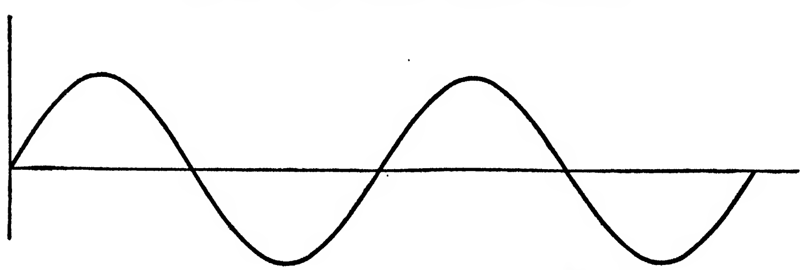


FIG. 468. AUDIO-FREQUENCY CURRENT PRODUCED IN A MICROPHONE BY A PURE TONE

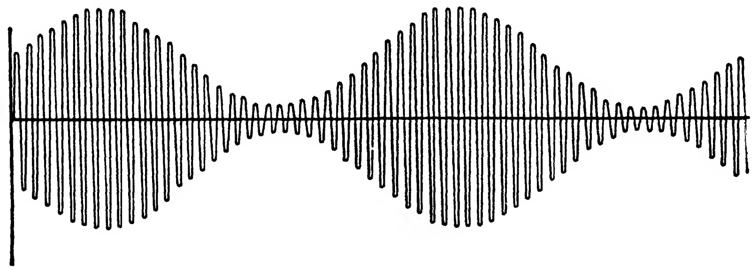


FIG. 469. CARRIER CURRENT MODULATED BY AUDIO-FREQUENCY CURRENT

audio-frequency range as contrasted with the radio-frequency range specified in the preceding paragraph. Figure 468 shows the form of the current produced in a microphone. A sinusoidal wave-form is selected for simplicity in making and interpreting the diagrams. In practice, the waves produced by sound currents are seldom as simple as this, but the descriptions about to be given would apply to any wave-form.

The process of modulation at the transmitter results in the amplitude of the carrier being controlled by the audio-frequency current flowing through the microphone. Thus, if the sinusoidal audio-frequency current of Figure 468 is used to modulate the carrier of Figure 467, there results a modulated carrier the envelope of which is a replica of the modulating current. The form of the resulting current in the antenna circuit of the receiving set is shown in Figure 469.

A very simple type of receiving set is shown in Figure 470. The antenna consists of a wire some distance above the earth and insulated from it. L_1 is a coil of wire wound on an insulating form. C_1 is a condenser the capacitance of which can be varied. These elements constitute the so-called antenna circuit. The alternating current flowing in this circuit because of the action of the carrier wave from the transmitter will be a maximum when the constants of the antenna circuit are adjusted so that its resonance frequency corresponds to that of the carrier wave. This adjustment, called tuning, is made by varying the capacitance of the condenser C_1.

Electromagnetically coupled with the coil L_1 is a second coil L_2. The high frequency current flowing through the coil L_1 induces a

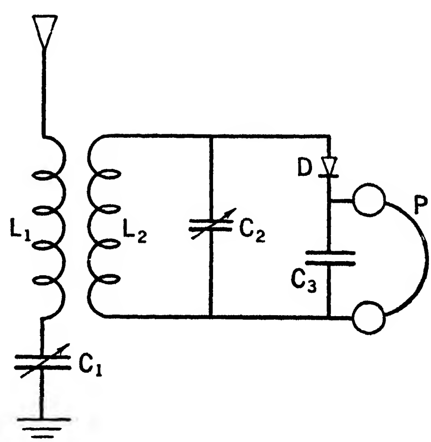


FIG. 470. A SIMPLE BROADCAST RECEIVER WITH A TUNED ANTENNA CIRCUIT COUPLED TO A TUNED CIRCUIT WITH A CRYSTAL DETECTOR

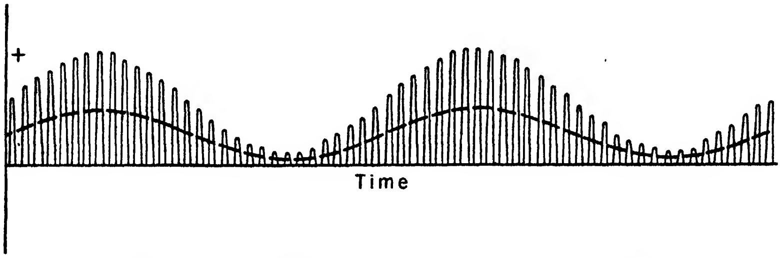


FIG. 471. THE CURRENT FLOWING THROUGH THE DETECTOR
(The current through the phone is the average value of this unidirectional high frequency current and is shown by the dotted line.)

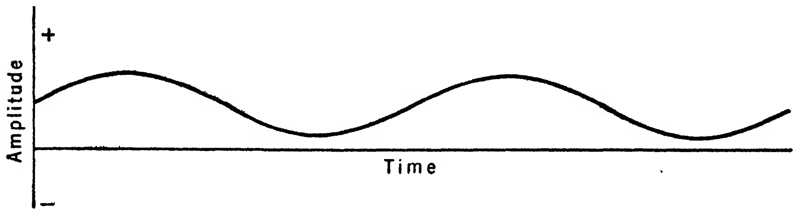


FIG. 472. THE CURRENT FLOWING THROUGH THE PHONES

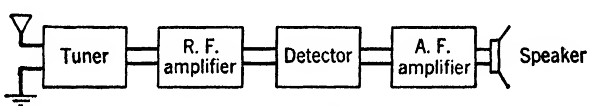


FIG. 473. SCHEMATIC DIAGRAM OF A RADIO RECEIVER

corresponding high frequency voltage in the coil L_2, the two coils together constituting a radio-frequency transformer. The coil L_2 and the variable condenser C_2 constitute an oscillating circuit in which current flows, the current being at maximum when the circuit is tuned to resonance by adjustment of the capacitance of the condenser C_2. The high frequency voltage across the condenser in this circuit is applied to the detector and the head-phones connected in series.

The detector has the property of rectification, that is, it permits current to flow in one direction and not the other. In the case under consideration the current flowing through the detector may be represented as in Figure 471. While this current is unidirectional, reference to Figure 471 shows that it may be considered as a sinusoidally fluctuating direct current of audio-frequency on which is superimposed a radio frequency of varying amplitude. The radio-frequency component passes through the small by-pass condenser of Figure 470, this condenser having a relatively low impedance for such a high frequency. The fluctuating direct current component, which in form and frequency is similar to that in the transmitter microphone, flows through the head-phones and is reproduced as sound.

In this receiver there are embodied the two fundamental operations of tuning and detection. In practice the modulated radio-frequency current is amplified before detection so as to be satisfactorily detected and again after detection so that a loud speaker can be used. As schematically represented in Figure 473, a modern radio receiver contains a tuner, a radio-frequency amplifier, a detector, an audio-frequency amplifier and a speaker.

The Thermionic Vacuum Tube

Without question the invention of the thermionic vacuum tube did more to make modern radio what it is today than any other single accomplishment since the birth of radio. Although volumes have been written on the theory and application of thermionic vacuum tubes, it will be advantageous to consider but a few of the applications, for example, the use of the two-electrode vacuum tube as a detector and of the three-electrode tube as an amplifier and as an oscillator.

In Chapter 46 the process of rectification of alternating currents by means of the diode, or two-electrode vacuum tube, has already been described. A diode could be used as the detector in the receiver of Figure 470. The student will have no difficulty in understanding its action there, especially if he refers to rectification as explained in Chapter 46.

The Triode

Before we take up the action of the triode, or three-electrode vacuum tube, as an amplifier, it will be helpful to examine one of its characteristics. The structure and basic principles of operation of the triode have already been given in the previous chapter. Figure 474 shows how the relation between the plate current and the grid voltage can be determined. The battery E_b serves as the source of potential for the plate circuit. The milliammeter I_p measures the current in the plate circuit. The battery E_c, together with the potential-dividing rheostat P, makes it possible to make the grid *negative* with respect to the cathode by any desired voltage within the range of the battery. The voltmeter E_g indicates the voltage applied to the grid. Moving the slider on the rheostat P changes this voltage and changes the current I_p in the plate circuit. Part (*b*) of Figure 474 shows the dependence of the plate current I_p on the grid voltage E_g. As explained in the former discussion of the three-electrode tube, the grid is nearly always kept negative with respect to the cathode. Use of this characteristic curve is made in the explanation of the triode as an amplifier.

With the triode as an amplifier (Figure 475), the straight-line portion of the characteristic curve is used. The average voltage of the grid is maintained at the midpoint of the straight-line portion of the characteristic curve. The voltage to be amplified is applied to the grid by means of the input transformer T_1, this voltage fluctuating about the average value of the grid voltage maintained by the battery E_c. As shown in Figure 476, this causes a fluctuating direct current to flow through the primary of the output transformer T_2 in the plate circuit of the tube and to induce an alternating voltage in the secondary of this transformer. This voltage is

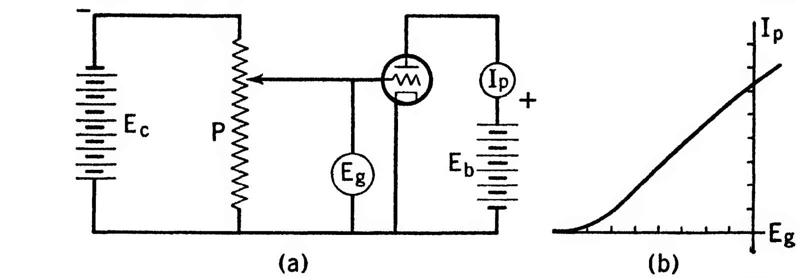


FIG. 474. EXPERIMENTAL DETERMINATION OF THE RELATION BETWEEN THE PLATE CURRENT AND THE GRID POTENTIAL OF A TRIODE

(Part (*b*) shows the characteristic I_p - - E_g curve of a triode.)

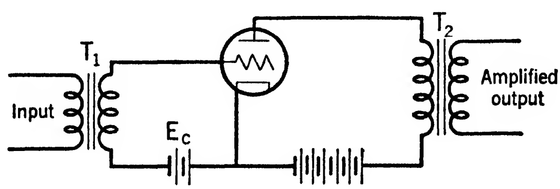


FIG. 475. THE TRIODE AS AN AUDIO-FREQUENCY AMPLIFIER

the same, in frequency and wave form, as that impressed on the primary of the input transformer, but is capable of furnishing more power. The output voltage may be further amplified or may be used in any manner desired, for example, to operate a loud speaker. The amplification of radio frequency voltages is very similar except that transformers are used which do not have iron cores.

The triode may be used as a generator either of audio-frequency or radio-frequency oscillations. In Figure 477 the inductance L_1 and the condenser C_1 form an oscillatory circuit. When oscillations are set up n this circuit they tend to "die down" in the manner shown in Figure 425 (Chapter

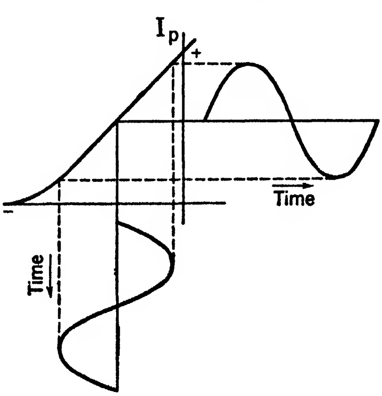


FIG. 476. THE PERFORMANCE OF THE TRIODE AS AN AMPLIFIER

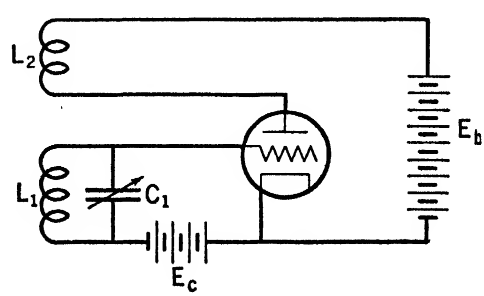


FIG. 477. THE TRIODE AS A RADIO-FREQUENCY OSCILLATOR

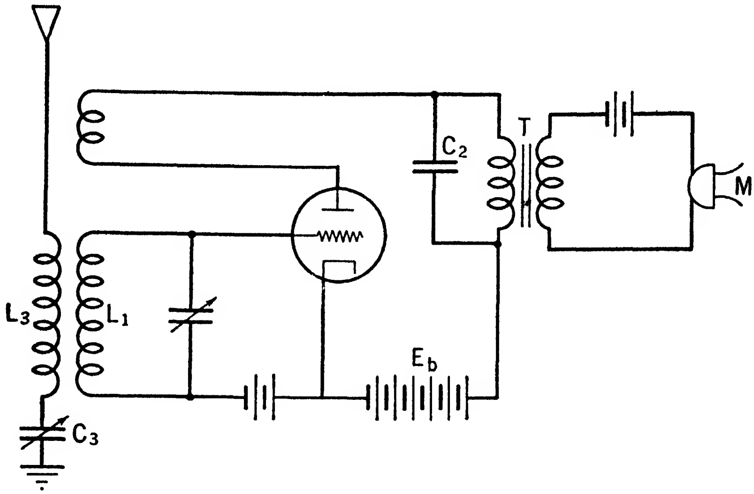


FIG. 478. A SIMPLE RADIO TELEPHONE TRANSMITTER

43) due to the dissipation of energy. But the alternating voltage across the condenser C_1 is impressed on the grid of the tube causing fluctuations of the same frequency in the plate current. The coil L_2 in the plate circuit is inductively coupled with the coil L_1, and the fluctuating current through the coil L_2 induces an alternating voltage in the coil L_1, which is of the same frequency and phase as the oscillations in the L_1C_1 circuit. Sufficient energy is fed into this circuit to overcome the losses, and continuous oscillations are maintained, the frequency of which is determined by the values of the inductance L_1 and the capacitance C_1. By the proper selection of L_1 and C_1, continuous oscillations may be obtained, from a few cycles per second to many billions. In the oscillator shown in the diagram the frequency can be changed by a factor of three or four by adjusting the variable condenser.

A Simple Radio Telephone Transmitter

The simplest type of radio transmitter requires a radio-frequency oscillator to generate the carrier, a means for modulating the carrier with the sound current, and an antenna to radiate the electromagnetic waves from the transmitter. A simple radio telephone transmitter is shown in Figure 478 and consists of the oscillator shown in Figure 477 with two additions, one for modulation, the other for radiation. The pulsating sound current controlled by the microphone *M* flows through the primary of the transformer *T* and induces an alternating voltage in the secondary of the transformer. This voltage is superimposed on that already applied to the tube by the battery E_b, the result being that the oscillations generated by the oscillator are caused to vary in amplitude in accordance with the fluctuations of the sound current. Thus the carrier is modulated. The function of the condenser C_2 is to by-pass the high frequency current around the secondary of the transformer. The antenna circuit consists of the antenna itself, the coil L_3 inductively coupled to the coil L_1, the variable condenser C_3, and a connection to the earth. When the condenser C_3 is adjusted so that the antenna circuit is resonant to the carrier frequency the current is a maximum. There is then radiated from the antenna a modulated electromagnetic carrier wave. It is this wave that is picked up by the antenna of the receiving set.

Continuous Wave Telegraphy

For the transmission of telegraph code signals a transmitter similar to that shown in Figure 478 may be used. However, instead of modulating the carrier wave, arrangements are made for starting and stopping the oscillations by means of a telegraph key. Pulses of radiation of the carrier, of long and short duration corresponding to dashes and dots of ordinary wire telegraphy, are radiated from the antenna of the transmitter. This type of signaling is used because the messages are "readable" over longer distances than the voice of a radio telephone.

The reception of interrupted continuous wave signals requires a slightly different type of receiver than that used for a modulated carrier. The receiver of Figure 479 is identical with that previously described for broadcast reception except that a local oscillator is inductively coupled to it by means of the coil L_3. Therefore there is induced in the coil L_2 of the receiver not only the high frequency current from the transmitting station but also that of the local oscillator which is tuned to a frequency of about 1000 cycles per second different from that of the transmitter. For example, if there were being received interrupted continuous-wave signals of a frequency of 1,700,000 cycles per second, the local oscillator might be tuned to 1,699,000 cycles per second. The effect of these two radio frequencies on the receiver can be explained by means of Figure 480; (*a*) represents the frequency of the local oscillator and (*b*) that of the incoming signal. Since these two frequencies differ by 1000 cycles per second, they will be exactly in phase 1000 times each second and their effects are additive; and 1000 times each second they will be exactly out of phase and their effects subtractive. The form of the resulting current in the L_2C_2 circuit is shown in (*c*) of the figure. This wave form is similar to that of a carrier

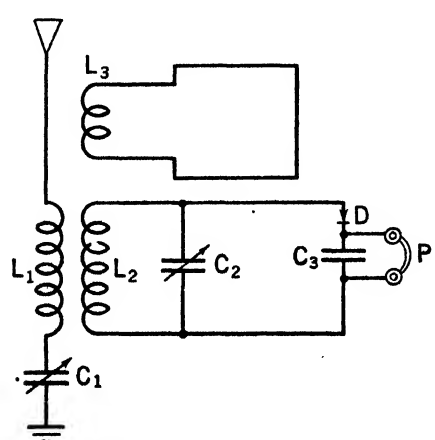


FIG. 479. A RECEIVER FOR INTERRUPTED CONTINUOUS WAVES

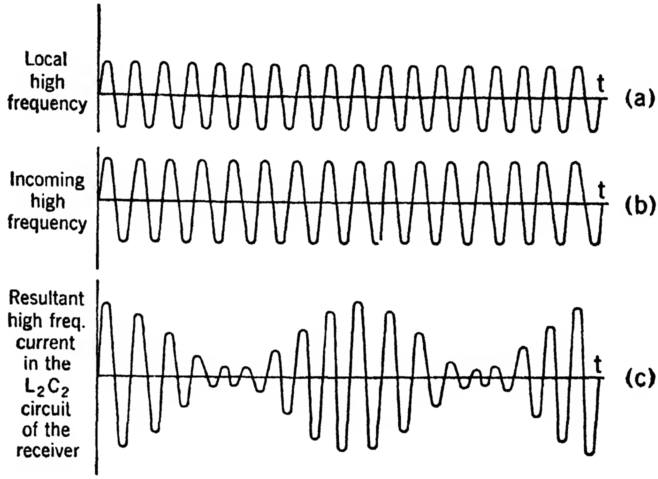


FIG. 480. THE FORMATION OF A BEAT WAVE OF AUDIO FREQUENCY BY INTERFERENCE BETWEEN TWO RADIO-FREQUENCY WAVES

modulated with a 1000 cycle per second frequency. When detected in the manner previously described in connection with the broadcast receiver, a tone of 1000 cycles per second will be heard in the phones. If either of the two high frequencies is lacking, no tone will be heard. Hence, the listener will hear dots and dashes as sent out by the operator of the transmitter.

Interference

A difference frequency similar to that of one thousand cycles described in the last paragraph is called a *beat frequency.* If the beat frequency of two radio carrier waves falling on the antenna of a receiver lies within the range of audibility, *interference*[1] results. The interference manifests itself as a "whistle," the pitch of which is determined by the two radio frequencies causing it. It is to avoid interference of this kind that the Federal Communications Commission has placed the frequency assignments of broadcasting stations at 10 kilocycle intervals.

In order that a radio receiving set may have the proper selectivity, that is, the ability to tune in one station to the exclusion of others, it is necessary to make it relatively insensitive to audio-frequencies above five thousand cycles per second. Since overtones above this frequency contribute greatly to the fidelity of the reproduction of music, present-day radio broadcasting and reception fall far short of perfection. Although broadcasting stations are generally equipped to reproduce programs of higher fidelity, it is impracticable with the present frequency allocations to have

[1] The term *interference* in this section refers to disturbances caused by beat frequencies, static, and so forth. It should not be confused with the term *interference* technically defined elsewhere in the text.

radio receivers sensitive to the higher frequencies because of interference.

Without question the most troublesome type of interference present in radio is that due to static. Electrical discharges of any kind create electromagnetic waves. Whether these originate from the dissipation of charges collected in the atmosphere, especially evident during electrical storms, or from the sparking of electrical devices, such as automobile ignition systems, brushes of electric motors, thermostats of heating appliances, or high tension electric light wires, there result irritating crackling noises in the radio receiver which cannot be eliminated by any sort of tuning. Some of the causes can be removed, but not all of them. Even if we had truly high fidelity reproduction, with our present system of broadcasting it would be marred by the presence of static.

The difficulty is that static interference consists of electromagnetic disturbances spread over a wide range of frequencies. It cannot be tuned out, being present throughout the entire range of frequencies used for radio broadcasting. Any attempt to increase the fidelity of radio reception by widening the frequency response of the receiver would result in an increase of the noise from static.

There is, however, a new system of broadcasting, invented by Major Armstrong and called *frequency modulation*, which does make it possible to minimize if not eliminate static and to have greatly improved reproduction.

The type of modulation described earlier in this chapter and used in present-day broadcasting is called *amplitude modulation*. In this type of modulation the *amplitude* of the carrier is made to fluctuate in accordance with the audio signal being transmitted. The magnitude of the fluctuation depends on the strength or intensity of the audio signal. The frequency of the carrier amplitude fluctuations depends on the frequency of the audio signal.

In frequency modulation the *frequency* of the carrier is made to fluctuate in accordance with the audio signal being transmitted. The magnitude of the carrier frequency fluctuation depends on the strength or intensity of the audio signal and the number of carrier frequency fluctuations per second depends on the frequency of the audio signal. For example, if a sinusoidal audio signal of 1000 cycles per second were used to frequency-modulate a carrier of 44,000 kilocycles, this carrier frequency would be made to fluctuate sinusoidally 1000 times per second between two frequencies, say 43,990 and 44,010 kilocycles per second. If the audio signal of 1000 cycles per second were made twice as great in amplitude the carrier frequency would be made to fluctuate 1000 times per second between 43,980 and 44,020 kilocycles per second. Thus the frequency and amplitude of the fluctuations of the carrier frequency are determined by the frequency and amplitude of the audio signal.

Throughout the entire frequency-modulation system of broadcasting and reception every attempt is made to eliminate amplitude modulation of the carrier. Most types of interference, including static, are similar to amplitude modulated signals. Therefore they are eliminated by a receiving system which is made insensitive to amplitude modulation.

The Radio Spectrum

High fidelity broadcasting requires the reproduction of audio-frequencies up to 15,000 cycles per second. Proper frequency modulation of the carrier under these circumstances requires a band of radio frequencies 200 kilocycles wide. This band is twenty times as wide as that now ordinarily used by an amplitude modulated transmitter. On the other hand, a television transmitter requires a band of radio frequencies from 4000 to 6000 kilocycles wide. The present band of frequencies used by broadcasting stations extends from 550 to 1590 kilocycles per second. Only five frequency modulated transmitters could operate simultaneously in this range, while it would take from four to six times this range to accommodate one television transmitter.

At the present time the assigned radio spectrum extends from 17.6 kilocycles per second to 401,000 kilocycles per second, that is, from a wavelength of about 17 kilometers to 75 centimeters. This spectrum is divided up into over a thousand bands which are allotted to the various services by the Federal Communications Commission. The assignment of these bands of frequencies is made somewhat easier because of the fact that the shorter radio waves are usable for comparatively short distances only.

The Propagation of Radio Waves

The energy radiated from a transmitter may be propagated as a *ground wave* along the surface of the earth or as a *sky wave* through the atmosphere. The ground wave becomes weaker as it travels from the transmitting antenna, the attenuation depending in a complicated manner on the distance, the frequency, and the electrical conductivity of the earth. The sky wave consists of energy radiated in directions other than along the ground. If these waves were to continue onward in their original direction they would of course have their energy lost in space.

There is, however, beginning some one hundred kilometers above the surface of the earth, a region in the atmosphere where the radiation from the sun has produced ionization to a degree sufficient to cause "reflection" of radio waves. This ionized region, or *ionosphere* (also termed the Kennelly-Heaviside layer after the two scientists who independently suggested its existence) does not really reflect the radio waves but rather bends them back to the earth by a continuous refraction process similar to that involved in the formation of mirages. The amount of penetration of the radio wave into the ionosphere before being bent back to the earth depends in a complicated manner on the direction and frequency of the wave and on the density of the free electrons in the ionosphere. That there are several effective "reflecting" layers in the ionosphere and that some of the layers undergo diurnal and seasonal variations as well as variations correlated with sun-spot cycles add to the complexity of the situation.

During the daytime, signals received on frequencies within the broadcast band are transmitted by the ground wave, the sky wave being almost completely attenuated. At night, however, the sky wave plays an important part and accounts for the reception of radio signals at distant points.

In the so-called short-wave region (1600 to 30,000 kilocycles) the ground wave is attenuated so rapidly as to be of no importance except for very short distances. Short-wave communication depends on the sky wave which may be "reflected" back and forth several times between the ionosphere and the earth before it is received at a far distant point. In the ultra-high frequency range (30,000 kilocycles and upward) the ground wave is very quickly absorbed and there is no "reflected" sky wave. Satisfactory communication in this region can be obtained only by using waves which pass from the transmitter to the receiver in a straight line. Due to the curvature of the earth this limits the range to a few miles. For this reason the reception of television and frequency modulated signals, which are allocated to the ultra-high frequency region, is limited to a distance of approximately thirty-five miles from the transmitting antenna. This means that for the near future, at least, these services will be limited to metropolitan areas.

Questions for Self-Examination

1. Outline the stages involved in transmitting sound over a beam of light and compare them with radio transmission.
2. Compare the contributions of Maxwell, Hertz and Marconi to the birth of radio communication.
3. Distinguish between a transmitted radio signal and the accompanying carrier wave.
4. Draw the circuit of a simple broadcast receiver, without amplification, and tell the function of each part.
5. Describe the "triode" and tell how it functions (*a*) as an amplifier, (*b*) as an electric oscillator.
6. Why can a continuous (unmodulated) radio wave not be heard? How can it be made audible for purposes of radio telegraphy?
7. Distinguish between amplitude modulation (the older method) and frequency modulation. What is the advantage of the latter?
8. What causes radio waves to take the curved path necessary to travel around the earth?

CHAPTER 48

Cathode Rays and the Electron

Instrumentation as an Element in Scientific Discovery

The history of scientific discovery is to a great extent the history of scientific instruments and improvements in experimental technique. This is especially true of those remarkable discoveries of the twentieth century that are usually classified under the general heading of "Modern Physics." Although the year 1895 is generally considered to be the dividing line between the new physics and the old, actually the transition was a gradual process extending over a period of forty years. The immediate cause of the rapid transition which occurred in 1895–1900 was the discovery of X-rays and electrons. It must be kept in mind, however, that the investigations which led to these discoveries originated in the study of electrical discharges through rarefied gases. These experiments, in turn, were made possible by improvements in the method of producing high vacuums.

The Evolution of Vacuum Pumps

The first mechanical air pump, which was invented by von Guericke in 1650, has been mentioned on page 85. Except for improvements in mechanical details no advance in vacuum technique was made in the next two hundred years. It was in 1855 that Heinrich Geissler, a skillful glass blower in Bonn, Germany, and the originator of the Geissler discharge tube, devised a new type of vacuum pump.

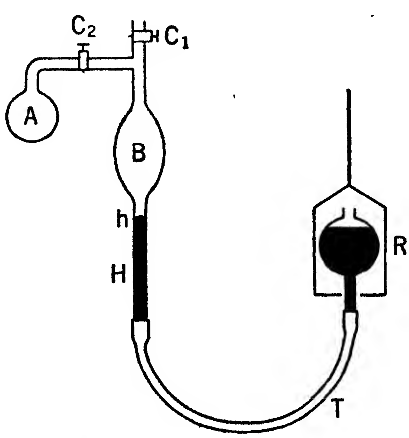


FIG. 481. GEISSLER'S VACUUM PUMP

In Geissler's apparatus all mechanical plungers and leather valves were eliminated, and the only moving part was a column of mercury. The action of the pump depends upon the existence of the Torricellian vacuum in the top of a barometer tube (page 85). The operation of Geissler's apparatus is easily understood with the aid of Figure 481. *A* is the glass vessel which is to be evacuated. B is a bulb blown in the top of a barometer tube, which is connected to a reservoir of mercury R by means of the rubber tube T. This reservoir can be raised or lowered by means of a cord and windlass.

With the stop-cock C_2 closed and C_1 open, R is raised until the level of mercury reaches C_1. C_1 is then closed and R lowered until the mercury level is at h and B is evacuated. C_2 is now opened, and the air originally contained in A is distributed between A and B. By repeating the above process, the air in B is forced out through C_1. In each operation a certain fractional part of the remaining air in A is removed. After several hours of slow and monotonous work, a high degree of vacuum can be attained — the final pressure being determined by the tightness of fit of the stop-cocks and the vapor pressure of mercury. All the early discharge tubes and X-ray tubes were evacuated by this slow, laborious method.

The next important advance was the invention of the diffusion pump by Gaede and Langmuir in 1913. This pump has no moving parts and is very fast. In a few minutes it can produce a higher vacuum than it is possible to obtain in several hours with the older type of apparatus. The essential part of the Langmuir pump is a jet of mercury or oil molecules which are moving with a high speed. This molecular stream tends to drag along adjacent layers of air so that a pumping action results.

With modern technique, pressures as low as 10^{-8} millimeters of mercury can be maintained in large vessels. It must be remembered, however, that current expressions such as "vacuum tube" and "high vacuums" must be interpreted in a relative sense. Even at the very lowest pressures now obtainable there are still approximately one hundred million molecules of air in each cubic centimeter of the "vacuum." Although this number sounds very large, the molecules are so small that at this pressure an atom of electricity can travel several feet without colliding with a single molecule.

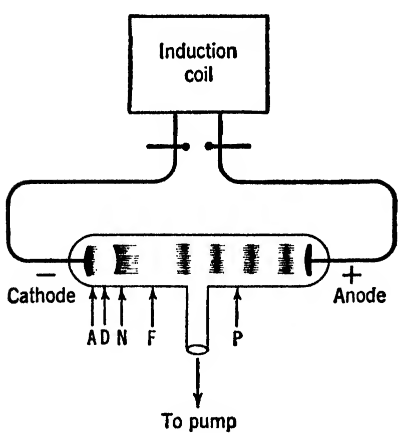


FIG. 482. APPEARANCE OF ELECTRIC DISCHARGE THROUGH A PARTIAL VACUUM

The Electric Discharge in Rarefied Gases

When an electric discharge passes through a gas at low pressure, several beautiful and interesting effects can be observed. A convenient arrangement for studying these effects is shown in Figure 482. The induction coil supplies the high potential required to maintain the discharge. If the glass tube is relatively long — four or five feet — the display is more spectacular.

At atmospheric pressure, there is no discharge. But when the pressure is reduced to 7 or 8 centimeters of mercury, a violet-colored thread of light extends from one electrode to the other. As the exhaustion proceeds the tube becomes filled with a soft glow, and at pressures of from 1 to 2 millimeters of mercury the discharge has the characteristics shown in Figure 482. The cathode appears to be covered with a velvety, luminous coating A called the cathode layer. Beyond this is a dark region D known as the Crookes' dark space. Next is a luminous patch N, called the negative glow, which is followed by the Faraday dark space F. The remainder of the tube is filled with a luminous band P, known as the positive column, which is divided into sections or striae of unequal intensity.

The length of the Crookes' dark space is independent of the distance from cathode to anode but is a function of the pressure. As the exhaustion of the tube proceeds, the Crookes' dark space becomes longer, while the positive column becomes shorter and less luminous. At a pressure of approximately .001 millimeter of mercury the positive column disappears and the Crookes' dark space seems to fill the tube. At the same time the glass walls of the tube in the region surrounding the anode glow with a bright green fluorescent light.

It is natural that the first detailed study of discharge phenomena should have been made in Bonn, the home of Geissler. Three years after Geissler's invention of his vacuum pump, Pluecker, professor of Physics at the University of Bonn, published a paper entitled *The Influence of a Magnet upon the Electric Discharge in Rarefied Gas*. Using a highly evacuated tube he observed the green fluorescent glow described above. But his most important observation was that when the tube was brought near an electromagnet the position of the glow shifted, changing from one side to the other when the polarity of the magnet was reversed. He also observed that the effect of the magnetic field upon the luminosity was independent of the nature of the gas in the tube and the material of the cathode. Gold, silver, and copper cathodes gave identical effects.

In 1869, Pluecker's student and colleague at Bonn, W. Hittorf, began a new series of investigations on discharge phenomena. He made the interesting discovery that all objects, solid or liquid, conductor or insulator, cast a well-defined shadow when placed before the cathode of a discharge tube. This shadow of an opaque object was later demonstrated in a spectacular manner by the English chemist Sir William Crookes. The type of tube used by Crookes, which is still in common use for demonstration purposes, is shown in Figure 483. When the electrode C is used as the cathode or negative terminal for the discharge a shadow of the cross is visible as shown in the figure.

The existence of these sharp shadows suggests that something is coming from the cathode which travels along straight lines. In order to verify this hypothesis Pluecker devised the tube shown in Figure 484. He observed that when the electrode b was used as the cathode, the negative glow filled

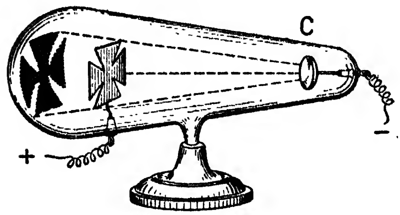


FIG. 483. TYPE OF DISCHARGE TUBE USED BY CROOKES

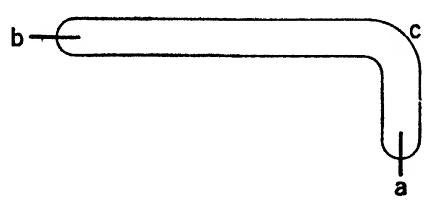


FIG. 484. TYPE OF DISCHARGE TUBE USED BY PLUECKER

the long arm of the tube. On the other hand, when *a* was used as the cathode the glow was confined to the short arm; the luminosity did not bend around into the arm *cb*. Hittorf therefore concluded that the glow was propagated along straight lines or *rays* and that each point of the cathode acted as the source of a cone of rays.

He observed also that if the cathode was a plane disk, a magnetic field at right angles to the line of discharge caused the negative glow to shift to the edge of the disk. He interpreted this shift correctly as an example of the motor principle.

Rival Theories Concerning Cathode Rays

During the next decade the principal studies of discharge phenomena were made by Goldstein in Berlin and Crookes in England. In addition to verifying Hittorf's work, Goldstein showed that the luminous "rays" given off by the cathode were perpendicular to the emitting surface. This is quite different from the emission of ordinary light, where the radiation is emitted in all directions. The study of this effect led Goldstein to waste a great deal of time and energy in making several hundred cathodes of various geometrical forms. The ultimate scientific value of all this work was negligible. He did, however, make two contributions which influenced later developments. One of these, the discovery of positive rays, will be discussed in Chapter 51. The other was the proposal and defense of the hypothesis that cathode rays were a type of electromagnetic radiation similar to light. The term "cathode rays" — *kathodenstrahlen* in German — was introduced by Goldstein although it seems to have been suggested orginally by Wiedemann.

Meanwhile Sir William Crookes, who had acquired experience in high vacuum technique in connection with his radiometer experiments, had begun to work on cathode rays. Crookes was largely self-taught and, like Faraday, had had no training in mathematics. But these deficiencies were outweighed to a certain degree by his experimental ingenuity and his remarkable imagination. Another important factor which contributed to his productivity and success was the great technical skill of his assistant and glassblower, G. C. Gimingham. Some of the tubes that he built have never been improved upon.

Crookes adopted the hypothesis — first proposed by Varley in 1871 — that the cathode rays were tiny corpuscles shot off in straight lines from the cathode. Following a suggestion that Faraday had made in 1816, he called these corpuscles "radiant matter" or

> matter in a fourth state or condition which is as far removed from the state of a gas as a gas is from a liquid.

Consequently all his experiments were designed for the purpose of proving the corpuscular nature of the rays. First, he verified and extended earlier observations on shadows and magnetic deflection. Then he attempted to prove that the cathode rays have momentum.

The Momentum and Kinetic Energy of Cathode Rays

According to Newtonian mechanics, a stream of material particles moving with high speed must have momentum and kinetic energy. In order to demonstrate the momentum of the rays Gimingham constructed the tube shown in Figure 485. The paddle wheel *W* has light mica vanes and rolls freely along the horizontal glass rails *g*. When electrode *a* is the cathode the wheel rotates to the right, and when *b* is the cathode it rotates in the opposite direction. At the time, this experiment was considered by many to be conclusive evidence in favor of the corpuscular theory. Several years later, however, J. J. Thomson showed that the observed rotation is produced by the heating of the vanes and not by the transfer of momentum through direct impact. Since ether waves also produce a heating effect the experiment is not decisive.

The transmission of energy by the cathode ray beam was demonstrated by the tube shown in Figure 486. The cathode *c* is a spherical cap which focuses the rays at the point *b*. When a discharge is sent through the tube the platinum disk at *b* becomes white hot. When the cathode rays are deflected by a magnet the heating effect disappears.

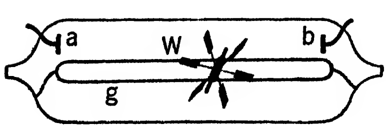


FIG. 485. TUBE DESIGNED TO SHOW THAT CATHODE RAYS POSSESS MOMENTUM

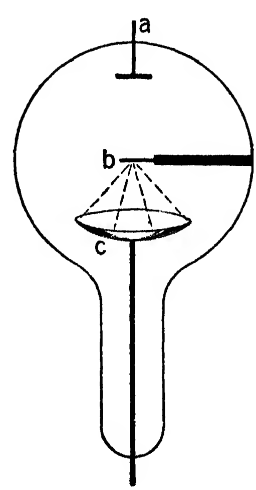


FIG. 486. TUBE DESIGNED TO SHOW THAT CATHODE RAYS POSSESS KINETIC ENERGY

Early Reactions to the Study of Cathode Rays

Since Crookes was an able and vigorous expositor of scientific subjects, his work soon attracted attention. In 1879 he was invited to give a general lecture on cathode rays — or radiant matter as he called them — before the British Association. In this address, he gave spectacular demonstration of all the known properties of radiant matter. The list of his demonstrations is instructive and worth repeating. He showed that cathode rays produced fluorescence, traveled in straight lines, were deflected by a magnetic field, carried momentum and transmitted energy. Furthermore, he emphasized the fact that all these effects were entirely independent of the chemical nature of the gas in the tube and the material of the cathode. Crookes concluded his address with the following enthusiastic and prophetic statement:[1]

> In studying this fourth state of matter we seem at length to have within our grasp and obedient to our control the little indivisible particles which, with good warrant, are supposed to constitute the physical basis of the universe. We have seen that in some of its properties radiant matter is as material as this table, whilst in other properties it almost assumes the character of radiant energy. We have actually touched the borderland where matter and force seem to merge into one another.
>
> I venture to think that the greatest scientific problems of the future will find their solution in this border land, and even beyond; here, it seems to me, lie Ultimate Realities, subtle, far-reaching, wonderful.

Crookes' demonstrations were generally taken to substantiate his verdict that cathode rays were "matter," albeit in a "fourth state." But according to Maxwell's theory, electromagnetic waves possessed four of the five properties demonstrated. The one characteristic of the rays that was not shown by electromagnetic waves was deviability in a magnetic field.

Crookes' enthusiasm was not shared by the majority of English physicists. Many were apathetic and some actually hostile. The latter went so far as to advise young men not to enter this field of research. They maintained that the study of cathode rays had no future. Their attitude was well expressed by a wealthy manufacturer who had endowed a physics laboratory. When he was shown a long discharge tube in operation he remarked, "How beautiful and how useless."

In Germany the corpuscular idea encountered strong opposition because it was in direct conflict with Goldstein's ether theory. Helmholtz, however, who was the leading German theorist at that time, favored the corpuscular view. He made several unsuccessful attempts to convert Goldstein.

The strongest support for the ether theory came from the work of Hertz and Lenard during the period 1880–94. In order to appreciate Hertz's keen critical insight we must consider first an experiment of fundamental importance made by Rowland in 1876.

[1] *Report of the British Association for the Advancement of Science* (1879).

Rowland's Experiment

When Maxwell wrote his *Treatise on Electricity and Magnetism* in 1873, he was in doubt concerning the following question: Is a piece of matter that is charged electrostatically and moving with high speed equivalent to an electric current in a wire? If so, a magnetic effect similar to that existing around the wire should be produced. Maxwell calculated the anticipated magnetic effect as being within the range of experimental possibilities.

At the suggestion of Helmholtz, Rowland attempted to obtain an answer to Maxwell's question. He attached strips of tin foil to the outer part of a glass disk, charged the strips electrostatically and set the disk into rotation. When the speed of rotation was sufficiently high a magnet near the disk was deflected; reversing the direction of rotation produced a deflection in the opposite direction. Owing to this experiment and to additional support based on theoretical considerations all physicists at the time accepted the principle that a moving charged body is equivalent to an electric current.

The Experiments of Hertz and Lenard

Hertz saw that this principle provided the means of making a crucial test of the corpuscular theory. The presence of a magnetic field would be convincing evidence for moving corpuscles. But he was never able to detect such a field near the cathode ray beam. A few years later he devised another crucial experiment. The cathode rays were shot between two parallel plates, one of which carried a positive charge and the other a negative. The principles of electricity required that a charged particle must be deviated in passing between the plates. Hertz was unable to detect even the slightest deviation. These two experiments convinced him that the corpuscular theory was untenable. His conclusion was logically sound but unfortunately his experimental results were wrong. Later work with improved technique showed that both effects actually exist.

In 1892 the ether theory received additional support when Hertz showed that cathode rays can penetrate thin sheets of metal. This interesting effect was thoroughly examined by Lenard who used a tube similar to that shown in Figure 487. One end of the glass tube is closed with a metal plate *P*. A small hole in the plate is covered with aluminum foil *F*. A sheet of foil .003 mm. thick is transparent to cathode rays but impervious to air.

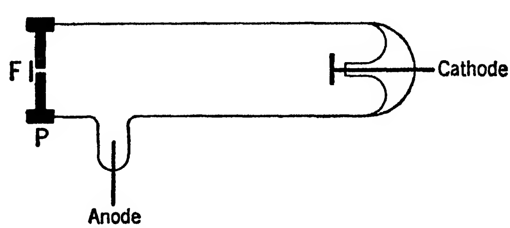


FIG. 487. LENARD TUBE

Lenard found that thin sheets of many substances, for example glass, mica, paper, gold, and copper were partially transparent to the rays. This showed that Crookes' corpuscles could not be charged atoms or molecules as had been assumed.

The Decisive Experiments of J. J. Thomson

Using improved technique, Thomson proved the existence of the electrostatic deflection that Hertz had looked for but had failed to find. His most important results, however, were the first reliable determinations of the speed of the rays and the ratio of their charge to their mass.

Concerning the cause of Hertz's failure, Thomson says (77:583):

> On repeating this (Hertz's) experiment I at first got the same result, but subsequent experiments showed that the absence of deflection is due to the conductivity conferred on the rarefied gas by the cathode rays. On measuring this conductivity I found that it diminished very rapidly as the exhaustion increased; it seemed then that on trying Hertz's experiment at very high exhaustions there might be a chance of detecting the deflection of the cathode rays by an electrostatic field.
>
> [The apparatus used is represented in Figure 488.]
>
> The rays from the cathode C pass through the slit in the anode A.... After passing through another slit they travel between two parallel aluminum plates about 5 centimeters long and 2 broad and at a distance of 1.5 centimeters apart; then they fall on the end of the tube and produce a narrow well defined phosphorescent patch.... At high exhaustions the rays were deflected when the aluminum plates were connected to a battery of small storage cells; the rays were depressed when the upper plate was connected with the negative pole of the battery, the lower with the positive and raised when the upper plate was connected with the positive, the lower with the negative pole. The deflection was proportional to the difference of potential between the plates, and I could detect the deflection when the potential difference was as small as two volts.

In discussing these results, Thomson says:

> As the cathode rays carry a charge of negative electricity, are deflected by an electrostatic force as if they were negatively electrified, and are acted on by a magnetic force in just the way in which this force would act on a negatively electrified body moving along the path of these rays, I can see no escape from the conclusion that they are charges of negative electricity carried by particles of matter. The question next arises, What are these particles? Are they atoms, or molecules, or matter in a still finer state of subdivision? To throw some light on this point, I have made a series of measurements on the ratio of the mass of these particles to the charge carried by it.

The same tube (Fig. 488) was used for the determination of the ratio mass/charge. When the plates D and E are at the same potential the rays strike the fluorescent screen at a. When a potential difference exists, and

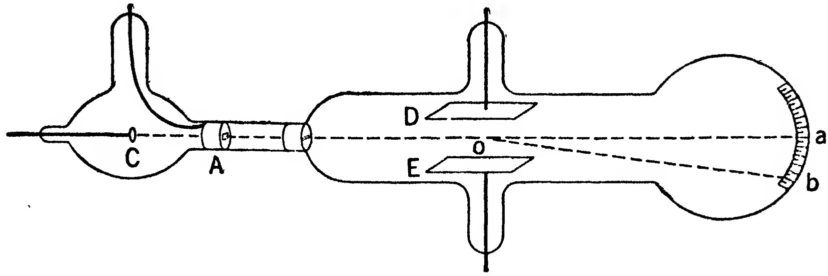


FIG. 488. INTERIOR OF THOMSON'S TUBE TO SHOW ELECTROSTATIC DEFLECTION OF CATHODE RAY STREAM

D is negative, the fluorescent spot is deflected to the position *b*. The amount of this deflection *ab* can be calculated from the principles of mechanics and electricity.

Assume a stream of identical charged particles moving to the right between the plates with a constant speed v. Let the charge of each particle be e and its mass m. If F is the intensity of the electric field between the plates the force on each particle is Fe. By Newtonian mechanics each particle will then have an acceleration a given by

$$a = \frac{Fe}{m}. \tag{1}$$

After an interval of time t, its velocity downward V_y will be

$$V_y = at = \frac{Fet}{m}. \tag{2}$$

Now let $t = t_1$ where t_1 is the time required for the particle to travel the distance l between the plates.

$$t_1 = \frac{l}{v} \tag{3}$$

and equation (2) becomes

$$V_{y1} = \frac{Fe}{m} t_1 = \frac{Fel}{mv} \tag{4}$$

where V_{y1} is the downward velocity of the particle as it leaves the region between the plates. The actual velocity of the particle $\vec{V}_r$ is the resultant of the horizontal velocity $\vec{v}$ and $\vec{V}_{y1}$ as shown in Figure 489. From the figure we see that the angle θ is determined by

$$\tan \theta = \frac{V_{y1}}{v} = \frac{Fel}{mv^2}. \tag{5}$$

From Figure 488 we find that the experimental value of $\tan \theta$ is ab/a_0. Since F and l can be measured, equation (5) gives us a numerical value for

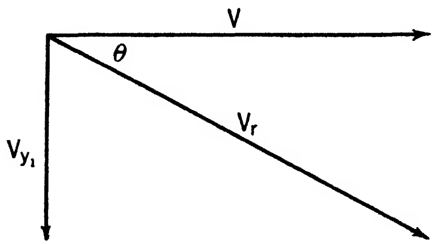


FIG. 489. COMPOSITION OF ELECTRON VELOCITIES

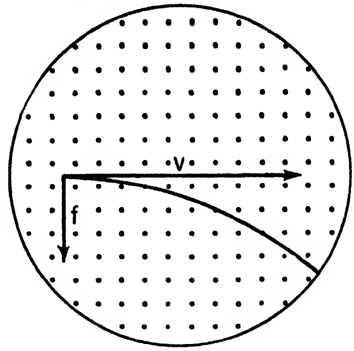


FIG. 490. ESTABLISHING ELECTRONIC CHARGES AS NEGATIVE

e/mv^2. To obtain the separate values of e/m and v it is necessary to have another independent relation between them.

Another Relation Yielding Information on Speed of Electrons

This desired relation can be obtained from another experiment in which the rays are deflected by a magnetic field. It will be recalled that both Rowland's experiment and the electromagnetic theory show that a stream of negatively charged particles moving to the right is equivalent to an electric current in the opposite direction. If, for example, the magnetic field in Figure 490 is perpendicular to the plane of the paper and downward, and if its intensity is H each particle will experience a force

$$f = \mu_0 Hev. \tag{6}$$

Since $\mu_0 H$ is equal to the magnetic induction B, equation (6) can be written,

$$f = Bev. \tag{7}$$

When B is expressed in webers/sq. meter, e in coulombs, and v in meters/sec, f is in newtons. The force deflecting a stream of electrons is identical with the force on an equivalent current. As Figure 490 indicates, the force f has the peculiar property of being perpendicular to both H and v.

This necessitates a continuous change in the direction of f, since the velocity $\vec{v}$ is always changing direction. The magnitude of f, however, does not change.

On page 110 it was shown that a particle moving in a circular path of radius R with a constant speed v has an acceleration of v^2/R, which is directed toward the center of the circle. This is equivalent to the statement that in uniform circular motion the acceleration of a particle is always at right angles to its direction of motion. Since acceleration is always proportional to force, the acceleration of the particle in Figure 490 will be perpendicular to the path. In order to comply with this condition and at the same time keep the magnitude of f constant, the path must be the arc of a circle. The radius R of this circle is given by the equation for centripetal force (page 110).

$$\text{Centripetal force} = \frac{mv^2}{R}. \tag{8}$$

Substituting the value of the force from equation (7), (8) becomes

$$Bev = \frac{mv^2}{R} \tag{9}$$

or

$$BR = \left(\frac{m}{e}\right) v. \tag{10}$$

(B in webers/sq. meter; R in meters; m in kgms; e in coulombs; v in meters/sec.)

Equation (10) has proved to be one of the most useful relations in modern physics. Since B and R can be measured readily it gives a convenient method of finding v if m/e is known or of comparing values of m/e for particles having the same velocity. We shall come across many applications of (10) in the chapters which follow. As a matter of historical accuracy it should be pointed out that J. J. Thomson was not the first to derive and use this equation. It was used in theoretical work by Riecke in 1881, and three years later Schuster suggested its application to the problem of finding m/e.

Thomson found the velocity of the cathode rays by adjusting the magnetic field B to such a value that the magnetic force Bev was equal and opposite to the electric force Fe. Under these conditions

$$Bev = Fe \quad \text{or} \quad v = \frac{F}{B} \tag{11}$$

(v in meters/sec; F in newtons/coulomb; B in webers/sq. meter). His original observations gave values of v between $2.2 \times .07$ and 3.6×10^7 meters/sec. Substituting the value of v in equation (5) he found m/e to be approximately 10^{-11} kgms/coulomb.

Surprising Conclusions from Thomson's Experiments

Thomson's value for m/e was surprising; for it was only $\frac{1}{1000}$ of the corresponding ratio for the hydrogen ion in electrolysis. At this time the hydrogen atom was believed to be the lightest particle of matter that could exist. In discussing his unexpected results Thomson said:

> The smallness of m/e may be due to the smallness of m or the largeness of e or to a combination of the two. That the carriers of the charges in the cathode rays are small compared with ordinary molecules is shown, I think, by Lenard's results.

In other words, the results of Thomson and Lenard can be reconciled and explained in a satisfactory manner by making the radical assumption that atoms can be subdivided into smaller particles. The year 1897, therefore, may be considered as marking the end of the age-old concept of the atom as an indivisible unbreakable entity. Since that time, the elastic billiard ball atom of kinetic theory has been replaced by a complex structure of smaller components. The significance of the new point of view was summarized by Thomson as follows:

> Thus on this view we have in the cathode rays matter in a new state, a state in which the subdivision of matter is carried very much further than in the ordinary gaseous state; a state in which all matter — that is matter derived from different sources such as hydrogen, oxygen, etc. — is of one and the same kind; this matter being the substance from which all chemical elements are built up.

The Irish theorist G. F. Fitzgerald was the first to appreciate fully the fundamental significance of Thomson's work. In 1897 [1] he said:

> As regards the calculation of the ratio of the numerical measure of the mass of the corpuscle to the electric charge it carries, there are two suggestions that can be made in respect to it. The first is that we are dealing with free electrons in these cathode rays.

After pointing out that Thomson's results do not actually prove that atoms have been dissociated, he makes the following prophecy:

> In conclusion, I may express the hope that J. J. Thomson is quite right in his by no means impossible hypothesis. It would be the beginning of great advances in science, and the results it would be likely to lead to in the near future might easily eclipse most of the other great discoveries of the nineteenth century.

The profound and extensive alterations in the structure of physics which resulted from the discovery of the electron, X-rays, and radioactivity will be treated in the following chapters.

The Zeeman Effect

A few months before Thomson solved the problem of cathode rays a remarkable discovery was made by P. Zeeman in Leyden, Holland. Placing a sodium flame between the poles of a magnet, he observed that the spectral lines which under normal conditions are sharp and narrow, were broken into two or three lines very close together. An explanation of the effect was given immediately by the Dutch theorist Lorentz in terms of his electron theory of matter. Lorentz assumed that just as the electrical oscillations in an antenna give rise to radio waves so the vibrations of ions (or electrons) in an atom produce light. When the atom is in a magnetic field the vibrating electron is subject to an electromagnetic force given by

[1] *39*, 103 (1897).

equation (7). This additional force produces a change in the frequency of vibration of the electron which is manifested by a change in the frequency and wave-length of the emitted light. By observing the change in wave-length and knowing the magnetic field intensity, Lorentz was able to calculate the value of m/e for the vibrating charge in a sodium atom. His result for this ratio was approximately 10^{-11} kgms/coulomb — which, within limits of experimental error, is precisely the same as the value obtained a few months later for cathode rays. Furthermore, Lorentz showed that the vibrating ion that was responsible for the emission of light carried a *negative* charge. When we recall that all cathode rays carry negative charges and are extracted from matter, it becomes evident that this tiny negatively charged corpuscle plays an important rôle in physical processes. To appreciate the excitement and enthusiasm which prevailed in scientific circles at this time it should be recalled that the discoveries of X-rays and radioactivity had been announced just the year before, 1896.

The Measurement of the Electronic Charge

Having made the assumption that the corpuscular mass m instead of the charge e was responsible for the small value of m/e, Thomson next undertook to prove it. Since a direct determination of the mass of a single atom is impossible, he attempted to measure the charge e. In this case also, it seemed impossible to proceed directly by observing the charge of a single ion. He therefore adopted the alternative of measuring the total charge carried by a known large number of ions. At first sight this seems to involve the impossible task of counting ions which are invisible. But this aspect of the difficulty was overcome in a very ingenious manner.

At that time a Scotchman in Thomson's laboratory, C. T. R. Wilson, was studying the formation of fogs. He found that if ions, that is, atoms or molecules carrying electrical charges, are present in saturated water vapor, they act as nuclei of condensation when the vapor is cooled. Around each ion a tiny water droplet is formed. With proper illumination the fog is clearly visible and its movement can be observed. Thomson found the total mass of a given fog by precipitating it and weighing the water. This weight divided by the mass of a single droplet gives the number of droplets and presumably the number of ions. But the weight of a single fog droplet is so small that it must be found by indirect methods.

Here again, Thomson's ingenuity and wide theoretical knowledge supplied the solution. Early in the nineteenth century, Sir George Stokes had shown that a small sphere of radius a moving through the air with a speed v experiences a resistance R given by

$$R = 6\,\pi a\eta v \qquad (12)$$

where η is the coefficient of viscosity of the air. It is owing to this resistance that rain drops fall with a constant speed. As a falling drop gathers speed, the resistance of the air increases until it becomes equal to the downward force of gravity on the drop. The resultant force now being zero, the velocity of the drop remains constant. Neglecting the slight correction arising from the buoyancy of the air, the downward force on a droplet of mass m is equal to mg. When the speed of the droplet has reached its constant value V_c, equation (12) becomes

$$6\pi a\eta V_c = R = mg. \tag{13}$$

But we know that

$$m = \tfrac{4}{3}\pi a^3 d \tag{14}$$

where d is the density of water. From (13) and (14) we find

$$a^2 = \frac{9}{2}\frac{\eta V_c}{dg}. \tag{15}$$

Having found a from (15) we can calculate the mass of a single droplet. This, in turn, enables us to find the total number of droplets in the cloud. The total charge of all the ions in the cloud can be measured with a sensitive electrometer. Thus all the necessary data for determining the charge of a single ion is now available.

Using this method and assuming that each droplet contained one ion, Townsend, in 1897, obtained a value of the ionic charge of approximately 1×10^{-19} coulombs. After making certain changes in technique, however, he repeated the experiment and gave his final result as 1.1×10^{-19} coulombs.

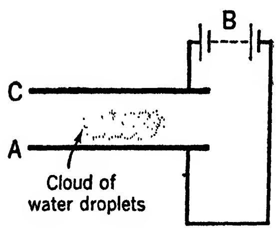


FIG. 491. WILSON'S CLOUD METHOD OF MEASURING ELECTRONIC CHARGE

The next important improvement in experimental procedure was made by H. A. Wilson in 1903. In his method a cloud of droplets is formed between horizontal parallel plates as shown in Figure 491. When the plates C and A are connected to a battery, an electrostatic field is established in the region between them. If the field strength is E, each ion with a charge e_1 will be subject to a force Ee_1. By selecting proper values of E it is possible to hold the droplets stationary or to cause them to fall or rise. In the last case the electrical force Ee_1 must be greater than mg. Assuming that this condition holds, equation (13) gives for the velocity of rise V_E

$$V_E = \frac{Ee_1 - mg}{6\pi a\eta}. \tag{16}$$

Similarly the velocity of fall without the electric field is

$$V_g = \frac{mg}{6\pi a\eta}. \tag{17}$$

Dividing (16) by (17) gives

$$\frac{V_E}{V_g} = \frac{Ee_1 - mg}{mg} \tag{18}$$

or

$$e_1 = \frac{mg}{E} \frac{(V_E + V_g)}{V_g}. \tag{19}$$

By measuring all the quantities on the right side of (19), Wilson obtained values of the average charge of a droplet, which varied from 0.7×10^{-19} to 1.4×10^{-19} coulombs. Although the results were significant they could scarcely be considered satisfactory.

The important improvements in Wilson's method which finally led to precise results were introduced by R. A. Millikan at the University of Chicago. Millikan saw that Wilson's difficulties arose from two sources. One was the impossibility of making accurate observations on the velocity of any body as extended and amorphous as a cloud; the other was the evaporation of the droplets. He removed the first by observing a single droplet and eliminated the second by using oil instead of water. The essential features of the final arrangement are shown in Figure 493. Oil is sprayed from an atomizer and a few droplets are allowed to fall through a small hole in the upper plate *C*. Under the intense illumination from the source of light *S*, the droplets are visible in the telescope *T*. Most of the droplets become charged when they are sprayed from the atomizer. When the

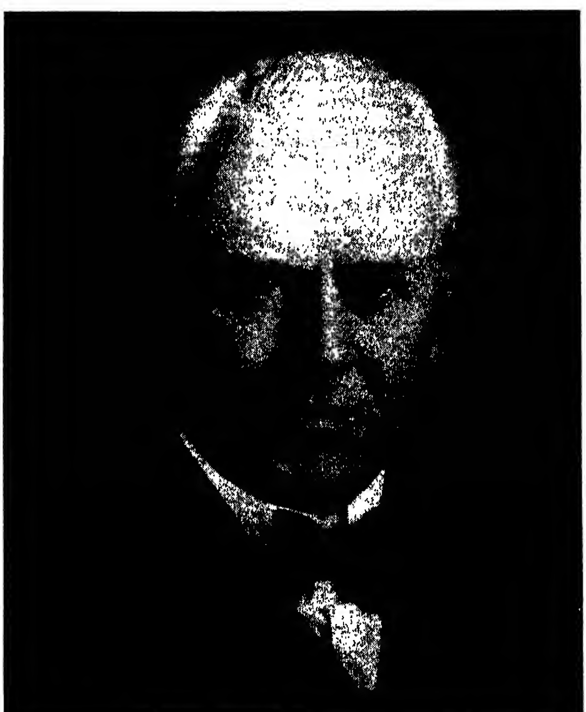


FIG. 492. R. A. MILLIKAN (1868–)

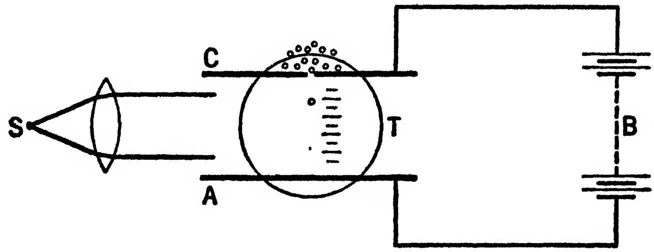


FIG. 493. MILLIKAN'S OIL-DROP EXPERIMENT

electric field is applied, some of those with a negative charge will move upward. Selecting one of these, the observer can measure its speed V_E with the electric field and its speed V_0 without the field. Equation (19) then gives e_1 which is the charge carried by this particular droplet and not an average value as was the case in Wilson's experiment. Millikan and his students made thousands of observations on individual droplets. In every case they found that the charge on the droplet e_1 was a certain value e or $2e$ or $3e$, etc., that is, e multiplied by an integer. The value of this smallest charge e was 1.60×10^{-19} coulombs.

The significance of this experiment was stated by Millikan (85b:77):

> [These results] place beyond all question the view that an electrical charge wherever it is found, whether on an insulator or a conductor, whether in electrolytes or in metals, has a definite granular structure, that it consists of an exact number of specks of electricity [electrons] all exactly alike, which in static phenomena are scattered over the surface of the charged body and in current phenomena are drifting along the conductor.

It must not be supposed, however, that the influence of the electron concept is confined to electrical phenomena. As our brief notice of the Zeeman effect has indicated, the electron and its behavior play a very important part in all questions concerning the structure of atoms and the emission of light. Recent indirect evidence indicates that Millikan's original value for the electronic charge is in error by about one half of one per cent. To-day the most probable value of e is taken to be 1.60×10^{-19} coulombs.

The Significance of the Determination of the Electronic Charge

Besides its intrinsic interest, the value of e determines other constants which are of the utmost importance in physics and chemistry. One of these is Avogadro's number N which is the number of molecules of any substance contained in a gram molecular weight of that substance. For example, 2 grams of hydrogen, 32 grams of oxygen, and 18 grams of water each contain N molecules. Faraday's law of electrolysis, page 435, shows that the deposition of one gram-atom of silver requires the passage of 96,490 coulombs of electricity. If the charge carried by each silver ion is the electronic charge, then Avogadro's number

$$N = \frac{96{,}490}{1.6 \times 10^{-19}} = 6.03 \times 10^{23}.$$

This number is of fundamental importance.

Knowing the ratio m/e and the value of e, we can now find the mass m of a cathode ray particle, or electron, for the correctness of Fitzgerald's conjecture that cathode rays are free electrons that have been released from matter has been firmly established. Substituting the numerical values of m/e and e we have

$$m = \frac{m}{c} \times c = 9.1 \times 10^{-31} \text{ kgms}$$

or 9.1×10^{-28} gms. The striking aspect of this result is that it is only $\frac{1}{1840}$ of the mass of the hydrogen atom.

Thus the long sequence of experiments, beginning with Pluecker in 1859 and culminating with Thomson and Millikan in 1900–15, have shown beyond any doubt that one of the constituents of all atoms is the electron. This concept has clarified and unified the fundamental laws of physics and chemistry to a remarkable degree.

Questions for Self-Examination

1. Outline the evolution of the vacuum pump.
2. What is the evidence that cathode rays are negatively charged particles?
3. What is the evidence that an electric current is simply static electricity in motion?
4. Describe Thomson's method of measuring the speeds of cathode particles and the ratio of their charge to their mass.
5. What support did the discovery of the "Zeeman effect" give to the hypothesis of the existence of "electrons"?
6. Describe Millikan's method of measuring the electric charge of the electron.

Problems on Chapter 48

1. (*a*) An electron that is accelerated in a vacuum through a potential difference of 1 volt is said to have kinetic energy of 1 electron-volt. Show that this unit of energy is 1.60×10^{-19} joules.
   (*b*) What is the velocity of an electron whose kinetic energy is 20,000 electron-volts? 8×10^{7} m/sec.
2. What is the velocity of a proton whose kinetic energy is 20,000 electron-volts? 1.9×10^{6} m/sec.
3. Electrons having a speed of 4×10^{7} m/sec move in a uniform transverse magnetic field of .001 webers/m^2. What is the radius of the path? .2 meters.
4. Electrons having a horizontal velocity of 4×10^{7} m/sec move between parallel plates as in Thomson's apparatus (Fig. 488). If the plates are 2 centimeters long and the transverse electric field is 10^{4} newtons/coulomb, what is the downward velocity of the electrons when they emerge from the field? 8×10^{5} m/sec.
5. In Millikan's oil-drop apparatus the intensity of the electric field is 10^{5} newtons/coulomb. An oil drop carrying a charge of two electrons is balanced in the field. What is the mass of the drop? 3×10^{-15} kgm.

CHAPTER 49

X-Rays and Radioactivity

The Announcment of the Discovery of X-Rays

On January 5, 6, and 7, 1896, newspapers in all parts of the world gave prominent notice to a "sensational scientific discovery." The discovery, which was made by Professor Roentgen of Würzburg, Germany, consisted of a new kind of light which was able to penetrate wood, human flesh, and most other opaque objects. With the new rays Roentgen had taken photographs of the bones in a human hand and of pieces of metal inside a closed wooden box. One paper closed its report with the following remark:

> The Press assures its readers that there is no joke or humbug in the matter. It is a serious discovery by a serious German Professor.

At first, scientists were skeptical; they considered it just "another newspaper story." Curious physicians and laymen, however, besieged the laboratories clamoring for X-ray photographs of various parts of the human body. The result was that within a few days practically every major laboratory in the world had verified the astounding discovery. Within one year over a thousand articles on X-rays were published.

Very few scientific discoveries are made in the way Roentgen discovered X-rays, though the general public seems to have an impression that most of them are made that way. The whole development occurred within a few weeks from the time Roentgen got his first "hunch"; it was all done by one man; it was immediately applicable to pressing problems outside of the field of physics; and it promptly caught the popular imagination. Most of the really important scientific discoveries proceed in the opposite patterns.

The Circumstances of the Discovery of X-Rays

When Roentgen began to investigate cathode rays in the autumn of 1895 he first repeated some of the experiments of Lenard and Crookes. His apparatus, a diagram of which is shown in Figure 494, consisted of a "fairly large induction coil" and a cathode ray tube of the Crookes type. It will be recalled that the principal method of observing cathode rays was by means of the fluorescence they produced. In order to be able to observe even a slight fluorescence, Roentgen was working in a dark room and had covered the tube with black paper. When a discharge was sent through the tube, he observed that a screen covered with barium platino-cyanide gave off fluorescent light. The effect was visible even when the screen was two meters away from the tube. Further, the fluorescence occurred *only* while the discharge was passing.

Although the new rays were first observed on November 8, 1895, Roentgen made no announcement of the discovery until December 28, when he submitted his first communication entitled, *On a New Kind of Rays (Preliminary Communication)*.

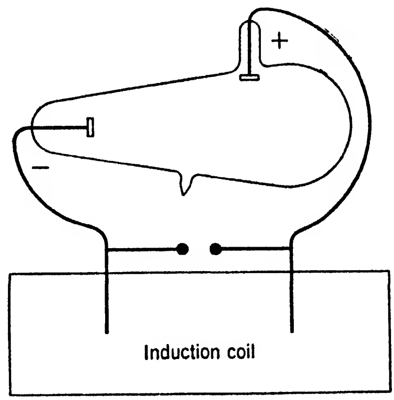


FIG. 494. THE ARRANGEMENT OF THE APPARATUS WITH THE AID OF WHICH ROENTGEN DISCOVERED X-RAYS

During this interval of seven weeks we have a situation which is almost, if not entirely, unique in the history of science. Roentgen was fifty years old and was universally recognized as a competent investigator. His publications, consisting of forty-eight papers, included several which were notable contributions to physics. He had just discovered an entirely new phenomenon and was completely aware of its extreme importance. Naturally he worked feverishly, yet carefully, to explore as many ramifications of the effect as possible.[1] Later

> Mrs. Roentgen said that she had to go through several terrible days. Her husband came late to dinner and usually was in a very bad humor; he ate little, didn't talk at all, and returned to the laboratory immediately after eating.

In his first paper on the new rays Roentgen said (109:3):

> For brevity's sake I shall use the expression, "rays"; and to distinguish them from others of this name I shall call them X-Rays.

He then reported the following observations:

1. All substances are more or less transparent to X-rays; they are able to penetrate several centimeters of wood and 1.5 centimeters of aluminum. Lead, however, only 1.5 millimeters thick, is practically opaque.
2. X-rays affect a photographic plate.
3. No perceptible refraction of the rays could be observed in water, carbon disulphide, hard rubber or aluminum. Hence X-rays cannot be focused by means of lenses. No evidence of regular reflection was found.
4. X-rays are not deviated in a magnetic field. Since cathode rays are deviated by a magnet, the new rays are of an entirely different type.
5. X-rays seem to originate in the point where the cathode rays strike

[1] From Glasser's *Wilhelm Conrad Röntgen*, pp. 129 ff. Courtesy of Charles C. Thomas, publisher.

the glass wall of the discharge tube. The glass, however, is not essential; X-rays are generated when any solid body is struck by cathode rays.

6. X-rays cause the discharge of electrified bodies in air. The rate of discharge increases with the intensity of the X-ray beam.

The Search for Further Properties of X-Rays

Thus in the brief period of two months Roentgen had determined the essential properties of the new rays as well as the method of producing them. The thoroughness of his work is shown by the fact that no new physical properties of X-rays — except perhaps polarization — were discovered in the next seventeen years.

During the next year, Roentgen made a more detailed study of the penetrating power of the rays in different substances. He found that rays from a "hard" tube — that is, one in which the vacuum was high — were more penetrating than those from a soft tube. Further, he showed that the absorbing power of a metal sheet was not simply proportional to its density. Dense metals such as lead absorbed more than aluminum but not in the ratio of their densities.

To Roentgen the fundamental and challenging question was: What is the physical nature of X-rays? Thus in the concluding paragraph of his third paper, which was submitted in March, 1897, he said (109:40):

> Since the beginning of my work on X-rays I have tried repeatedly to obtain diffraction phenomena with them; several times I have obtained with narrow slits etc., phenomena whose appearance reminded one, it is true, of diffraction images; but when by alteration of the conditions of experiment tests were made of the correctness of the explanation of these images by diffraction, it was refuted in every case... I have no experiment to describe from which, with sufficient certainty, I could obtain proof of the existence of diffraction by the X-rays.

There has been much speculation concerning the reasons for Crookes' and Lenard's failure to discover X-rays. In most of their experiments rather intense X-rays were produced, but they failed to detect their presence. Lenard's failure was in part accidental. To detect the cathode rays, he happened to use a different kind of screen than Roentgen used later, coated with a substance which, while sensitive to cathode rays, was not affected by X-rays. Other investigators immediately took up the search for interference and diffraction effects. In spite of their improved apparatus and more intense rays no interference could be found. It was not until 1912 that a brilliant suggestion was made that brought success after seventeen years of failure. This will be outlined on page 610, but first a description must be given of some of the great advances made in the technique of producing and measuring X-rays during the interval.

Roentgen's Tubes

The X-rays emitted by an ordinary Crookes tube are not very intense. Realizing that the X-rays are generated by the impact of cathode rays

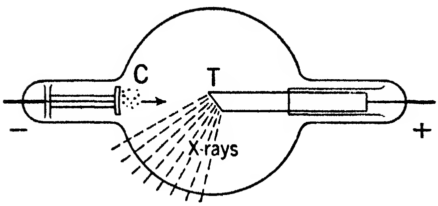


FIG. 495. AN EARLY X-RAY TUBE OF THE "GAS" TYPE

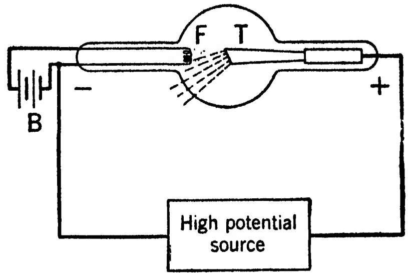


FIG. 496. THE COOLIDGE X-RAY TUBE

upon a solid target Roentgen designed a tube similar to that shown in Figure 495. High-speed electrons from the cathode *C* are focused upon a platinum target *T* which becomes the source of the X-rays. The operation of this type of tube is erratic because it depends upon the gas pressure in the tube. Unfortunately this pressure changes when the tube is used for long intervals. Gas may be adsorbed or emitted by the walls of the tube. If the pressure is too low the discharge will not pass; if the pressure is too high the discharge occurs at such low voltage that the X-rays are weak.

The Coolidge X-Ray Tube

The solution of the problem was supplied by two developments in experimental technique. As was noted in Chapter 48 the advent of the diffusion type of vacuum pump in 1912 made possible the production of higher vacua. With improved vacuums Langmuir had shown that pure tungsten filaments emitted electrons when heated to high temperatures. Encouraged by Langmuir's results, W. D. Coolidge of the General Electric Company designed the tube shown in Figure 496. The tungsten filament *F* is heated by a battery *B* until it emits electrons. When the target or anode *T* is raised to a high positive potential the electrons are attracted and strike the target with high speed. The number of electrons is controlled by the filament temperature and their speed at the instant of impact by the potential difference between *F* and *T*. Tubes of this type operating at high potentials emit much more intense X-rays than Roentgen was able to obtain. This ordinary type of Coolidge tube is used for voltages up to approximately 150,000 volts. Recent tubes for therapeutic work, which operate at a million volts or more, are much longer and require special design.

Measurement of X-Ray Intensities

Roentgen observed that X-rays passing through air rendered it conducting. This effect was studied more carefully by J. J. Thomson and Ernest Rutherford, who had just come from New Zealand to work in the Cavendish Laboratory at Cambridge University in England. They found that when the air between the plates *A* and *B* in Figure 497 was exposed to X-rays a

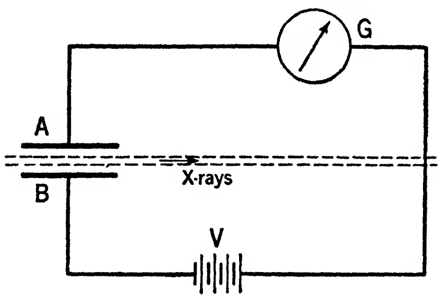


FIG. 497. IONIZATION OF AIR BY X-RAYS

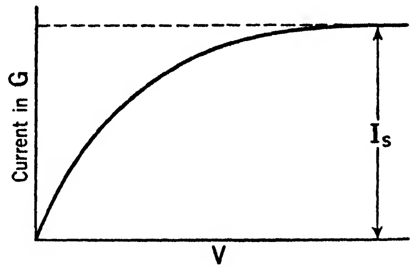


FIG. 498. CURVE OF IONIZATION OF AIR

small current passed through the sensitive galvanometer G. When the applied potential V was increased the current varied as shown in Figure 498. The maximum constant value, I_s, was called the saturation current. The existence of this saturation effect was explained by assuming that when V was high enough the ions produced by the X-rays were swept over to the plates just as rapidly as they were formed. Hence I_s was a measure of the intensity of the ionizing X-rays. This ionization method is still used to measure X-ray intensities.

Diffraction of X-Rays by Crystals

By 1912, most workers in the X-ray field were convinced that the rays were similar to light but of much shorter wave-length — approximately 10^{-9} centimeters in contrast to ordinary light which has a wave-length of about $5 \cdot 10^{-5}$ centimeters. The failure to observe X-ray diffraction patterns was attributed to this extremely small value of the wave-length. For, as optical interference theory shows (page 376), the width of the fringes depends upon the ratio $\frac{\text{wave-length}}{\text{slit width}}$. To make slits narrow enough, or to rule gratings closely enough, to give observable effects with X-rays seemed out of the question.

The difficulty was surmounted by a brilliant suggestion made by Max von Laue of Munich, Germany. One of Laue's friends was working upon the theory of the propagation of light in crystals. In discussing this problem the following question was proposed: Assuming X-rays to be light of very short wave-length, what would occur when a beam passes through a crystal? It happened that Laue was eminently qualified to answer this question. He was an authority on interference effects and knew something of crystallography. After making a careful analysis of the problem, he predicted that a narrow pencil of X-rays would give a definite diffraction pattern after passing through a crystal.

The Experiment of Friedrich and Knipping

Since Laue was a theorist, two of his colleagues, Friedrich and Knipping,

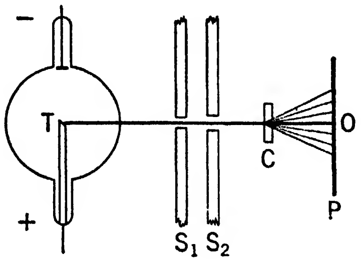


FIG. 499. DIAGRAM OF APPARATUS TO PRODUCE LAUE SPOTS

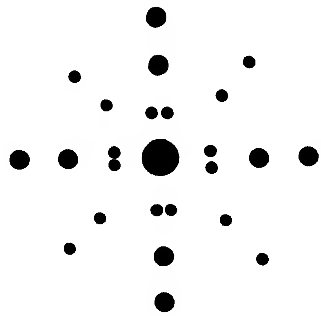


FIG. 500. A TYPICAL LAUE SPOT PATTERN

immediately undertook to test the prediction experimentally. Their apparatus is shown in Figure 499. X-rays from the target T pass through the pinholes in the lead screens S_1 and S_2, and then through the crystal C. Most of the rays strike the photographic plate P at the central point O. A small fraction, however, are diffracted and strike the plate at points which form a regular pattern. After a few unsuccessful attempts a photograph somewhat comparable to Figure 500 was obtained. This definite arrangement of spots on the plate constituted the first definite experimental proof that X-rays gave interference effects similar to those obtained with light. This is a striking example of the importance of a definite hypothesis in stimulating and guiding research. Many persons had sent X-rays through crystals before 1912, but no one had observed any diffraction effects. It was only after Laue's theory had predicted what to look for that the effect was found.

Bragg's Equation for X-Ray Interference

The theory of the formation of the Laue spots cannot be presented in elementary form. This difficulty does not apply, however, to the ingenious method devised by W. H. Bragg in England.

Bragg assumed that the atoms or ions in a crystal are arranged in a regular pattern such as that shown in Figure 501, which represents the sodium and chlorine ions in a crystal of sodium chloride (common salt). The distinction between crystalline and non-crystalline bodies is simply the difference between regular and irregular arrangement of the atoms or ions constituting them. Let us consider a single line of atoms as shown in Figure 502. Under the influence of the incident X-rays, the atoms emit weak X-ra-

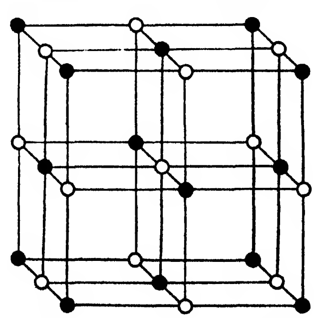


FIG. 501. A CRYSTAL IS MADE UP OF ROWS AND COLUMNS OF ATOMS

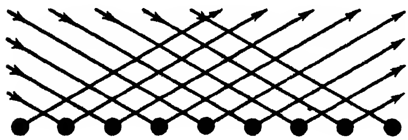


FIG. 502. REFLECTION OF X-RAYS FROM A SINGLE LAYER OF ATOMS IN A CRYSTAL

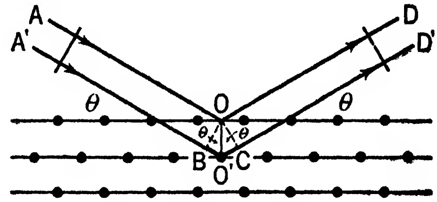


FIG. 503. INTERFERENCE BETWEEN X-RAYS REFLECTED FROM SUCCESSIVE LAYERS IN A CRYSTAL

diation in all directions. These scattered wavelets reinforce one another only in the direction shown, i.e., when the angle of reflection is equal to the angle of incidence. A very small fraction of the X-ray energy, therefore, is said to be "reflected" by a layer of atoms. In the crystal there are thousands of these reflecting layers evenly spaced as in Figure 503. That part of the incident beam which is reflected by the top layer of atoms will travel a distance equal to AOD. That reflected from the second layer will travel $A'O'D'$. The path difference is $BO' + O'C$ since OB and OC are perpendicular to $A'O'$ and $O'D'$. If this path difference is an integral number of wave-lengths of the X-rays, the two reflected beams will reinforce one another. The condition for a strong reflection therefore is that

$$n\lambda = BO' + O'C. \tag{1}$$

But from Figure 503, $BO' + O'C = 2(OO') \sin \theta$. If the distance between adjacent layers of the crystal is called d, the equation (1) can be written

$$n\lambda = 2\,d \sin \theta \tag{2}$$

where n is an integer and λ the wave-length of the X-rays that are strongly reflected at the angle θ. Equation (2), which is called Bragg's equation, is one of the most useful relations in contemporary physics. Having found θ experimentally we can calculate λ if d is known, or find d if λ is known. In this way we obtain information concerning the X-rays and detailed knowledge about the arrangement of atoms in the crystal. This type of information is of utmost importance in chemistry, physics, biochemistry, metallurgy, and mineralogy.

Bragg's Determination of X-Ray Wave-Lengths

In order to apply equation (2), Bragg constructed the X-ray spectrometer shown in Figure 504. The first slit, S, permits a narrow pencil of X-rays from the tube to fall upon the crystal C. The "reflected" rays pass through the second slit, S, and enter the ionization chamber, I. The intensity of this ionization current is indicated by a sensitive galvanometer. As the angle of incidence is varied, certain values are found which give intense reflections. Bragg interpreted these values as those which satisfy equation (2).

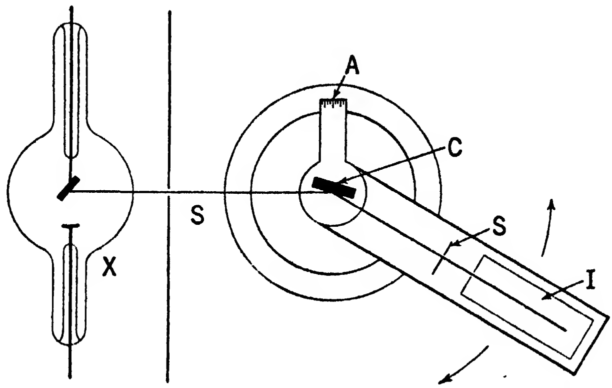


FIG. 504. THE PRINCIPLE OF THE BRAGG X-RAY SPECTROMETER

Knowing the density and structure of the crystal, he was able to calculate the distance d between the crystal planes. Referring to Figure 501, we see that a crystal of salt is made up of a large number of tiny cubes. The volume of each cube is d^3. In the interior of the crystal each ion is shared by eight cubes, and a complete cube requires eight ions. Therefore the total number of cubes is equal to the number of ions. Since a molecule of sodium chloride consists of one sodium atom and one chlorine atom, the number of molecules in the crystal is one half the number of atoms or ions. From the laws of chemistry we know that 6.03×10^{23} molecules of sodium chloride have a mass of 58.45 grams. A crystal of this mass has a volume $\frac{58.45}{2.17}$ cubic centimeters since its density is 2.17 gms/cc. The number of tiny cubes in this crystal is $2 \times 6.03 \times 10^{23}$. We have, therefore,

$$(2 \times 6.03 \times 10^{23})d^3 = \frac{58.45}{2.17},$$

$$\text{or } d = 2.81 \times 10^{-8} \text{ cms.}$$

Using this result and equation (2), Bragg was able to determine X-ray wave-lengths. He found that a rhodium target emitted two strong rays of wave-length 0.607×10^{-8} and 0.533×10^{-8} centimeters or 0.607×10^{-11} and 0.533×10^{-11} meters. Similarly palladium, copper, and nickel emitted two-line X-ray spectra. These intense X-ray lines of definite wave-length are called characteristic X-rays since they are characteristic of the target. A tube also emits general X-radiation which depends largely upon the tube voltage and only slightly upon the material of the target.

Absorption of X-Rays

In his pioneer experiments Roentgen observed that the intensity of a beam of X-rays suffers greater diminution in passing through a sheet of lead than through the same thickness of aluminum. The amounts of absorption, however, are not directly proportional to the densities of the two metals.

Later investigations using homogeneous beams — that is, X-rays of a single wave-length — showed that the intensity, I, of the beam after passing through x centimeters of absorbing material is

$$I = I_0 \epsilon^{-\mu x} \tag{3}$$

where I_0 is the intensity of the original incident beam and μ is a constant called the *linear coefficient of absorption*; "ϵ" is the base of natural logarithms (see exponential table in Appendix). Equation (3) can also be written in the form

$$I = I_0 \epsilon^{-\frac{\mu}{\rho}(\rho x)} \tag{4}$$

where ρ is the density of the absorbing screen. The quotient μ/ρ is known as the *mass absorption coefficient.* Its numerical value is a function of the atomic number (see page 639) of the absorbing material and the wave-length of the X-rays.

The Discovery of Radioactivity

The discovery of radioactivity by Henri Becquerel in 1896 is another instance of an important discovery resulting from a fallacious hypothesis. On January 20, 1896, the mathematician Poincaré reported Roentgen's discovery of X-rays to the French Academy of Science. In the discussion which followed it was brought out that the X-rays emanated from the part of the glass discharge tube which fluoresced strongly. It immediately occurred to Becquerel and others that X-rays might be related to fluorescence and phosphorescence. Becquerel was doubly fortunate in having the necessary materials and training to test this hypothesis. He was the son and grandson of eminent physicists who had made important contributions in the field of phosphorescence.

After several substances had given negative results, Becquerel happened to test a compound of uranium. His method is best described by his own words (77:610):

> I wrapped a Lumière photographic plate with bromized emulsion with two sheets of thick black paper, so thick that the plate did not become clouded by exposure to the sun for a whole day. I placed on the paper a plate of the phosphorescent substance, and exposed the whole thing to the sun for several hours. When I developed the photographic plate I saw the silhouette of the phosphorescent substance in black on the negative. If I placed between the phosphorescent substance and the paper a coin or a metallic screen pierced with an open-work design, the image of these objects appeared on the negative....
>
> We may therefore conclude from these experiments that the phosphorescent substance in question emits radiations which penetrate paper that is opaque to light, and reduces silver salts.

This result was reported on February 24, 1896. One week later, on March 2, he had more exciting results to report (77:611):

> I particularly insist on the following fact, which appears to me exceedingly important and not in accord with the phenomena which one might expect to observe: the same encrusted crystals placed with respect to the photographic plates in the same conditions and acting through the same screens, but protected from the excitation of incident rays and kept in the dark, still produce the same photographic effects. I may relate how I was led to make this observation: among the preceding experiments some had been made ready on Wednesday the 26th and Thursday the 27th of February and as on those days the sun only showed itself intermittently I kept my arrangements all prepared and put back the holders in the dark in the drawer of the case, and left in place the crusts of uranium salt. Since the sun did not show itself again for several days I developed the photographic plates on the 1st of March, expecting to find the images very feeble. The silhouettes appeared in the contrary with great intensity. I at once thought that the action might be able to go on in the dark.

Further tests showed that Becquerel was correct. The activity of the substance did not depend upon exposure to sunlight. It was soon found that all minerals containing uranium spontaneously emitted the penetrating rays. By March 9, he had shown that the uranium rays, like X-rays, were able to ionize air. This property gave a rapid and convenient method of testing samples for the presence of "Becquerel rays," as they were called.

The interest in X-rays was so intense in 1896 that the general public and most scientists paid little attention to Becquerel's discovery; it seemed to be of minor importance. In Paris, however, a young Polish physicist, Marie Curie, became interested in the new rays and set out to test the activity of all the elements and of many minerals. She soon found that thorium was the only other known element that was active. While observing the ionization produced by certain uranium ores she noticed that the effect was three or four times as great as that normally obtained from the same amount of uranium. She attributed this intense activity to the presence of a new, hitherto unknown element, and began the difficult task of separating it. Her husband, Pierre Curie, realizing the importance of this work gave up his own research in piezoelectricity and assisted her.

After months of tedious work, they separated a substance which had an activity about four hundred times as great as that of uranium. This new metal, which was similar to bismuth in its chemical properties, was called ***polonium***, after the native country of Marie Curie.

The Discovery of Radium

A few months later, in December, 1898, the Curies reported the discovery of a new element which was much more active than polonium (77:615–16):

> We believe . . . that this substance, although for the most part consisting of barium, contains in addition a new element which gives it its radioac-

tivity and which furthermore is very near barium in its chemical properties.... M. Demarcay has found in the spectrum a ray which seems not to belong to any known element. This ray (of the new substance), which is scarcely visible in the chloride that is 60 times more active than uranium, becomes strongly marked in the chloride that was enriched by fractionation until its activity was 900 times that of uranium. The intensity of this ray increases at the same time as the radioactivity, and this, we think, is a strong reason for attributing it to the radioactive part of our substance.

The various reasons which we have presented lead us to believe that the new radioactive substance contains a new element, to which we propose to give the name *radium*.

The Nature of the "Rays" from Radium

The availability of more active sources of Becquerel rays containing polonium and radium led to a further study of the nature of the rays. At first Becquerel and others had assumed that the radiation consisted of weak X-rays. In 1899, however, Giesil in Germany and Becquerel showed that part of the radiation emitted by radium was deviated in a magnetic field. The direction of this deviation suggested that this part consisted of high-speed electrons. By deflecting the rays in an electric field as Thomson did with cathode rays, he was able to calculate the ratio $\frac{\text{charge}}{\text{mass}}$ and to prove definitely that these rays were electrons. His first approximate value for the velocity of the emitted electrons was $1.6 \cdot 10^8$ meters/sec, which is about one half the velocity of light.

Alpha, Beta, and Gamma Rays

In January, 1899, the young New Zealander, Rutherford,[1] found that the rays from uranium were complex and that

> there are present at least two distinct types of radiation — one that is very easily absorbed, which will be termed for convenience α radiation, and the other of a more penetrating character which will be termed the β radiation.

Later Villard observed some very penetrating radiations emitted by radium and called them γ (gamma) rays. Although later experimental evidence has shown that the α, β, and γ rays are not new and distinct entities, the nomenclature has remained in common usage. A substance which spontaneously emits any of these rays is said to be *radioactive*. It should be emphasized that "radio" in this case refers to radium and does not imply any relationship whatever with wireless telephony, which we in America call radio.

The Nature of Alpha Rays

Because they were unable to deflect the α rays in a magnetic field Becquerel and others assumed that the rays were easily absorbed X-rays.

[1] *Philosophical Magazine*, January, 1899, p. 109.

Since Roentgen had shown that X-rays are produced when fast electrons are suddenly stopped it was natural to assume that the β particles from radium would be accompanied by X-rays.

For several years no one questioned this interpretation of the nature of α rays. In 1902, however, Rutherford was impressed by the fact that the inert gas helium was always found occluded in minerals which contained uranium or thorium. This suggested that helium was related in some way to radioactivity. He suspected that the alpha rays were helium atoms or ions and set out to prove it.

There is no evidence that Rutherford or any one else at that time thought that this search would lead to results of extraordinary importance. Yet such proved to be the case, for the study of the nature of alpha rays and their interaction with matter led Rutherford step by step to his nuclear model of the atom and the transmutation of elements. As we shall see later, these new ideas have altered our fundamental concepts in physics and chemistry as profoundly as Newton's mechanics altered natural philosophy in the seventeenth and eighteenth centuries.

By using electric and magnetic fields of greater intensity Rutherford, in 1903, succeeded in deflecting the α rays. The observed deviations showed that the α particles were positively charged and that their ratio of charge to mass was $4.8 \times 10^7 \frac{\text{coulombs}}{\text{kgm}}$. This ratio is approximately the same as that for an atom of helium which has acquired a charge of $+ 2\,e$ by losing two of its normal quota of electrons. The experimental evidence, however, was not conclusive. For a molecule of hydrogen that had lost one electron would have the same value of $\frac{\text{charge}}{\text{mass}}$. This alternative explanation did not deter Rutherford from proceeding on the assumption that the α rays were high-speed helium atoms carrying a charge of $+ 2\,e$.

The year 1903 was notable for two other important advances in the new field of radioactivity — the heating effect of radium and the theory of radioactive transformation. Curie and Laborde showed that a small amount of a radium salt placed in a cavity in a block of lead maintained itself at a higher temperature than the surrounding air. Since the heat is conducted and radiated away by the lead, the existence of this temperature difference shows that heat energy is being given off continually and spontaneously by the radium. Measurements of temperature difference and rate of heat loss showed that one gram of radium gave off heat at the approximate rate of .1 calorie per hour. Rutherford immediately suggested that this heating effect was simply the manifestation of the kinetic energy of the α particles emitted by the radium. Knowing approximately the number of α particles emitted per second by one gram of radium, he was able to calculate their total kinetic energy. This energy was found to be of the same order of magnitude as the heat given off, as he had suspected. This numerical

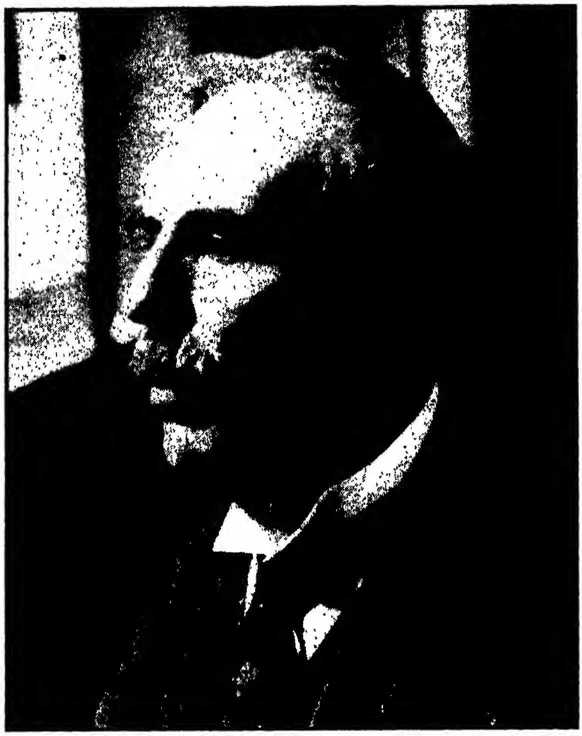


FIG. 505. SIR ERNEST RUTHERFORD (1871–1937)

agreement gave further support to Rutherford's hypothesis that helium was given off in radioactive processes.

The Theory of Radioactive Transformations

By 1903 a great mass of unrelated and puzzling facts concerning radioactive change had been accumulated. Many of the observed phenomena seemed strange and inconsistent. This chaotic material was reduced to order by the theory of radioactive transformation proposed by Rutherford and Soddy. This theory, which is the product of a remarkable scientific imagination, is notable for its originality, its completeness, and its simplicity. The fundamental idea is that the atoms of a radioactive substance are unstable. In a particular time interval a certain per cent of the atoms will explode, ejecting high-speed α or β particles in the process. The ejection of one of these particles, however, leaves a residue which differs from the original atom. The theory assumes that this residue is a new chemical element. Thus we have chemical elements changing into different elements by a process of spontaneous explosions. Although we are able to observe and follow the interesting transformations we are not able to control them. The rate at which they explode is entirely independent of temperature, pressure, and chemical forces. This shows that the radioactive

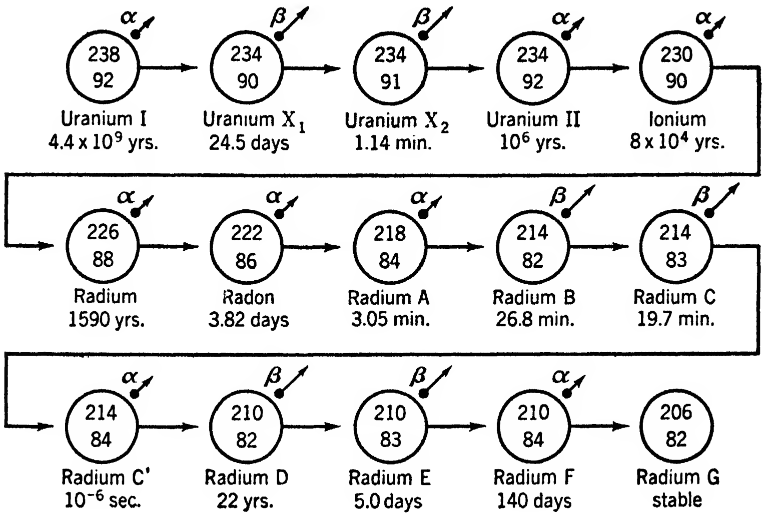


FIG. 506. TRANSFORMATION SERIES OF THE URANIUM FAMILY

(The upper number in each circle is the atomic mass, the lower the atomic number. Below each circle is the name and the "half life" of the corresponding element.)

processes are quite different from ordinary chemical processes. According to the theory the complexity of the observed facts arises from long series of disintegrations which follow one another in a particular family of active elements. The sequence of transformations in the case of the uranium family is shown in Figure 506. From the figure we see that when a uranium atom, atomic weight 238, emits an α particle, mass 4, the residue is an atom of uranium X_1, mass 234. Similarly, uranium X_1 emits a β particle and creates a new atom uranium X_2 which has different properties than the mother substance. The final product of the sequence of changes is radium G, which is not radioactive and cannot be distinguished chemically from ordinary lead.

Although we cannot retard or accelerate the transformation process, the rate of decay of a particular element can be represented by an equation of the form

$$N = N_0 \epsilon^{-\lambda t} \tag{5}$$

where N = number of atoms which have not exploded after a time interval t, and N_0 is the number of atoms at the beginning of the interval. ϵ is 2.718 (the base of natural logarithms) and λ is a constant which is characteristic of the active element.[1] For substances such as radium which transform slowly, λ is equal to the fraction of the atoms which break up in one second. Since λ is constant, if one half of the original atoms remain unchanged after a time T, one fourth will remain after $2\ T$, one eighth after $3\ T$, and so on. The time T is called the *half-period* of the substance. As shown in Figure

[1] For values of this exponential function see the table on page xxiv of the Appendix.

506, the half-periods of the members of the uranium family vary enormously. For radium it is 1590 years; for radium A 3.05 minutes; and for radium C' only one millionth of a second.

The Energies of Alpha Particles

When an atom of radium explodes, the α particle is always ejected with approximately the same speed, $1.52 \cdot 10^9$ cms/sec. This is about one twentieth of the velocity of light or over 9000 miles per sec. In other words, in a vacuum the α particle would travel this distance in one second. In air at atmospheric pressure, however, it travels only 3.3 centimeters. Owing to its positive charge it removes electrons from many of the molecules which it encounters and thus ionizes the air.

Although the speed of the α particle is not as great as that of electrons in an X-ray tube, the kinetic energy is much greater. For the mass of the α particle is 7320 times the electronic mass. To give an electron a kinetic energy equal to that of an α particle from radium would require a tube operating at 4,800,000 volts. We say, therefore, that the energy of the α particle is 4.8 million electron-volts. The α particles from radium C' have energies of over 7 million electron-volts. Using these energetic particles Rutherford later was able to probe the central cores of atoms and to determine their structure.

The Significance of the Transformation Theory

The importance of the Rutherford-Soddy theory of radioactive transformation cannot be overemphasized. Before 1903 it was considered definitely established that atoms were constant and fixed. The idea that they spontaneously exploded and changed into new elements was contrary to all the principles of chemistry. It is not surprising, therefore, to find that the theory encountered strong opposition. It was too radical for the chemists and the older generation of physicists. Instead of deterring Rutherford, the criticism seemed to stimulate him. Assisted by capable and enthusiastic students, he piled up such an overwhelming mass of experimental evidence that within a few years even his most severe critics were convinced. But even more remarkable is the fact that all the new data which have come to light since 1903 are in complete harmony with the theory as it was orginally proposed. No essential modifications or alterations have been necessary.

Questions for Self-Examination

1. Describe the circumstances leading up to Roentgen's identification of X-rays.
2. Outline the evolution of X-ray tubes.
3. Tell how natural crystals gave evidence of wave-length structure of X-rays.
4. How did Bragg deduce the distances between atoms of crystals?

5. Describe the circumstances leading up to Becquerel's discovery of radioactivity and tell how the Curies extended that discovery.
6. Describe the characteristics of the three types of radioactive radiation.
7. Describe successive stages in a representative series of radioactive transformations.

Problems on Chapter 49

1. What is the value of the ratio $\frac{\text{charge}}{\text{mass}}$ for an α particle? $4. \times 10^7 \frac{\text{coulomb}}{\text{kgm}}$.
2. Alpha particles emitted by radium *A* have a velocity of 1.7×10^7 m/sec. What intensity of magnetic field will cause them to move in a circle of .50 meters radius? .7 webers/m^2.
3. Express the kinetic energy of the α particles from radium *A* in electron-volts. 6×10^6 electron-volts.
4. In making observations with a Bragg spectrometer (p. 613), an intense reflected beam is obtained when the angle $\theta = 7° \ 16'$. If the reflecting crystal is rock-salt, what is the wave-length of the X-rays? $.7 \times 10^{-8}$ cms.
5. The decay constant, λ in equation (3), for radon is .181 day^{-1}. What percentages of a given amount of radon will remain after 2, 4, 6, and 10 days respectively? (A table of ϵ^{-x} is given in the Appendix.) .69, .48, .34, .16.

CHAPTER 50

Quantum Theory and Atomic Structure

The "Black Body"

In 1859, G. Kirchhoff showed that a good absorber of radiant heat is also a good emitter. It follows from Kirchhoff's analysis that a perfect *black body*, that is, one which absorbs all the radiation that falls upon it, is the best possible emitter of radiant energy. The theory also showed that the total radiation from a black body depends only on the temperature of the body and not on its chemical or physical nature.

Applying thermodynamical methods to radiation, Stefan and Boltzmann showed in 1879 and 1884 respectively that the radiant energy emitted per second by a black body was proportional to the fourth power of its absolute temperature. Certain applications of this law were discussed on page 188.

No natural solid substance is known which absorbs *all* the radiation that falls upon it; even lampblack reflects about 1 per cent of the incident radiant energy. Consequently, when Lummer and Wien in 1895 began to investigate black body radiation, their first problem was to devise a satisfactory black body. Their ingenious solution of the problem is shown in principle in Figure 507. The device consists of a small furnace whose walls are maintained at a uniform temperature. If radiation enters the enclosure through the small opening *O*, it is scattered by repeated reflections from the walls, and only a very small fraction of it will escape through *O*. The opening *O*, therefore, has the properties of a black body. When equilibrium is reached, the walls are emitting radiation as well as absorbing it. The radiation which escapes through the opening has the properties of black body radiation. It depends only on the uniform temperature of the walls of the enclosure, not on the nature of the furnace walls.

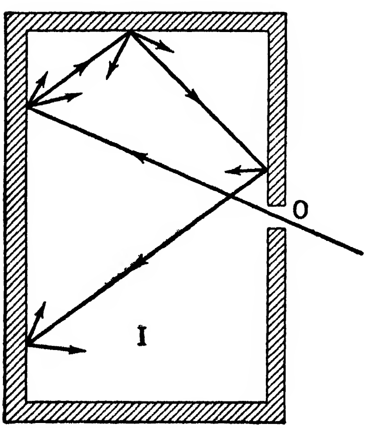


FIG. 507. LABORATORY "BLACK BODY"

Experimental Results of Lummer and Pringsheim

In the years 1897–99, Lummer and Pringsheim in Berlin made careful measurements of the distribution of the energy of black body radiation through the spectrum. The radiation from a black body at high temperature fell upon a prism spectrometer and was spread out into a spectrum. The intensities of various regions of the spectrum were then measured with a sensitive thermopile. Graphical representation of their results gave curves similar to those shown in Figure 508.[1]

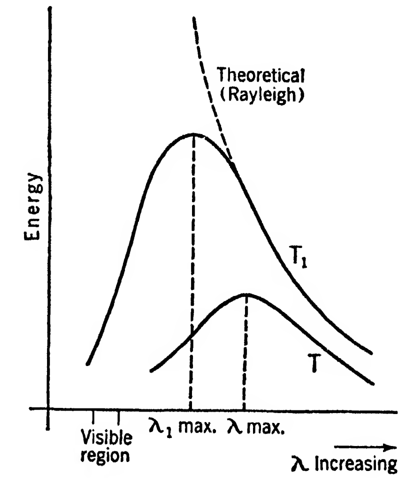


FIG. 508. SPECTRAL ENERGY DISTRIBUTION OF BLACK BODY RADIATION

The point of principal present importance is not so much the different heights of the curves for different temperatures as the fact that the wave-length at which the maximum energy concentration occurs is progressively shorter as the temperature rises. For each temperature there is a definite wave-length $\lambda_{max.}$ which has maximum energy. The progress of the maximum energy region toward the shorter wave-lengths accounts for the fact that at moderate temperature a furnace appears red, but when the temperature increases, it appears yellow and finally white. In 1893 Wien showed by thermodynamical arguments that the wave-length of the maximum in Figure 508, $\lambda_{max.}$ was inversely proportional to the absolute temperature, or

$$\lambda_{max.} T = A \qquad (1)$$

where A is a constant. This relation is known as Wien's Displacement Law.

The experiments of Lummer and Pringsheim verified Wien's Law and gave the numerical value of A as 2.89×10^7 if $\lambda_{max.}$ is expressed in angstroms and T in absolute Centigrade degrees. If it is assumed that the sun radiates as a "black body," equation (1) gives a convenient method of calculating the temperature of that body. Measurement of the energy distribution of sunlight gives a value of $\lambda_{max.}$ of 4700 A. Substituting this value in equation (1) we find the temperature of the sun's surface to be 6150° absolute. This result cannot be considered exact since the distribution of energy in sunlight differs slightly from that of a black body.

[1] Compare Figure 317. Curves of black body radiation for temperatures from 373° K to about 6000° K are shown in that figure. Actual temperatures are not stated there and the vertical energy scale is not uniform. If it had been, the curve for the sun would have been many thousands of times as high as that for boiling water.

Planck's Quantum Theory of Radiation

Although Wien's Law is very important, it gives no information whatever about the detailed shape of the energy distribution curve. This question was carefully studied by Lord Rayleigh in England, who assumed that the vibrating electrons in the atoms emitted and absorbed radiation in accordance with Maxwell's laws, which had been verified in regard to radio waves. His further assumptions involved well-established principles of heat and mechanics. The final result, however — shown by the broken curve in Figure 508 — was in striking disagreement with the experimental results. Other theorists examined the question, and all agreed that Rayleigh's method and results were theoretically sound. The laws of electricity, heat, and mechanics established during the nineteenth century could give no other answer than that obtained by Rayleigh — and it was definitely wrong.

In 1899 the contradiction between theory and experiment was removed by a brilliant and radical suggestion made by Max Planck. Planck assumed that the atomic vibrator could absorb or emit energy only in finite amounts or *quanta*. Further, the size of this quantum of energy was not the same for all vibrators but was equal to a constant h multiplied by the frequency ν of the oscillator. The constant h is called Planck's constant and has the same value for all substances. It is one of the so-called universal constants.

The quantum theory of Planck gave the correct answer, but it represented a drastic break with traditional methods in physics. According to Planck, a vibrating electron in an atom could have amounts of energy equal to $h\nu$, $2\ h\nu$, $3\ h\nu$, and so forth, but could never have a total energy equal to, say, $2.2\ h\nu$ or any other *non-integral multiple* of the fundamental quantum. This concept that energy can be added or removed only in finite "chunks" or quanta is in direct contradiction with classical Newtonian mechanics and Maxwell's electromagnetic theory. One of Newton's fundamental maxims was that "Natura non saltus facit." (Nature does not provide discontinuities.)

Naturally such a radical theory encountered strong opposition. All efforts to find a less violent remedy, however, were unsuccessful. The quantum idea spread to other fields of physics, and when the first Solvay Conference met in Brussels in 1911 to discuss the major problems of physics, the topic was "The Quantum Theory of Radiation." By this time the leading theorists of the world were convinced that it was not possible to derive the correct formula for black body radiation without making use of Planck's quantum concept.

Photoelectric Effect

The most direct and convincing evidence for the quantum theory came from the study of the *photoelectric effect*, the principle upon which the light-beam receiver of Chapter 47 was based. In the course of his experiments on electric waves in 1887, Hertz observed that a spark discharge between two metal spheres was able to jump a greater distance when the spheres were exposed to the light from another spark. A year later, Hallwachs showed that a negatively charged zinc plate lost its charge when it was illuminated with ultra-violet light. If the plate were charged positively, however, the loss of charge did not occur. Later experiments indicated that the effect existed even when the plate was in an evacuated tube.

After the discovery of the electron in 1897 the general features of photoelectric emission became clear. When light of wave-length less than a certain limiting value falls upon a metal, electrons are emitted from the surface of the metal. Radiation of wave-length greater than this limiting value, or *threshold*, as it is called, produces no emission. The threshold frequency or wave-length is characteristic of the emitting substance. For the more common metals such as copper, zinc, and silver it is in the ultraviolet region, between 2500 and 3900 angstrom units. For the alkali metals, sodium, potassium, and caesium, it is in the region 7000 to 20,000 angstroms. Hence, these metals are sensitive to visible light and are therefore commonly used in photoelectric devices.

Einstein's Theory of the Photoelectric Effect

In 1902, Lenard made the astonishing discovery that the maximum velocity of the photoelectrons was independent of the intensity of incident light. It is true that the number of electrons emitted per second was found proportional to the light intensity, but increasing or decreasing the intensity had no effect upon the speed of emission of the fastest electrons. This maximum speed, however, did depend upon the frequency of the incident light. Three years later Albert Einstein, who was then an examiner in the Swiss Patent Office, pointed out that Lenard's results could be explained in a very simple manner by making use of Planck's quantum concept. Einstein's explanation involved three assumptions. First, that the energy in a beam of light was concentrated in very small bundles or quanta. These quanta are now called photons. Second, that the energy of an individual photon was $h\nu$ where ν is the frequency of the incident light. Third, that in the photoelectric effect all the energy of one photon was transferred to a single electron, that is, one electron could never absorb half the energy of a photon or that of two or three photons. On the basis of these hypotheses the conservation of energy can be written in the form

$$\frac{1}{2} mv^2_{\text{max.}} = h\nu - W \tag{2}$$

where $v_{\text{max.}}$ is the maximum speed of the emitted electrons, ν is the frequency of the incident light, and W is the work required to remove an electron from the metal. If ν_0 is the frequency of the light corresponding to the threshold, equation (2) becomes

$$0 = h\nu_0 - W$$

or

$$W = h\nu_0 \tag{3}$$

and equation (2) becomes

$$\tfrac{1}{2}\, mv^2{}_{\text{max.}} = h\nu - h\nu_0. \tag{4}$$

Equation (3) shows that the work required to remove an electron from a metal is equal to the threshold frequency ν_0 multiplied by Planck's constant h. This quantity is called the *work function* of the metal.

Since the kinetic energy of an electron can never be negative, equation (4) explains the observed fact that when the frequency of the incident light ν is less than ν_0 no photoelectrons are emitted.

Experimental Support for Einstein's Equation

Lenard's original experiments were not very precise. Owing to this fact and to the radical nature of Einstein's assumptions, the theory was not generally accepted for several years.

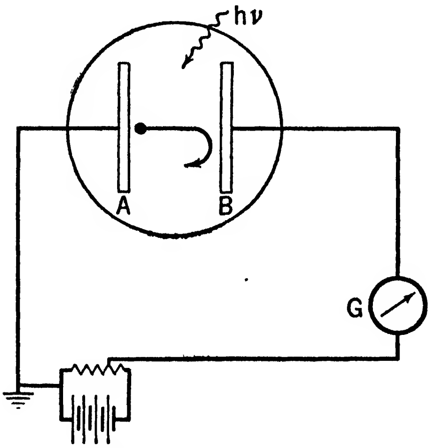


FIG. 509. STUDY OF PHOTOELECTRIC EMISSION

The first convincing experimental test of equation (4) was reported by Millikan in 1916. The principal features of his method are shown in Figure 509. Light that has been made monochromatic by passing through a prism falls upon the plate A. When the potential of the plate B is positive the electrons are attracted and the number striking B per second is measured with a sensitive galvanometer G. If, however, the potential of B is made more and more negative, a value V_s is reached where all the electrons including the fastest ones are turned back and the current in G is zero. This stopping potential V_s enables us to evaluate the kinetic energy of the fastest electrons. For the principle of conservation of energy gives

$$eV_s = \tfrac{1}{2}\, mv^2{}_{\text{max.}} \tag{5}$$

where e is the electronic charge. Equation (4) now becomes

$$eV_s = h(\nu - \nu_0). \tag{6}$$

By observing V_s for different values of ν, Millikan found that the stopping potentials were directly proportional to the frequency of the light as is predicted by equation (6). If V_s is plotted as a function of ν as in Figure 510, the slope of the line is h/e. Since the electronic charge was known from the oil-drop experiment, Millikan found h to be 6.55×10^{-27} erg sec or 6.55×10^{-34} joule sec. The agreement between this result and Planck's value of 6.52×10^{-34} joule sec, which was obtained sixteen years earlier, is a very convincing argument in favor of Einstein's photoelectric equation.

At first sight, it is surprising that a constant so small as h can be detected experimentally and can be a significant factor in most physical and chemical processes. In this connection, it must be remembered that h usually occurs in the product $h\nu$. The frequency ν can be very large. For visible light having a wave-length of 6000 angstroms, ν is 5×10^{14}, and the quantum of energy $h\nu$ is approximately 33×10^{-20} joules. Since one electron-volt is 1.6×10^{-19} joules, the energy of this quantum is 2.06 electron-volts. In the case of penetrating X-rays and of gamma rays from radium, the energy of a quantum of radiation may be of the order of a million electron-volts.

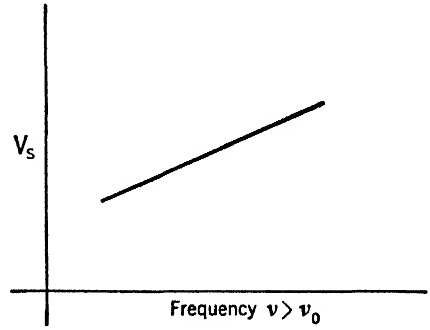


FIG. 510. GRAPH OF EINSTEIN'S PHOTOELECTRIC EQUATION

Einstein's Equation in the X-Ray Region

When they are irradiated with X-rays, all substances emit photoelectrons. In this case the electron speeds are so high that they are easily detected. By applying a magnetic field and observing the curvature of the electron paths, De Broglie in 1921 determined the velocities of emission and showed that equation (4) held for X-ray wave-lengths. In the case of this high frequency radiation the term $h\nu_0$ in equation (4) is negligible in comparison with $h\nu$, and equation (4) becomes

$$\tfrac{1}{2}\, mv^2_{\text{max.}} = h\nu. \tag{7}$$

The Inverse of the Photoelectric Effect

In the phenomena described above, a metallic surface emitted electrons when it was illuminated. The question naturally occurs: Does the inverse effect exist? That is, if electrons having kinetic energy $\frac{1}{2}mv^2$ strike a metallic surface, will radiation be produced having a frequency ν which satisfies equation (7)? The discoveries of Laue and Bragg described in Chapter 49 enable us to answer this question. If we observe the shortest wave-length, and hence the highest frequency, of X-rays emitted by a tube when the target is bombarded with electrons having a known kinetic energy, it is possible to calculate h from equation (7). These values of Planck's constant are in excellent agreement with those found by other methods. All the experimental evidence shows conclusively that equation (7) represents a very general relationship. It applies to all transformations between radiant energy and electronic energy irrespective of the direction of the process. Thus Planck's quantum concept was found to have a much broader field of application than was originally suspected. Full appreciation of its significance, however, did not come till Bohr applied it to the problem of atomic structure.

Counting Atoms

In following the development of physics in the eighteenth and nineteenth centuries, we have encountered the word "atom" several times. The concept of tiny perfectly elastic billiard-ball atoms proved to be very fruitful in interpreting many phenomena in the subject of heat. Yet many scientists refused to admit the existence of atoms. To them the atom was a theoretical concept and nothing more. Furthermore, they challenged the enthusiastic atomists to demonstrate the existence of individual atoms. Owing to the experimental genius of two men, C. T. R. Wilson and Hans Geiger, the challenge has been met. As K. K. Darrow has aptly remarked (34:107):

> One of the things which distinguishes ours from all earlier generations is this, that *we have seen our atoms.*

This statement requires some explanation. It does not mean that we can observe a single stationary atom with a microscope as we can a microorganism or grain of sand. But it does mean that the path of a single high-speed atom can be observed and photographed just as the trail of a shooting star can be observed.

The Scintillation Method

In Chapter 49 we saw that the kinetic energy of a single alpha particle was very great. In 1903, Sir William Crookes, and also Elster and Geitel, observed faint short flashes of light when alpha particles bombarded a zinc sulphide screen. The short flashes, or scintillations, appeared to occur at random and could easily be seen with a low-power magnifying glass. It was assumed that each flash was produced by a single alpha particle. This assumption was later confirmed by Geiger's electrical method of detecting the particles.

It is a curious fact that not all zinc sulphide crystals show these scintillations. Pure crystals show none at all since a very small proportion of certain impurities must be present for the scintillations to be visible. Crystals containing one ten-thousandth part of certain copper salts give best results. The flashes are so faint and so short in duration that it is difficult to count them accurately. For many years, however, it was the only method of investigating the details of atomic phenomena.

The Cloud Chamber of C. T. R. Wilson

In Chapter 48 reference was made to C. T. R. Wilson's method of producing clouds in ionized air. As early as 1899, he had found that in a supersaturated vapor every ion becomes the center of a tiny droplet, which was visible under proper illumination. His first cloud expansion chamber, which has served as the prototype for all subsequent developments, was built in 1912. The principle of the chamber is shown in Figure 511. The chamber *C* consists of a short cylinder closed at the top with a glass plate and at the bottom with a movable piston *P*. The saturated air in *C* is cooled and becomes supersaturated when the piston *P* is suddenly moved downward. Water droplets form rapidly on the ions produced by the alpha and beta particles from the minute amount of radioactive material on the wire *R*. Under the intense illumination from *L* the droplets appear as bright points against a black background and may be observed or photographed.

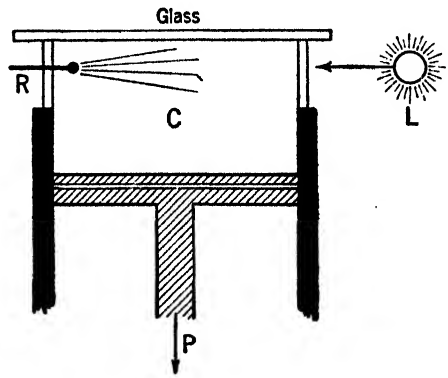


FIG. 511. PRINCIPLE OF WILSON'S CLOUD EXPANSION CHAMBER

As indicated in Figure 512 the tracks produced by alpha particles are straight lines radiating from the source *R*. Owing to their large mass and great energy, these particles are able to ionize the molecules which they encounter without being deviated from their course. In a few instances, however, a sharp kink is observed near the end of a track. The study of these kinks has yielded important information concerning the inner structure of atoms.

Figure 513 is a photograph of tracks produced by beta rays from Radium *E*. The beta particles are moving upward and the tracks are curved owing to the presence of a magnetic field of approximately .03 webers/m^2. Since the number of ions produced per centimeter by an electron is only a small fraction of the number produced by an alpha particle these tracks are very fine broken lines instead of heavy continuous ones.

Wilson's apparatus enables us not only to detect and observe the behavior of a single atomic particle, but to identify it as well. In commenting on the cloud chamber method, Lord Rutherford[1] said:

> To the period 1895–1912 belongs the development of an instrument which to my mind is the most original and wonderful in scientific history. I refer to the cloud or expansion chamber of C. T. R. Wilson.... It was a wonderful advance to be able to see, so to speak, the details of the adventures of these particles in their flight through the gas. Any one with imagination who has seen the beautiful stereoscopic photographs of the trails of swift alpha particles, protons and electrons cannot but marvel at the perfection of detail with which their short but strenuous lives are recorded.

As Figure 513 shows, when the expansion chamber is placed in a magnetic

[1] *Nature, 138,* 865 (1936).

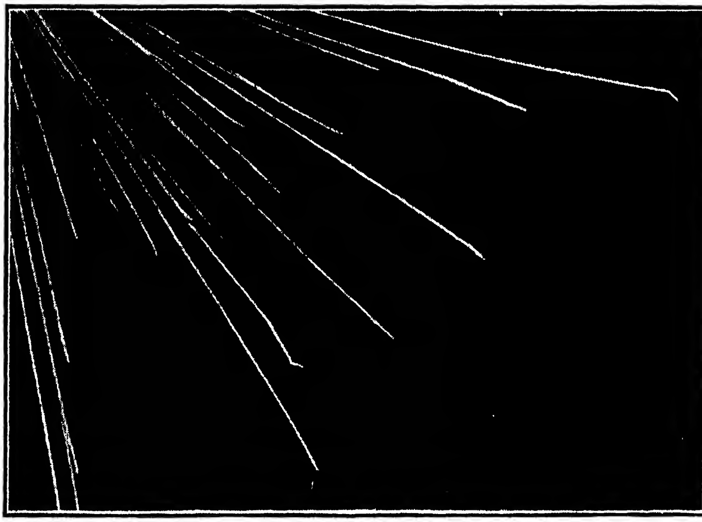


FIG. 512. CLOUD CHAMBER PHOTOGRAPH OF TRACKS OF ALPHA PARTICLES

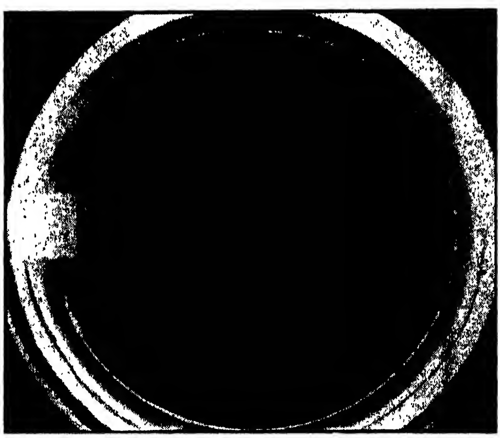


FIG. 513. PHOTOGRAPH SHOWING TRACKS OF THE BETA RAYS OF RADIUM E IN A FIELD OF ABOUT 300 GAUSS
(Photograph provided by F. Rasetti.)

field of intensity B which is parallel to the direction of motion of the piston, the particles which move in a plane perpendicular to B describe circular arcs. The radius R of a track is related to B, e/m, and velocity v of the particle by

$$v = BR\left(\frac{e}{m}\right) \tag{8}$$

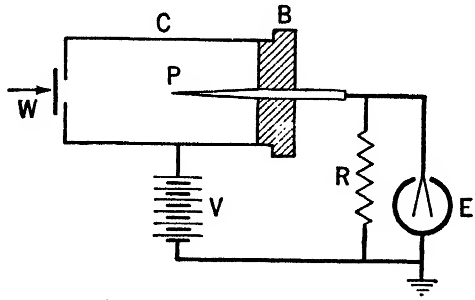


FIG. 514. GEIGER POINT COUNTER

as was pointed out in Chapter 48. If the identity of the ionizing particle is known, by measuring B and R we can find the speed v. It is also important to recall that particles with charges of opposite sign will curve in opposite directions.

Geiger Counters

Although several thousand ions are produced by a single alpha particle or fast electron, the resultant ionization current is not large enough to be detected electrically. For electrical detection some process of amplification is necessary.

The first successful method, which was devised by Geiger in 1908, is shown in Figure 514. One electrode of the ionization chamber is a short cylinder C; the other is the point of a needle P. The ionizing particles enter through the very thin window W. The applied potential V is just below the value that will produce a glow discharge in the chamber. When an ionizing particle enters and produces ions, these ions are accelerated by the electric field between C and P to such an extent that they produce other ions by collision. A momentary current flows through the high resistance R, which causes a visible "kick" of the electrometer E. In this manner the entrance of each individual particle can be observed. In recent years the electrometer E has been replaced by an amplifier of the type used in radio receivers, and the output of this is recorded by an automatic device which registers the total number of pulses.

In 1928, Geiger and Mueller developed the *tube counter*, which is usually referred to as the *Geiger-Mueller counter*. It operates on the same principle, but has a different arrangement of the electrodes as shown in Figure 515. The outer electrode is a metal cylinder as before, but the central electrode is a wire along the axis of the cylinder. In the one shown the electrodes are sealed into a glass bulb which is filled with gas at a reduced

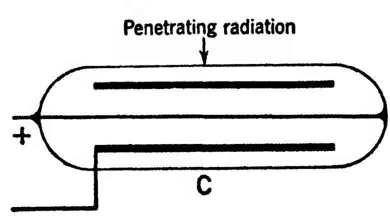


FIG. 515. GEIGER-MUELLER COUNTER

pressure — usually from 2 to 10 centimeters of mercury. This counter is very efficient as a detector of cosmic rays and weak gamma radiation. When the incident radiation ejects a photoelectron from the walls of the cylinder C, a "count" is recorded. As we shall see in the next chapter, most of the investigations of cosmic rays and of artificial radioactivity have been made with counters of this type.

Scattering of Alpha Particles by Heavy Atoms

In 1906, while studying the paths of alpha particles in magnetic fields, Rutherford observed that the particles were slightly deviated or "scattered" by very thin sheets of mica. Since the particles have a relatively large mass and enormous speeds, he realized that the forces responsible for this scattering must be very intense. A few years later we find Geiger, who was working in Rutherford's laboratory, making a detailed study of the scattering of alpha particles in thin films of gold. In reporting this work Geiger remarked that a very small number of the particles, about one in ten thousand, were deviated a surprisingly large amount.

At this point we have another illustration of Rutherford's genius. Instead of dismissing the few observations of large scattering as accidental, he continued to puzzle over a possible explanation. He asked Geiger and Marsden to investigate the large angle scattering in detail. A few days later they reported that some of the particles turned around in the foil and emerged from the same side at which they had entered. Twenty years later, in speaking of his reactions to this information, Rutherford said (88:68):

> It was quite the most incredible event that has ever happened to me in my life. It was almost as incredible as if you fired a 15-inch shell at a piece of tissue paper and it came back and hit you. On consideration I realized that this scattering backwards must be the result of a single collision, and when I made the calculations I saw that it was impossible to get anything of that order of magnitude unless you took a system in which the greater part of the mass of the atom was concentrated in a minute nucleus. It was then that I had the idea of an atom with a minute massive center carrying a charge. I worked out mathematically what laws the scattering should obey, and I found that the number of particles scattered through a given angle should be proportional to the thickness of the scattering foil, the square of the nuclear charge, and inversely proportional to the fourth power of the velocity.

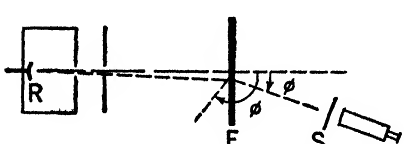


FIG. 516. SCATTERING OF ALPHA PARTICLES BY A THIN FOIL

These deductions were verified by Geiger and Marsden. Their simple and beautiful experiments may be considered the experimental basis for the nuclear model of the atom which is now the foundation of modern physics and chemistry. Their experimental method is shown schematically in Figure 516. A narrow beam of alpha particles from the radioactive source R passes through the foil F and produces scintillations when it strikes the small zinc sulphide screen S. M is a low-power microscope for observing the scintillations. By moving S and M along the arc of a circle, the number of particles deflected through various angles ϕ can be observed. In this manner all of Rutherford's theoretical deductions were verified.

The Nuclear Atom

Rutherford envisaged the atom as consisting of a small, concentrated, positively charged nucleus surrounded by a number of electrons, different for the atoms of different chemical elements, but the same for each atom of a given element. The radius of the nucleus of a gold atom was found to be of the order of 10^{-12} centimeters. All the facts of chemistry and physics indicate that the radius of the whole atom is of the order of 10^{-8} centimeters. Thus the nucleus was only $\frac{1}{10,000}$ the size of the atom itself. In spite of its extreme smallness, the nucleus was assumed to be the seat of practically the entire mass of the atom. The nuclear charge was estimated to be $+Ze$, where Z is an integer equal to about one half the atomic weight of the atom and e is the numerical value of the electronic charge, 1.6×10^{-19} coulombs.

Later experiments have shown that if the elements are arranged in the periodic table and numbered consecutively, beginning with hydrogen as unity, the nuclear charge of the Zth element is $+Ze$. The integer Z is called the ***atomic number*** of the element. Hence, referring to the periodic table in the Appendix, we see that for helium $Z = 2$, for lithium $Z = 3$, for sodium $Z = 11$, etc. Further inspection of the table shows that, while Z is approximately one half the atomic weight for the light elements — as Geiger and Marsden had found — this relationship does not hold for the heavy elements. In the case of mercury, for example, the atomic weight is 200.6, but Z is only 80. Since an atom in its normal state is electrically neutral, it is possible, and often useful, to think of the atomic number as determining the number of electrons moving about the nucleus of the normal atom.

In April, 1912, a young Dane, Niels Bohr, who had just received his doctorate in physics at Copenhagen, arrived at Manchester to work with Rutherford. After starting some experimental work he became so interested in trying to formulate a theory of the nuclear atom that he asked permission to devote all his time to this theoretical problem. Rutherford granted his request and gave him every encouragement.

In order to appreciate the difficulty and magnitude of Bohr's problem, it is necessary to recall some of the requirements of a satisfactory atomic theory. In addition to explaining the scattering of alpha particles, it must account for the emission and absorption of radiation in both the visible and the X-ray region, ionization by various agents, radioactivity, the general features of the periodic arrangement of elements, and the laws of chemical valence.

Bohr's Explanation of the Spectrum of Hydrogen

From the point of view of nineteenth-century physics, Rutherford's nuclear atom model was not satisfactory. For according to Maxwell's electromagnetic equations, the electrons circulating about the central positive charge must radiate energy, therefore lose speed, and in a short interval spiral into the nucleus. His theory also shows that such an atom is unstable and cannot emit the types of spectra that are actually observed.

The concept of a planetary atom was not new; it had been suggested by Perrin in 1901 and Nagaoka in 1904. Owing to the above defects, however, it had never received serious consideration. Bohr was fully aware of these previous failures when he attacked the problem in 1912.

All the experimental facts reviewed in Chapter 34 indicate that the spectrum of a substance is determined by the structure of the emitting atom or molecule. It will be recalled that in 1885 Balmer found that the wavelengths of one group of spectral lines emitted by hydrogen could be represented by the equation

$$\lambda = 3645.6 \times \frac{m^2}{m^2 - 4}, \tag{9}$$

where λ is in angstrom units and m takes the integral values 3, 4, 5, 6, . . . (page 398). Since that time these lines have been known as Balmer's series.

Later, the Swedish physicist Rydberg wrote equation (9) in the following form:

$$\frac{1}{\lambda} = \frac{1}{3645.6}\left(\frac{m^2 - 4}{m^2}\right). \tag{10}$$

Multiplying the numerator and denominator of the right side of (10) by 4, we have

$$\frac{1}{\lambda} = \frac{4}{3645.6}\left(\frac{1}{2^2} - \frac{1}{m^2}\right). \tag{11}$$

If λ is expressed in centimeters this becomes

$$\frac{1}{\lambda} = 109{,}720\left(\frac{1}{2^2} - \frac{1}{m^2}\right). \tag{12}$$

The factor 109,720 is known as Rydberg's constant and is usually designated by the letter R. Later more precise determinations of R gave the value 109,677.8 cms^{-1}, or 1.096778×10^7 reciprocal meters.

Lyman, in 1906, and Paschen, in 1908, discovered two other series in the spectrum of hydrogen which can be represented respectively by

$$\frac{1}{\lambda} = R\left(\frac{1}{1^2} - \frac{1}{m^2}\right) \tag{13}$$

where $m = 2, 3, 4 \ldots$, and

$$\frac{1}{\lambda} = R\left(\frac{1}{3^2} - \frac{1}{m^2}\right) \tag{14}$$

with m taking integral values beginning with 4.

Thus all the spectral lines emitted by hydrogen can be represented by an equation of the type

$$\frac{1}{\lambda} = R\left(\frac{1}{n^2} - \frac{1}{m^2}\right) \tag{15}$$

where n and m are integers. If (15) is written in the form

$$\frac{1}{\lambda} = \frac{R}{n^2} - \frac{R}{m^2}, \tag{16}$$

the right side is the difference of two quantities called spectral terms. Bohr's great contribution to modern physics consisted in devising an atomic theory which gave a simple physical interpretation of these terms.

Bohr's Theory of the Hydrogen Atom

Bohr's first step was to propose a theory of the hydrogen atom that would account for Balmer's series. Accepting Rutherford's idea of a small massive nucleus he assumed that the hydrogen atom consisted of a tiny nucleus with a charge $+e$, called a proton, about which a single electron was rotating. The letter e represented the magnitude of the electronic charge.

From Coulomb's law of electrostatics it was known that the force of attraction between the proton and electron varied inversely as the square of the distance between them. In general therefore, the electron would describe an elliptical orbit about the proton just as the planets move in ellipses around the sun.

In order to simplify the calculations Bohr first considered the special case where the electron moved in a circular orbit. Since the proton mass is 1840 times the mass of the electron it is assumed that the proton remains at rest, only the electron being in motion. The electrostatic attraction between the proton and electron is $\dfrac{e^2}{4\,\pi k_0 r^2}$. This force directed toward the center of the orbit gives the electron a centripetal acceleration $\omega^2 r$. Applying Newton's second law we have

$$m\omega^2 r = \frac{e^2}{4\,\pi k_0 r^2}, \tag{17}$$

where m is the mass of the electron, ω its angular velocity, and k_0 is the dielectric constant of empty space.

According to Maxwell's electromagnetic theory this circulating electron *must* radiate energy and describe a spiral path until it collides with the proton. The observed spectrum of hydrogen shows that this gradual collapse of atoms does not occur. Bohr therefore assumed that Maxwell's theory does not apply to atomic processes, although we know that it is valid for large scale phenomena such as radio.

With the insight of genius Bohr made two assumptions or postulates which enabled him to solve the problem. The first postulate specified that only those electronic orbits were possible in which the angular momentum of the electron was $nh/2\pi$, where n was an integer, that is, 1, 2, 3, 4, ... The permissible orbits, which were the ones that occurred in nature according to the theory, were called stationary motions or stationary states. In any one of these states the energy of the atom was assumed to be constant; no radiation was emitted. The quantum number of the state or orbit was called n. The second postulate stated that when an electron made a transition or jump from a state having a quantum number n_2 and total energy W_{n2} to a lower state where $n = n_1$ and the energy was W_{n1}, the energy loss $W_{n2} - W_{n1}$ was given off as radiation. It was further assumed that the frequency ν of this radiation was given by

$$\nu = \frac{W_{n2} - W_{n1}}{h}, \tag{18}$$

where h was Planck's constant.

Applying these postulates to the hydrogen model, Figure 517, Bohr derived the correct equation for all the spectral series of hydrogen. Even more striking was the fact that the theory gave the numerical value of R. To derive Rydberg's formula we start with (17) which can be written

$$m\omega^2 r^3 = \frac{e^2}{4\pi k_0}. \tag{19}$$

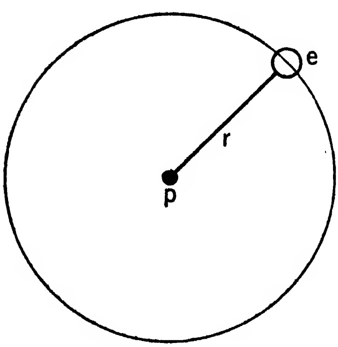


FIG. 517. MODEL OF HYDROGEN ATOM

The first postulate requires that

$$m\omega r^2 = \frac{nh}{2\pi} \tag{20}$$

since the angular momentum of the electron is $m\omega r^2$. (In Chapter 15 angular momentum is defined as $I\omega$, and in this case $I = mr^2$.) Squaring both sides of (20) and dividing by (19) we have

$$mr = \frac{(4\pi k_0)n^2h^2}{4\pi^2e^2}$$

or

$$r = \frac{(4\pi k_0)n^2h^2}{4\pi^2e^2m}, \tag{21}$$

where n can have only integral values. Hence, the radii of the possible stationary states are in the ratio 1, 4, 9, ... In order to find the frequency of the emitted light it is necessary to calculate the energy of the atom, W_n, when the electron is in the nth orbit. This total energy is equal to the sum of the kinetic and potential energies. From electrostatics we have for the potential energy U of the electron

$$U = -\frac{e^2}{4\pi k_0 r_n}, \tag{22}$$

where r_n is the radius of the nth orbit. The negative sign comes from the negative charge of the electron. From (19) the kinetic energy is found to be

$$\tfrac{1}{2} mv^2 = \tfrac{1}{2} m\omega^2 r^2_n = \frac{e^2}{(4\pi k_0)2\, r_n}. \tag{23}$$

The total energy can now be written

$$W_n = U + \tfrac{1}{2} mv^2 = -\frac{e}{(4\pi k_0)2\, r_n}. \tag{24}$$

Substituting for r_n, from equation (21),

$$W_n = \frac{-2\pi^2 e^4 m}{(4\pi k_0)^2 h^2 n^2}. \tag{25}$$

The values of W_n are the energy states or energy levels of the atom. When the electron falls from a higher level n_2 to a level n_1 the frequency of the emitted radiation is given by (18).

$$\nu = \frac{W_{n2} - W_{n1}}{h} = \frac{2\pi^2 e^4 m}{(4\pi k_0)^2 h^3}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right), \tag{26}$$

or the reciprocal wave-length is

$$\frac{1}{\lambda} = \frac{\nu}{c} = \frac{2\pi^2 e^4 m}{(4\pi k_0)^2 c h^3}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right). \tag{27}$$

If we make

$$R = \frac{2\pi^2 e^4 m}{(4\pi k_0)^2 c h^3} \qquad (28)\ [1]$$

equation (27) is identical with Rydberg's expression for Balmer's series. Substituting the numerical values of e, m, k_0, c, and h in (28) we find for the value of R, 1.09×10^7 reciprocal meters which agrees within the limits of experimental error with the value obtained from spectra. This agree-

[1] In the c.g.s. system of units $4\pi k_0 = 1$, so that the expression for Rydberg's constant becomes

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

which is the equation given in most textbooks.

ment is quite remarkable since the measurements of the five constants e, m, k_0, c, and h are entirely independent of the spectrum of hydrogen. Another success of the theory is that the value of the radius of the first orbit, $n = 1$, calculated by equation (25) is $.53 \times 10^{-10}$ meters, which agrees with the results obtained from the kinetic theory of gases.

In the normal state of the atom the electron is in the state of minimum energy, that is, $n = 1$. As long as it remains in this state no radiation is emitted. When a hydrogen atom is bombarded by electrons or ions the electron will be raised to an orbit further removed from the center. It remains here a very short interval of time and then falls to a lower orbit. As indicated in Figure 518 the transitions which end on the level $n = 2$ give rise to the lines of the Balmer series. Those ending on the level $n = 1$ produce the Lyman series, and for $n = 3$ the so-called Paschen series. For each transition the wavelength of the emitted radiation is given by equation (27).

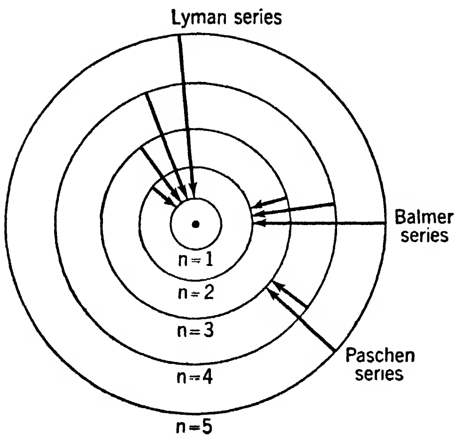


FIG. 518. CIRCULAR ORBITS IN THE HYDROGEN ATOM
(Not drawn to scale.)

The cardinal feature of Bohr's contribution to atomic physics is his concept of stationary states or energy levels. In its normal state the atom is in the lowest level. No radiation can be emitted until its energy is increased to a higher level. This increase may be brought about by collisions due to high temperature, by electronic impacts, or by absorption of radiation. The amounts of energy in electron-volts required to raise the atom to the higher levels are called the *excitation potentials* of the atom. The energy necessary to remove one electron is called the first ionization potential. According to the theory the ionization potential is numerically equal to the energy of the atom in the normal state. Similarly, differences between various excitation potentials should be equal to the differences in energy levels as found from spectra.

By bombarding gases and vapors with electrons of different energies Franck and Hertz in 1913 were able to verify all these predictions of the theory. They showed further that if an atom was struck by an electron which did not have sufficient energy to raise the atom to its first excited state, the collision was elastic. Practically no energy was transferred to the atom and no radiation emitted. But in case the electronic energy was great enough to excite the atom, a large fraction of the collisions would result in excitation and emission of light. Since atoms can exist only in a discrete number of energy states they can accept energy only in finite chunks or *quanta*. The Bohr theory, then, can be considered an extension of Planck's quantum idea to all atomic processes.

Application of Bohr's Theory to More Complex Atoms

It will be recalled that the atomic number, Z, fixes the number of electrons in a given atom, which, in turn, determines its chemical behavior. Thus any atom with $Z = 10$ is neon irrespective of its atomic mass. As we shall see in the next chapter, atoms having the same value of Z but different masses are known as isotopes.

As an example of the extension of Bohr's theory to an atom with several electrons, we shall consider an atom of sodium. The atomic number of sodium is 11; hence, there are eleven electrons moving about the nucleus. The electrons are attracted by the nucleus and repelled by one another. From the optical and X-ray spectra of sodium and its chemical behavior, Bohr and others have proposed an arrangement of the eleven electrons which seems to meet the requirements imposed by our knowledge of the properties of the sodium atom. The general features of the arrangement are shown schematically in Figure 519. Two electrons remain relatively close to the nucleus; they are in quantum states having $n = 1$ and are said to be in the K shell. Eight electrons are in states having $n = 2$, which constitute the L shell; states having $n = 3$ constitute the M shell; and so on. It should be pointed out that the eight electrons of the L shell — and the electrons of any single shell except the K — do not all have the same energy and are not all in identical orbits. The word *shell*, therefore, is not synonymous with energy level. The eight electrons which have $n = 2$ form what is called a closed shell. Hence, the eleventh electron must remain in the M shell. This single outermost electron determines the optical and chemical properties of sodium. When it is raised to a higher, normally unoccupied, energy level, it falls back, and the atom emits radiation in accordance with Bohr's second postulate. Conversely, if light of all wave-lengths is passed through sodium vapor, those wave-lengths will be absorbed which are emitted by electron jumps ending on the lowest or normal energy level. This group of lines is known as the principal series of sodium. The other series observed in emission do not occur in the absorption spectrum.

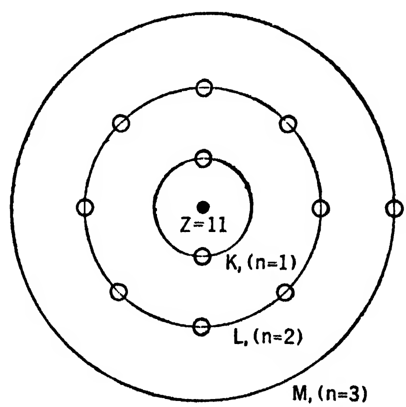


FIG. 519. ARRANGEMENT OF ELECTRONS IN THE SODIUM ATOM
(Not drawn to scale.)

X-Ray Spectra; Moseley's Law

The optical spectrum of an element is a complex arrangement of many lines. Its interpretation is often difficult. X-ray spectra, on the other hand, are exceedingly simple, consisting of only a few lines which can be easily classified and interpreted by Bohr's theory.

The remarkable simplicity of X-ray spectra was first observed and interpreted in 1913 by a young graduate student, H. G. J. Moseley. Working in Rutherford's laboratory in Manchester he determined the wave-lengths of the characteristic X-rays emitted by successive elements in the periodic table. Moseley's experimental method was essentially the same as Bragg's (page 611). The X-rays of the element which was being investigated were excited by electron bombardment. They were reflected from a rotating crystal and their wave-lengths computed by Bragg's equation (page 612).

Moseley found that each element gave two groups of characteristic X-ray lines. Adopting the nomenclature of Barkla he called the group having the shorter wave-lengths, K lines, the other L lines. The two strongest lines in the K group were designated as K_α and K_β. In his first paper[1] wave-lengths were given for the K_α and the K_β lines of the elements from calcium to zinc, whose atomic numbers extend from 20 to 30. Although the optical spectra and chemical properties of these elements vary greatly, their X-ray spectra are all similar. The only change with increasing atomic number is a shift of the lines toward the shorter wave-lengths. Moseley found that the frequencies of the K_α lines for these elements were given accurately by the equation

$$\nu_{K_\alpha} = b\,(Z-1)^2 \tag{29}$$

where b is a constant and Z is the atomic number. Thus the frequency increases regularly with the nuclear charge Z. There is no periodicity in X-ray spectra corresponding to the chemical periodic table.

As Moseley himself pointed out, equation (29) can be readily interpreted in terms of Bohr's theory. Assuming that the K_α line is emitted when an electron jumps from a state $n = 2$ to one where $n = 1$, the theory gives approximately

$$\nu_{K_\alpha} = cR\,(Z-1)^2\left(\frac{1}{1^2} - \frac{1}{2^2}\right), \tag{30}$$

which is of the same form as equation (29).

X-Ray Spectra and Electron Shells

A few years after Moseley's pioneer experiments Kossel gave a detailed explanation of X-ray spectra in terms of the Bohr atom model. This explanation can be best presented by considering a special case, say copper.

[1] *Philosophical Magazine* (6), *26*, 1024 (1913).

Since the atomic number of copper is 29, the atom has 29 electrons. Chemical and spectroscopic evidence indicates that two electrons are in the *K* shell, eight in the *L* shell, eighteen in the *M* shell, and one in the *N* shell. To excite the K_α lines of an atom it is necessary to remove one electron from the *K* shell. This requires bombardment with fast electrons since the *K* electrons are well shielded by other electrons and are strongly attracted by the nuclear charge of +29 *e*. In the case of copper, energy of approximately 9000 electron-volts is required to remove a *K* electron; for tungsten ($Z = 74$) 69,000 electron-volts is necessary. Thus an X-ray tube with a copper target must be operated at a potential of 9000 volts or more in order to excite its characteristic K_α X-rays.

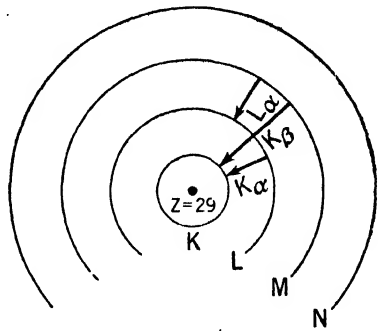


FIG. 520. SOME POSSIBLE ELECTRON TRANSITIONS IN PRODUCING X-RAYS OF COPPER

When one of the *K* shell electrons has been removed the vacancy is filled by a transition from the *L* or *M* shell. As shown in Figure 520, a jump from the *L* shell gives rise to the K_α X-ray line, while one from the *M* shell produces K_β. The L_α line originates in a transition from *M* to *L*. If Kossel's interpretation of these processes is correct, we must have

$$h\nu_{K_\alpha} + h\nu_{L_\alpha} = h\nu_{K\beta}. \tag{31}$$

For according to Bohr's theory, both sides of equation (31) are equal to the difference in energy of an electron in the *M* shell and one in the *K* shell. Taking account of the slight differences in energy of the different electrons in a single shell, equation (31) is found to be true.

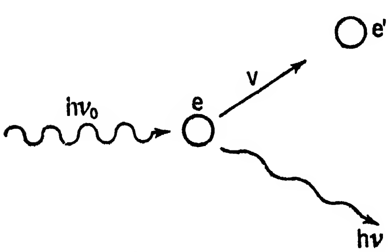


FIG. 521. COMPTON'S THEORY OF A COLLISION BETWEEN A PHOTON AND AN ELECTRON

Although an element emits an X-ray spectrum consisting of sharp lines it does not have an absorption spectrum of the same type. The reason for this can be understood with the aid of Figure 520. Assume that copper *Kα* X-rays pass through a sheet of copper. The energy *hν* of the rays is just sufficient to raise a *K* electron to the *L* shell. As was pointed out above, however, the *L* shell is filled so that the transfer cannot take place. Radiation having a value of *hν* sufficient to remove an electron from the atom will be strongly absorbed, but there is no sharp absorption line like that obtained in the optical case. Thus the Bohr theory, which was designed to interpret optical spectra, proved to be just as successful in interpreting the emission and absorption of X-rays.

Photon Hypothesis and the Compton Effect

In his interpretation of the photoelectric effect (page 625), Einstein assumed that light consisted of quasi-corpuscles, called ***photons.*** The energy of an individual photon was taken to be $h\nu$. At about the same time Einstein was led by his special theory of relativity to propose a fundamental relationship between mass and energy, namely,

$$\text{mass} = \frac{\text{energy}}{c^2} \tag{32}$$

where c = velocity of light. According to this viewpoint a photon of radiation can be considered as having a mass of $\frac{h\nu}{c^2}$ and momentum $\frac{h\nu}{c^2}\cdot c = \frac{h\nu}{c}$. Owing to the large value of ν in the X-ray region an X-ray photon may have a mass comparable to that of an electron at rest.

When a beam of light passes through a fog some of it is diffused or scattered in all directions. It is by means of this scattered light that we are able to see the beam. The diffused light has the same wave-length as the original beam, and the mechanism of the scattering can be explained by Maxwell's electromagnetic theory.

In 1922, the American physicist A. H. Compton made the surprising discovery that when X-rays are scattered by a light substance such as carbon, the wave-length of the radiation is slightly increased. This change of wave-length is contrary to all predictions of the wave theory of light. The photon hypothesis, however, gives a beautifully simple explanation of the phenomenon.

In Compton's theory of the process the collision between a photon and an electron is treated in the same manner as the collision between two perfectly elastic particles. As was shown in Chapter 13, such a collision is characterized by the conservation of momentum and the conservation of energy. Let us consider a photon $h\nu_0$, Figure 521, which collides with an electron e, originally at rest. After the impact the electron will have a certain velocity v in the direction ee'. Since the electron has acquired an amount of energy, $\frac{1}{2}mv^2$, the energy of the scattered photon is

$$h\nu = h\nu_0 - \tfrac{1}{2}mv^2. \tag{33}$$

The frequency ν being less than ν_0, the wave-length of the scattered X-rays must be greater than that of the incident beam. Applying equation (33) and the condition for conservation of momentum, Compton was able to calculate the exact change in wave-length for a given angle of scattering. His experimental values were in excellent agreement with the theoretical predictions.

The Compton effect is often presented as conclusive proof of the corpuscular nature of radiation. A more careful analysis, however, shows that although the photon is quasi-corpuscular in interacting with matter it cannot be considered a corpuscle in the sense in which Newton used the term. There is no experimental evidence that photons collide with one another. In order to determine the mass and energy of the photon its wave-length must be measured by a process in which the photon is considered to be a wave motion and nothing more. Further, it travels with only one velocity, the velocity of light *in vacuo*, whereas the velocity of a particle varies with its energy. All the experiments on interference and diffraction of Chapter 33 can be simply and accurately interpreted in terms of a pure wave theory. Thus the experimental facts require us to adopt the view-point that under certain experimental conditions light can be treated as a purely wave phenomenon; under other conditions it must be considered quasi-corpuscular.

Matter-Waves of De Broglie

In attempting to reconcile the wave and corpuscular features of light Louis de Broglie was led to suggest in 1924 that particles of matter should have wave properties. The general ideas which guided him in developing his theory have been clearly stated by De Broglie himself in his Nobel Prize address, given in 1929 (35:168):

> When I began to consider these difficulties I was chiefly struck by two facts. On the one hand the Quantum Theory of Light cannot be considered satisfactory, since it defines the energy of a light corpuscle by the equation $W = h\nu$, containing the frequency ν. Now a purely corpuscular theory contains nothing that enables us to define a frequency; for this reason alone, therefore, we are compelled in the case of Light, to introduce the idea of a corpuscle and that of periodicity simultaneously.
>
> On the other hand, determination of the stable motion of electrons in the atom introduces integers; and up to this point the only phenomena involving integers in Physics were those of interference and normal modes of vibration. This fact suggested to me the idea that electrons too could not be regarded simply as corpuscles, but that periodicity must be assigned to them also.

By applying Einstein's special theory of relativity De Broglie derived the following relation between the motion of a particle and its associated wave:

$$\lambda = \frac{h}{mv} \tag{34}$$

where λ is the wave-length of the wave, mv is the momentum of the particle, and h is Planck's constant. It is interesting to note that equation (34) holds for photons. For electromagnetic waves we have

$$\lambda = \frac{c}{\nu} = \frac{hc}{h\nu} = \frac{h}{\left(\frac{h\nu}{c}\right)}. \tag{35}$$

As was shown on page 642, $h\nu/c$ is the momentum of the photon. Thus the relationship $\lambda = \frac{h}{\text{momentum}}$ seems to be a very general one which holds for both matter and radiation.

Experiments of Davisson and Germer

Equation (34) shows that the wave-lengths of the De Broglie waves associated with the smallest particle that can be observed and weighed are very much shorter than that of any known radiation. They are so short that there seems to be no possibility of detecting them experimentally. In the case of electrons, however, the situation is different. If we consider an electron which has a kinetic energy of 100 electron-volts, its momentum mv is $9.1 \times 10^{-31} \times 5.93 \times 10^{6} = 5.40 \times 10^{-24}$ kgm $\frac{m^4}{sec}$.

Hence,
$$\lambda = \frac{6.62 \times 10^{-34}}{5.40 \times 10^{-24}} = 1.22 \times 10^{-10}\text{m}$$
$$= 1.22 \text{ angstroms,}$$

which is in the region of X-ray wave-lengths.

In 1927, the American physicists Davisson and Germer showed that beams of electrons are diffracted by a crystal in precisely the same manner that X-rays are. Applying Bragg's equation (2), page 612, they were able to determine λ for electrons of known energy. The experimental values agreed with those calculated by De Broglie's equation. In this experiment we have the surprising phenomenon of a beam of electrons showing interference effects similar to those exhibited by X-rays.

Interpretation of Wave-Particle Duality

One must not conclude from Davisson and Germer's results that "electron waves" and X-rays are identical in nature. As J. J. Thomson showed in 1895–97, fast electrons have properties which are entirely absent in light waves. In most situations electrons can be considered small particles, each having a definite mass and charge. Their behavior can be predicted by the equations of Newton and Maxwell. But when they interact with atoms of a gas or crystal it is necessary to use the wave-equations of De Broglie and Schroedinger. These equations predict interference effects which are actually observed. We cannot, however, give a "physical picture" of these waves in the sense that we can of sound waves, for example. At the present time we must simply regard them as mathematical symbols which are used in predicting the outcome of certain types of experiments.

Although this wave-particle duality is found in radiation phenomena and in experiments involving only matter, the two cases are far from identical. Photons differ fundamentally from electrons. They have zero rest mass, no electric charge, and always travel with the same speed. Their mass and momentum seems to be concentrated, yet interference effects are obtained with only a single photon passing through the apparatus at a time. The photon interferes with itself, so to speak. A single photon must therefore cover a whole grating or lens. All these apparently contradictory results show that our simple intuitive concepts based upon ordinary mechanics are not able to deal with the interactions of radiation and matter. We cannot visualize processes on the atomic scale.

Questions for Self-Examination

1. What was the central point in Planck's revolutionary hypothesis on the nature of radiant energy?
2. What is the "photoelectric effect" and what was Einstein's hypothesis as to its nature?
3. What new type of phenomenon did Wilson's cloud chamber disclose?
4. What was the significance of the scattering of alpha particles first observed by Rutherford?
5. Tell how Bohr combined the quantum hypothesis with Rutherford's picture of the atom to account for the spectrum of hydrogen.
6. Compare a typical X-ray spectrum with the spectrum of hydrogen and tell why the similarity exists.
7. What is the evidence that light waves possess some of the properties of flying particles?
8. What is the evidence that electrons possess some of the properties of waves?

Problems on Chapter 50

1. If the photoelectric threshold of a metal is 3000 angstroms, what energy in electron-volts is required to remove an electron from the metal? 4 electron-volts.
2. Using equation (21) calculate the radius of the first Bohr orbit for an atom of hydrogen. $.5 \times 10^{-10}$ m.
3. Calculate the wave-lengths of the first two lines of the Balmer's series using equation (27). Compare your results with the experimental values on page 399.
4. What is the shortest wave-length in angstroms emitted by an X-ray tube that is operating with a potential difference of 30,000 volts between filament and target? .4 angstrom.
5. What is the wave-length of the De Broglie waves of an electron whose kinetic energy is 10,000 electron-volts? 1.2×10^{-11} meters.

CHAPTER 51

Nuclear Physics

The Nucleus of the Atom

After the remarkable successes of the Bohr theory in interpreting optical and X-ray spectra there was little doubt concerning the validity of its principal features. The existence of energy levels and the arrangement of electrons in shells were firmly established.

It must be kept in mind, however, that the tiny positively charged nucleus is the real controlling element. Its charge determines the number of electrons and therefore the chemical and optical properties of the atom.

All the experiments on the scattering of alpha particles and Moseley's law for X-ray wave-lengths are consistent with the view that the nuclear charge of any atom is given by $+ Ze$, where Z is the atomic number of the element and e is the magnitude of the electronic charge. Thus, referring to the table of elements in the Appendix, for hydrogen $Z = 1$, helium $Z = 2$, lithium $Z = 3$, neon $Z = 10$, and so on.

Any atom with a nuclear charge $+ 10\,e$ will have ten extra nuclear electrons which give it the spectrum and chemical behavior of neon. Removal of one of these outer electrons does not create a new type of atom. The resulting positive ion captures an electron and reverts to a normal atom of neon. To change an atom into another chemical type it is necessary to change the nuclear charge Ze.

Chemical Elements and Types of Atoms

The periodic table shows ninety-two chemical elements with values of Z from 1 to 92 and atomic weights from 1.008 for hydrogen to 238.14 for uranium. Chemical atomic weights are relative. All oxygen atoms are assumed to have a weight of 16.00 atomic weight units, and the other atomic weights are relative to this value. For example, the atomic weight of sodium is given as 23.00; this means that the ratio

$$\frac{\text{weight of } N \text{ sodium atoms}}{\text{weight of } N \text{ oxygen atoms}} = \frac{23}{16}$$

where N is any integer.

Forty years ago the two questions: how many chemical elements are there? and how many types of atoms are there? would have been considered synonymous. It was assumed that each element consisted of a single type of atom. This assumption of ninety-two distinct types of atoms is not very satisfactory to those who believe that all natural phenomena can be interpreted in terms of a *few* fundamental concepts. In 1813, ten years after Dalton had proposed his atomic theory, Dr. Prout, an Edinburgh physician, observed that the atomic weights of all the elements were integers, taking hydrogen as *one*. This suggested that all atoms are composed of a number of primordial atoms, and therefore all atomic weights should be integers.

Later accurate determination of atomic weights showed that Prout's hypothesis was not tenable. For, as the periodic table shows, the atomic weights of the great majority of the elements are not integers.

After the discovery of radioactivity, elements were found which could not be separated by chemical means but which had distinct radioactive properties. This led Soddy to state in 1910 (88:100–101):

> Chemical homogeneity is no longer a guarantee that any supposed element is not a mixture of several [elements] of different atomic weights, or that any atomic weight is not merely a mean number.

Soddy's suggestion was considered quite radical at the time. It proposed that an element such as neon, atomic weight 20.18, for example, consisted of a mixture of atoms having different weights, say 20, 21, and 22. Although these three types of atoms have different weights it is not possible to separate them by chemical means. Since they always occur in nature in the same proportions, the atomic weight of the mixture is always found to be 20.18. In accordance with Soddy's suggestion atoms having different masses but identical chemical properties are called ***isotopes*** (meaning ***same place*** in the periodic table).

Precise Measurements of Atomic Masses

As was pointed out in connection with J. J. Thomson's experiment (page 596), masses of particles even smaller than atoms can be determined if the particle has an electric charge. Although atoms normally are not charged, they acquire a charge when ionized in a discharge tube.

In 1886, in the course of his work on cathode rays, Goldstein observed rays behind a perforated cathode. Several years later Wien and J. J. Thomson showed that these so-called positive rays were atoms which had lost one or more electrons. By bringing these ions into a magnetic field it is possible to determine their relative masses with high precision.

The first precise results by this method were obtained by F. W. Aston in the Cavendish laboratory. His apparatus, called a ***mass spectrograph***, is rather complicated in design, so we shall describe a more recent instrument designed by K. T. Bainbridge.

The essential parts of the Bainbridge apparatus are shown in Figure 522. The ions to be "weighed" are formed in a discharge tube above the slit S_1.

After passing through the slit S_2 and between the parallel plates P_1 and P_2, they emerge through another slit S_3 into a region in which there is a uniform magnetic field B perpendicular to the plane of the diagram. In this field the ions describe a circular path of radius R given by

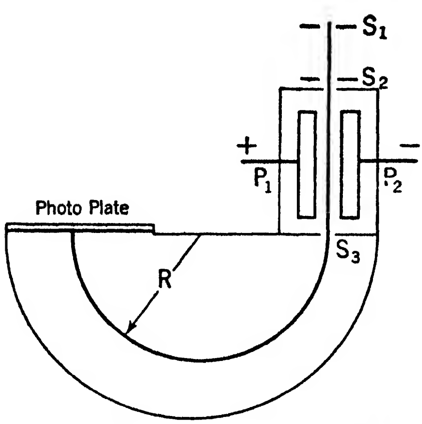


FIG. 522. DIAGRAM OF BAINBRIDGE'S MASS SPECTROGRAPH

$$R = \frac{mv}{eB} \tag{1}$$

where m = mass of ion, e = charge, and v = velocity. (Equation (10), page 599.)

After traversing the semicircle the ions strike the photographic plate, producing a trace or line. From the positions of these lines the values of R can be found. If all ions have the same velocity and the same charge, equation (1) shows that the ionic masses are proportional to R. If there is an electrostatic field between the plates P_1 and P_2, and a magnetic field perpendicular to the electric one, only ions with a certain specific velocity can pass through the slit S_3. This insures the constancy of v in equation (1).

The results of Aston, Bainbridge, and others have shown that Prout's hypothesis is essentially correct. All elements having fractional atomic weights are found to consist of a mixture of several isotopes. For example, lithium, atomic weight 6.94, consists of a mixture of two types of atoms having atomic masses of 6 and 7. Assuming 7.9 per cent of the atoms have the smaller mass, the average mass of the mixture is 6.94.

In light of our present information, therefore, the two questions proposed above are not identical. The number of chemical elements is 92, but 287 different atoms having distinct masses have been observed.[1]

Notation for Atomic Nuclei

In order to identify the various isotopic nuclei a specific notation has been adopted. The chemist uses an abbreviation or symbol for each element. Thus H represents hydrogen, Ne neon, Li lithium, and so forth. The complete list is given in the periodic table. To represent the nuclei the atomic number Z is written as a subscript at the left of the chemical symbol and the mass number as a superscript at the right. For example, the three types of neon nuclei are written ${}_{10}Ne^{20}$, ${}_{10}Ne^{21}$, and ${}_{10}Ne^{22}$. Since the atomic mass is very nearly the same as the nuclear mass, the same notation is used for atoms.

[1] *Reviews of Modern Physics*, **12**, 31 (1940).

Relation Between Mass and Energy

As the precision of mass spectrographs was increased it became evident that the whole-number rule for atomic masses was only approximate. The more accurate masses of the first light elements are hydrogen, 1.00813; the hydrogen isotope known as heavy hydrogen or deuterium, 2.01473; helium, 4.00389; $(\text{lithium})^6$, 6.0168; $(\text{lithium})^7$, 7.01818. As we shall see later, another fundamental particle is the ***neutron***, having no charge and a mass of 1.00893 units.

At first sight the above values seem to contradict Prout's hypothesis. If we assume, for example, that the helium nucleus is composed of two neutrons and two ***protons*** (hydrogen nuclei), the resultant structure has the correct charge $+2\,e$. The total mass of the four particles, however, is 4.03422 while the observed mass of helium is 4.00389. An appreciable amount of mass has disappeared.

This difficulty can be eliminated by accepting the relation between mass and energy which was proposed by Einstein in 1905. In his special theory of relativity, Einstein put forward the hypothesis that the inertial mass m of any particle of matter varies with the velocity v of the particle according to the equation

$$m = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{2}$$

where c is the velocity of light, and m_0 is a constant known as the ***rest mass*** of the particle. In ordinary mechanics the ratio v/c is so small that the mass can be considered constant and equal to m_0. In the case of high-speed electrons, however, v is of the same order of magnitude as c and the variation of mass with speed must be taken into account. All experimental observations in this field are in harmony with equation (2).

If we retain the Newtonian definition of force as the rate of change of momentum and the usual definition of work, it can be shown that the kinetic energy of a particle with rest mass m_0 and velocity v is given by

$$\text{K.E.} = m_0c^2\left(\frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} - 1\right). \tag{3}$$

From equations (2) and (3) we have

$$\text{K.E.} = (m - m_0)c^2 \tag{4a}$$

or

$$m = m_0 + \frac{\text{K.E.}}{c^2}. \tag{4b}$$

Equation (4b) states that the increase in mass of a moving particle is equal to its kinetic energy divided by the square of the velocity of light.

But there seems to be no good reason for assuming that this relationship is limited to *kinetic* energy. Consequently, Einstein and others adopted the hypothesis that the relation is a general one — that with an amount of energy, W, in any form, kinetic, potential, thermal, etc., there is always associated a mass, m, equal to W/c^2.

According to this point of view the mass of a molecule must be less than the mass of its component atoms. Consider, for example, a diatomic molecule. Let m be the mass of the molecule, m_{01} and m_{02} the rest masses of the *atoms*, and D the work done against attractive valence forces in separating the atoms. Since the work D must be added to the molecule in order to produce two atoms, equation (4b) gives

$$m + \frac{D}{c^2} = m_{01} + m_{02} \qquad (5a)$$

or

$$m = m_{01} + m_{02} - \frac{D}{c^2}. \qquad (5b)$$

In ordinary chemical reactions D is so small that the change in mass is imperceptible. In nuclear reactions, however, D, which is called *binding energy*, is large enough for the change in mass to be observed. All the data available at the present time are in harmony with the relation (5b) and the experimental values of losses in mass are used to calculate nuclear binding energies.

Radioactivity as a Nuclear Process

The nuclear atom model gives a direct and satisfactory explanation of the principal facts of radioactive disintegration. A radioactive element is one whose nucleus is unstable. It eventually explodes ejecting fast alpha or beta particles in the process. In symbolical notation the transformations from radium to radium C[1] can be written

$$_{88}\text{Ra}^{226} = {}_{86}\text{Rn}^{222} + {}_{2}\text{He}^{4} \qquad (6)$$

$$_{86}\text{Rn}^{222} = {}_{84}\text{Ra}A^{218} + {}_{2}\text{He}^{4} \qquad (7)$$

$$_{84}\text{Ra}A^{218} = {}_{82}\text{Ra}B^{214} + {}_{2}\text{He}^{4} \qquad (8)$$

$$_{82}\text{Ra}B^{214} = {}_{83}\text{Ra}C^{214} + {}_{-1}e^{0} \qquad (9)$$

where $_{-1}e^0$ represents an electron or beta particle. The mass of the electron is not actually zero, but it is very small in atomic weight units. Equation (6) states that when a radium nucleus having a mass of 226 units and charge of $+88\ e$ explodes, it ejects an alpha particle, $_2\text{He}^4$, having a mass 4 and charge $+2\ e$, leaving a residual nucleus of mass 222 and charge $+86\ e$. Conservation of mass is insured by balancing the sum of superscripts on both sides of the equation; conservation of charge by balancing

[1] Compare these equations with the diagrammatic representation of the successive changes involved in radioactive disintegration in Figure 506.

the subscripts. As equation (9) shows, when an electron is ejected from a nucleus a new element is formed with a higher nuclear charge.

Since some of the observed beta rays come from the nucleus it was originally assumed that nuclei were composed of alpha particles, protons, and electrons. More recent experimental and theoretical work, however, favors the view that the only constituents of atomic nuclei are protons and neutrons. It is now assumed that the nuclear beta particles are created when the radioactive change occurs. Many of the beta particles which are observed come from the electron shells. These seem to be emitted by a kind of "internal photoelectric effect."

The above equations of radioactive change give no account of the origin of gamma rays. After the remarkable success of the Bohr theory in accounting for X-ray spectra it was a natural step to extend the concept of energy levels to the nucleus. Thus, in certain radioactive transformations, the newly formed nucleus is assumed to be in an excited state. As it falls to the normal state its excess energy is given off as gamma radiation. This accounts for the fact that gamma rays are always associated with the emission of alpha or beta particles.

The First Transmutation of a Stable Element

During 1914–18, Rutherford's research efforts turned from atomic nuclei to submarine detection. During his spare moments in 1917–18, however, he made observations on the passage of alpha particles through gases. He devised an experimental approach to the problem which is illustrated in Figure 523. The radioactive source R was placed in a tube T which could be filled with different gases. At one end was a thin window and a zinc sulphide screen, S. The impacts of alpha particles on this screen produced scintillations not unlike the sparks from a blacksmith's anvil. The scintillations were observed with the microscope M.

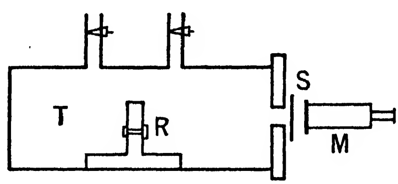


FIG. 523. RUTHERFORD'S APPARATUS FOR THE TRANSMUTATION OF NITROGEN

The alpha particles themselves could travel 7 centimeters in air. But a few scintillations were observed when the distance RS was as large as 40 centimeters. Rutherford suspected that these particles striking the screen were hydrogen nuclei. By deflecting them in a magnetic field he was able to show that the scintillations were produced by protons. The most surprising fact, however, was that the greatest number of long-range particles were observed when the tube was filled with pure nitrogen. In reporting his results Rutherford said:[1]

> It is difficult to avoid the conclusion that the long-range atoms arising

[1] *Philosophical Magazine*, *37*, 581 (1919).

> from the collision of alpha particles with nitrogen are not nitrogen but probably atoms of hydrogen or atoms of mass 2. If this be the case we must conclude that the nitrogen atom is disintegrated under the intense forces developed in a close collision with a swift alpha particle and that the hydrogen atom which is liberated formed a constituent part of the nitrogen nucleus.

These observations may be considered the origin of the science of nuclear physics. His simple apparatus presents a striking contrast to the gigantic "atom smashers" used in nuclear research today.

Further work in Cambridge and Vienna showed that all the elements between boron ($Z = 3$) and potassium ($Z = 19$) with the exception of oxygen and carbon could be disintegrated by fast alpha particles. In each case protons were emitted.

The disintegration of nitrogen can be represented symbolically by

$$_7N^{14} + {_2He^4} = {_8O^{17}} + {_1H^1}. \tag{10}$$

This indicates that one chemical element, namely, nitrogen, was transmuted into another distinct element, oxygen, in the process. In a limited sense the dreams of the alchemists had been achieved.

The Neutron

In 1930–31, Bothe and Becker in Germany showed that the element beryllium emitted a penetrating radiation when it was bombarded with alpha particles. Because of its great penetrating power they assumed that this radiation consisted of gamma rays of very short wave-length.

A year later, Irène Curie, daughter of the famous Curies who discovered radium, made a detailed study in collaboration with her husband, F. Joliot, of the absorption of this new radiation. Their experimental arrangement is shown schematically in Figure 524. Alpha rays from polonium P struck the beryllium B, producing the penetrating rays which were detected by means of the ionization chamber I and a very sensitive electrometer. While testing various absorbing screens S, they made the surprising discovery that if S was a thin sheet of paraffin the ionization current was increased, being approximately doubled.

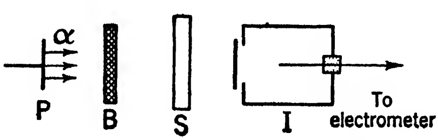


FIG. 524. THE CURIE-JOLIOT ARRANGEMENT FOR THE STUDY OF NEUTRONS

They suspected that the increased ionization was produced by fast protons being knocked out of the paraffin which is rich in hydrogen. By placing the paraffin inside a Wilson expansion chamber and photographing tracks they showed that their suspicions were correct. The momentum of a gamma ray photon is not high enough to give a proton the velocity that they observed. It was evident that this mysterious radiation coming from the beryllium was not ordinary gamma rays, as Bothe and Becker assumed.

As long as twelve years before this, Rutherford had pointed out that many facts concerning atoms could be interpreted in a very satisfactory manner by assuming the existence of a particle having the same mass as the proton but no charge. For such a particle he proposed the name neutron. Two separate attempts to detect the neutron in 1920 failed. When J. Chadwick saw an account of the experiments of Curie-Joliot he immediately realized that fast neutrons striking the hydrogen in the paraffin would give the observed effects. After further experiments which confirmed his theory he announced that the experimental evidence furnished conclusive proof of the existence of the neutron.

Since the neutron has no charge it experiences very little retardation and produces no ions in passing through a gas. In this respect it differs from alpha particles, beta particles, and protons. Owing to their charges, the latter remove electrons from atoms producing ions. The ions manifest their presence by electrical effects or by the tracks observed in the Wilson cloud chamber. Neutrons produce no cloud tracks. However, when they collide directly with protons, the latter are driven back with sufficient speed to produce a track.

Even dense solids, such as lead, show very slight absorption of neutrons. Slow neutrons can penetrate several inches of lead but are completely absorbed by a sheet of cadmium as thin as writing paper. Neutron behavior varies widely with speed. There is as much difference between the properties of fast and slow neutrons as between visible radiation and hard X-rays. Since neutrons are very penetrating and can knock protons out of molecules, they cause serious damage to living cells. Exposure to intense beams of neutrons produced with modern cyclotrons[1] may be fatal. To safeguard workers these machines are surrounded by tanks of water giving a screen several feet in thickness.

At first, attempts were made to consider the neutron a closely combined proton and electron. But in light of later experimental and theoretical information it seems preferable to consider it one of the fundamental constituent units of matter, on equal footing with the proton. The present generally accepted view is that all atomic nuclei are composed of protons and neutrons. This gives a simple and satisfactory interpretation of isotopes. Thus in the case of lithium, for example, the isotope of mass 6 has 3 protons and 3 neutrons while that of mass 7 has 3 protons and 4 neutrons. Adding a neutron to a nucleus does not change the atomic number Z or the number of electrons in an atom. It simply increases the mass by one unit producing an isotope of the original element.

Before 1932 it was supposed that all matter was composed of protons and electrons. Chadwick's discovery added a third fundamental particle, and a few months later Anderson in America discovered the positron, a particle similar to the electron but having a positive instead of a negative charge.

[1] See page 657.

As often happens in scientific research, Anderson's discovery was a "by-product" in his investigations on cosmic rays.

Concerning Cosmic Rays

During the last decade cosmic ray studies have enlisted the attention of a large number of scientists. As K. K. Darrow [1] has so aptly said,

> This field is unique in modern physics for the minuteness of the phenomena, the delicacy of the observations, the adventurous excursions of the observers, the subtlety of the analysis and the grandeur of the inferences.

The study of these rays began with the observation by Elster and Geitel and C. T. R. Wilson in 1900 that a closed ionization chamber filled with dry dust-free air showed a very weak but definite conductivity. Wilson suggested that the ionization might be produced by radiation similar to X-rays or cathode rays coming from outside the earth's atmosphere.

In order to find out whether this radiation came from the earth or from outer space the Swiss physicist Gockel in 1910 made balloon observations up to altitudes of 4500 meters. At this altitude the ionization was greater than at the earth's surface. Since 1922, Millikan and his collaborators and many others have made careful measurements of the intensity of this penetrating radiation at different altitudes and at different points on the earth's surface. All observers agree that the intensity increases with altitude reaching a maximum at approximately 31,000 meters above the earth's surface, where the atmospheric pressure is about 8 centimeters of mercury. Further, the intensity, especially at the higher altitudes, is much less near the equator than at high latitudes.

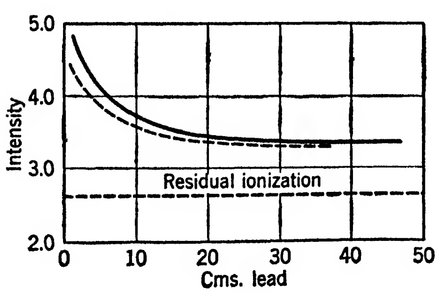


FIG. 525. PENETRATING POWER OF COSMIC RAYS IN LEAD

The great penetrating power of these cosmic rays is shown in Figure 525 which gives the intensity of the rays after passing through lead screens of different thickness. The first 10 centimeters of lead reduce the intensity about 25 per cent. Beyond this thickness the absorption is slight, indicating a very penetrating or hard radiation.

As Darrow has stated, the nature of cosmic rays has been the source of much discussion and speculation. Their penetrating power suggests gamma rays of extremely short wave-length. On the other hand their deviation in the earth's magnetic field indicates that they are electrically charged particles.

When the cosmic rays pass through matter, high-speed electrons are

[1] *Bell System Technical Journal, 11,* 148 (1932).

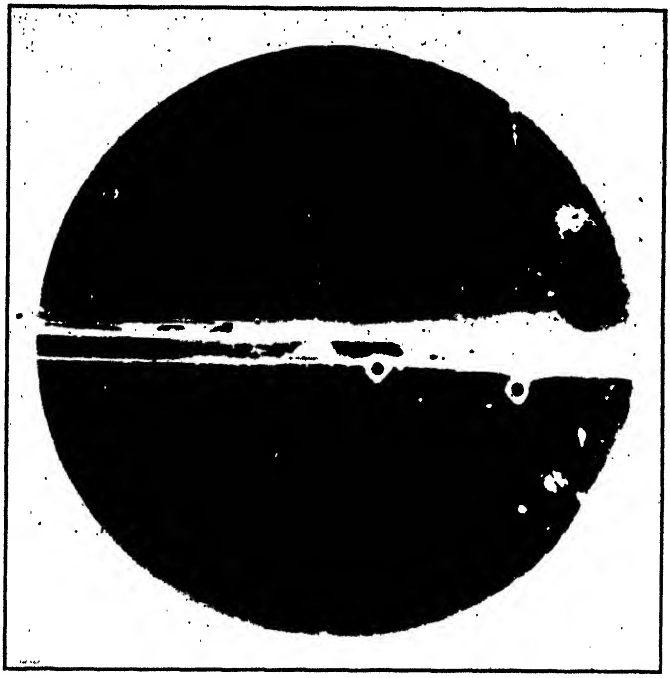


FIG. 526. ANDERSON'S DISCOVERY OF THE POSITRON
(From *The Physical Review*. Photograph provided by Dr. Carl D. Anderson.)

ejected which can be detected with Geiger-Mueller counters or with a Wilson cloud chamber. By placing the cloud chamber in a magnetic field the speeds of the electrons can be determined from the curvatures of their tracks. In August, 1932, while studying tracks by this method, C. D. Anderson, of the California Institute of Technology, obtained the photograph shown in Figure 526. The track is identical with those produced by electrons. But assuming that the direction of motion is upward it curves in the wrong direction. As shown in the photograph, the particle passed through a lead sheet six millimeters thick. If the assumption is made that the track was made by an electron traveling in the opposite direction it is impossible to avoid the embarrassing conclusion that it gained 40 million electron-volts of energy in passing through the lead. This is contrary to all the experimental and theoretical properties of the electron. The density of the track and the ability of the particle to penetrate six millimeters of lead definitely eliminates the hypothesis that the track was made by a proton. Hence, Anderson concluded that the track was produced by a particle of electronic mass but having a positive instead of a negative charge.

Additional work by Anderson and by Blackett showed that the positron tracks are usually accompanied by numerous electron tracks. These groups or "showers" of particles are produced when cosmic rays pass through matter.

Transmutation of Atoms by Fast Protons

In 1932, Cockroft and Walton, working in Rutherford's laboratory, showed that lithium, boron, and other light elements disintegrate with the emission of fast alpha particles when they are bombarded with high-speed protons.

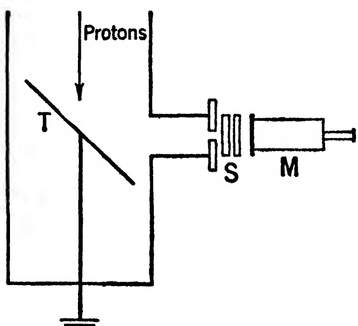


FIG. 527. THE FIRST OBSERVATION OF TRANSMUTATION BY PROTONS

After several years of research Cockroft and Walton constructed a high-voltage generator capable of producing a potential difference of 700,000 volts. Protons, generated in a hydrogen discharge tube, were accelerated in a high-voltage tube and struck the lithium target T, Figure 527. When the accelerating potential reached 125,000 volts, bright scintillations were observed on the zinc sulphide screen S. The scintillations were identical in appearance with those produced by alpha particles. Observation of the tracks in a cloud chamber showed that the particles coming from the lithium were actually alpha particles. The energy of these particles was comparable to that of the fastest alpha rays from radioactive substances.

In nuclear notation the transmutation can be written

$$_3Li^7 + {}_1H^1 = {}_2He^4 + {}_2He^4 + Q \tag{11}$$

where Q is the energy released in the process. In this case Q is 17 million electron-volts. This great release of energy shows that the incident protons must penetrate the lithium nucleus forming a structure composed of four neutrons and four protons. Such a nucleus, having a charge $+4\,e$, would be beryllium $_4Be^8$. The only known stable beryllium atom, however, is $_4Be^9$, which has a nuclear structure of 4 protons and 5 neutrons. The $_4Be^8$ is unstable and explodes producing two high-speed alpha particles.

The important difference between this disintegration and that achieved by Rutherford in 1919 is that the fast particles in this case were produced and controlled by the experimenter. Before this the only particles with sufficient energy to penetrate the nucleus were obtained from radium products and were, therefore, very limited in intensity. With larger and more powerful machines the transmutation of all elements was now within the range of possibilities.

The Cyclotron

The most ingenious and most successful device for producing high-speed ions is the cyclotron developed by E. O. Lawrence at the University of California. The cyclotron consists of two hollow *D*-shaped plates, called "dees" (Fig. 528, *A* and *B*), which are placed between the poles of a large electromagnet. The plates are in an evacuated chamber and are connected electrically to the terminals of a high frequency generator. In this manner an alternating potential difference of approximately 100,000 volts is applied to the "dees" producing an intense alternating electric field across the gap between *A* and *B*. If positive ions are produced at *P* near the center and the potential of *A* is negative, they will be accelerated toward *A*. Owing to the magnetic field perpendicular to the plane of the figure they will move in a circular path inside *A*. When they reach the gap at G_1 the potential has reversed, *C* being negative and *A* positive. Hence, they are accelerated at G_1, describe a larger semicircle, and are accelerated again at G_2. As their speed increases the ions describe larger and larger semicircles until they reach the periphery where they are deflected to the window *W* by a negatively charged plate *D*.

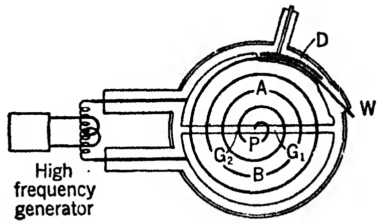


FIG. 528. THE CYCLOTRON

The successful performance of the apparatus depends upon the fact that the time required for an ion to describe a semicircle in the magnetic field is *independent* of the radius of the circle and is determined only by the ratio $\frac{\text{charge}}{\text{mass}}$ of the ion and the magnetic field intensity. This characteristic of the motion enables the ions to remain "in step" with the alternating field and acquire higher and higher velocities. The energy acquired by a given type of ion is proportional to B^2R^2 where B is the magnetic flux density and R the radius of the ion path just before it strikes the window. Since the maximum value of B is fixed by the magnetic saturation of iron, the only method of obtaining higher energies is to increase the radius R. With a radius of 15 inches Lawrence obtained protons, deuterons, and alpha particles (helium nuclei) having respectively energies of 4, 8, and 16 million electron-volts. His new cyclotron with $R = 30$ inches has already produced protons and deuterons with energies of 8 and 16 million electron-volts respectively. Theory indicates possible energies of 25 million electron-volts with deuterons and 50 million with helium nuclei.

Induced Radioactivity

In 1934, Irène Curie and F. Joliot observed a new type of nuclear disintegration. While investigating the emission of positrons from aluminum bombarded with alpha particles they found that positrons were emitted for several minutes after the source of alpha particles had been removed. The number of positrons emitted per second after the bombardment had ceased decreased with time according to the exponential law found in radioactivity (page 619). In other words, the aluminum nuclei were transmuted into radioactive nuclei by the alpha particle impacts. The half period of this radioactive aluminum was found to be 3 minutes 15 seconds. Similar effects were observed with boron and magnesium.

By chemical analysis the Joliots were able to show that the active substance in the case of aluminum was an isotope of phosphorus. Hence, the equations for the process are

$$_{13}Al^{27} + {_2He^4} = [_{15}P^{30}] + {_0n^1} \tag{12}$$

followed by

$$[_{15}P^{30}] = {_{14}Si^{50}} + {_{+1}e^0} \tag{13}$$

where $_{+1}e^0$ represents the positron. The square brackets indicate an unstable or radioactive nucleus.

Since 1934 radioactive isotopes of practically all the elements have been produced and studied. A recent survey by J. J. Livingood and G. T. Seaborg [1] catalogues the type of radiation, half life, and method of excitation for several hundred radioactive nuclei.

One of the most interesting cases of induced radioactivity is that of sodium discovered by E. O. Lawrence. He bombarded rock salt with deuterons of 2 million electron-volts energy and obtained radioactive sodium. The reaction can be written

$$_{11}Na^{23} + {_1H^2} = [_{11}Na^{24}] + {_1H^1}, \tag{14}$$

followed by

$$[_{11}Na^{24}] = {_{12}Mg^{24}} + {_{-1}e^0}, \tag{15}$$

with a half life of fifteen hours. The magnesium nucleus, however, is in an excited state and in falling to the normal state emits gamma rays. Hence, radiosodium emits beta and gamma rays identical in nature with those from natural radioactive substances. These gamma rays are more penetrating than those from radium. In most instances the amounts of radioactive substances produced by bombardment are very minute. The activity is detected with a Geiger-Mueller counter which counts *individual* electrons. With the recent large cyclotrons, however, it is possible to produce radiosodium in sufficient quantities for medical purposes.

The advent of induced radioactivity creates new possibilities in biological and chemical research. The active atoms are so easily detected that they can be traced through a process or reaction. They serve as marked or "tagged" atoms which can be located by their effect upon the counter.

Nuclear Fission

The nuclear transformations described so far have been effected by fast alpha particles, protons, or deuterons. The most generally effective transmuting agent, however, is the neutron. In the case of heavy elements the

[1] *Reviews of Modern Physics*, *12*, 30 (1940).

nuclear charge Ze is so large that electrostatic repulsion prevents the proton or deuteron from making a close encounter with the nucleus. There is no such difficulty with the neutron. Owing to its lack of charge it can easily penetrate a nucleus and provoke a transformation. The struck nucleus may eject an alpha particle, a proton, two neutrons, or gamma rays. An example of the first type is nitrogen:

$$_7N^{14} + {_0n^1} = {_5B^{11}} + {_2He^4}. \tag{16}$$

Many of the induced radioactive isotopes of the heavier elements have been produced by neutron bombardment.

In all the reactions described above, one of the final products of the transformation is a light particle; that is, an alpha particle, a proton, a neutron, or an electron. The other disintegration product has a mass which is not very different from that of the original bombarded element. In 1939, Hahn and Strassman found that, when a uranium nucleus is bombarded with neutrons, it may break up into two or more heavy fragments. Their observations indicated that the process was probably

$$_{92}U^{238} + {_0n^1} = {_{56}Ba^{148}} + {_{36}Kr^{91}} + Q. \tag{17}$$

These isotopes of barium and krypton are very unstable and emit electrons until a stable nucleus is obtained. Neutrons also are emitted in the process. The energy released, Q, is exceptionally large — more than 150 million electron-volts. Since the nuclear fragments are so large, this process is called ***nuclear fission*** in order to distinguish it from the more common type of reaction. Another astonishing observation was that the fission of uranium could be produced by either fast neutrons or slow thermal neutrons, but not by those of intermediate speed. Bohr attributed this to the existence of two isotopes of uranium, U^{235} and U^{238}. Later experiments have shown that Bohr's hypothesis is correct and that slow neutrons produce fission in U^{235}, but fast neutrons are required to break up U^{238}.

In each fission of uranium two or three neutrons are emitted. If each of these neutrons could produce fission in neighboring nuclei a "chain reaction" might occur with explosive violence. If the rate of reaction could be controlled a practical source of atomic power is within the range of possibilities. Calculation shows that the energy released by a pound of uranium in such a reaction would be a million times that given off by the combustion of a pound of coal. Although many laboratories are investigating possible methods of producing and controlling nuclear "chain reactions," no successful procedures have been announced.

A very interesting type of reaction was observed by Chadwick and Goldhaber in 1932. They found that the hard gamma rays from Thorium C″ caused the deuteron to break into a proton and neutron. The reaction can be written:

$$_1H^2 + h\nu = {_1H^1} + {_0n^1} \tag{18}$$

where the incident photon is represented by $h\nu$. In the photoelectric effect the energy of a photon is transferred to a single electron. In equation (18) the energy is transferred directly to a nucleus. By determining the lowest frequency at which the reaction (18) occurs it is possible to calculate the energy required to separate a proton and neutron. In this way the energy of cohesion of the deuteron was found to be 2.2 million electron-volts.

Production of Electron-Positron Pairs

After the discovery of the positron it was observed that hard gamma rays passing through matter often produced an electron-positron pair, the two particles having the same origin and equal energies. All the observations indicate that the photon disappears completely in the process. A little consideration shows that the production of pairs is consistent with the relation between mass and energy discussed on page 649. For in equation (4a), page 649, $(m - m_0)c^2$ is equal to an amount of kinetic energy. Hence, m_0c^2 can be considered as the intrinsic energy of the particle at rest due to, or associated with, its rest mass m_0. In order to be able to produce a pair, therefore, the frequency, ν, of the gamma rays must satisfy the relation

$$h\nu > 2\, m_0c^2 \tag{19}$$

where m_0 is the rest mass of the electron or positron. The experimental evidence indicates that the relation (19) is always satisfied. As might be expected, the observations show that the reverse process also is possible. That is, an electron and positron may combine and their total rest masses and kinetic energies be converted into two photons.

Since the rest mass of the photon is zero, some physicists speak of the above processes as the "interconversion of matter and radiant energy." This statement is probably too general since the word matter connotes more than electrons and positrons.

This interpretation of pair production, however, does emphasize the great change that has taken place in the last forty years concerning fundamental concepts in physics. At the close of the last century all physicists accepted the viewpoint that matter and energy were distinct entities, each having its own independent law of conservation. But in the last twenty years the atomic physicist has found it increasingly difficult to maintain a clear-cut distinction between those properties which are characteristic of matter and those which are characteristic of energy. In the fields of atomic and nuclear physics, Einstein's relation between mass and energy has proved to be a most useful guiding principle. For the ordinary large-scale phenomena of physics and chemistry, however, the conservation of energy and the conservation of mass are still independently valid and fundamental.

The Mesotron

The most recent addition to the list of fundamental particles is the *mesotron* (called *meson* by some writers) or heavy electron. Its existence was postulated in order to explain the great penetrating power of the "hard" cosmic rays. There is considerable evidence that a particle exists which has a mass between 150 and 250 times the electronic mass. According to the theory mesotrons may have either a positive or negative charge of electronic magnitude or may be neutral. It is further assumed that the particles of intermediate mass disintegrate producing electrons.

This rapid survey of nuclear physics shows that in this field, as in all other fields of science, the development has been one of rapid evolution, not revolution. The growth has occurred step by step, involving a continual interplay between theory and experimental technique. These characteristics have been well described by two of the most influential physicists of modern times, the Dutch theorist, H. A. Lorentz, and the founder of nuclear physics, Lord Rutherford.

In a lecture given at the Royal Institution in London in 1923 Lorentz said (71:17):

> One of the lessons which the history of science teaches us is surely this, that we must not too soon be satisfied with what we have achieved. The way of scientific progress is not a straight one which we can steadfastly pursue. We are continually seeking our course, now trying one path and then another, many times groping in the dark, and sometimes even retracing our steps. So it may happen that ideas, which we thought could be abandoned once for all have again to be taken up and come to new life.

Lord Rutherford, in surveying the growth of atomic physics in the twentieth century concluded with the following statement (88:73):

> I have also tried to show you that it is not in the nature of things for any one man to make a sudden violent discovery; science goes step by step, and the work of every man depends on the work of his predecessors. When you hear of a sudden unexpected discovery — a bolt from the blue as it were — you can always be sure that it has grown up by the influence of one man on another, and it is this mutual influence which makes the enormous influence of scientific advance. Scientists are not dependent on the ideas of a single man but on the combined wisdom of thousands of men, all thinking of the same problem, and each doing his little bit to add to the general structure of knowledge which is being gradually erected.

Questions for Self-Examination

1. Describe the principle of the mass spectrograph.
2. What is the evidence on the interconvertibility of mass and energy in the case of the electron?
3. What was the significance of Rutherford's disintegration of nitrogen?
4. What are neutrons and under what circumstances are they produced?
5. Describe the nature of cosmic rays.

6. Describe the cyclotron. Why is it such an important scientific instrument?
7. What is the nature and what are the potentialities of the process termed nuclear fission?
8. What can be said about two-way interconversion of matter and radiant energy in general?

Problems on Chapter 51

1. Show that the energy released in the nuclear reaction equation (10) is equivalent, within the limits of experimental error, to the mass that disappeared in the process. Precise nuclear masses are given on page 650.
2. Singly charged ions of Li^6 and Li^7 emerge from the slit S_3 in a Bainbridge mass spectrograph (Fig. 522) with the same velocity. If the Li^6 ions describe a circle of 20 centimeters radius, what is the radius of the path of the Li^7 ions? How far will their respective traces be separated on the photographic plate?
   23 cms and 6 cms.
3. The cyclotron has produced deuterons having energies of 16 million electron-volts. If the radius of the path described by the deuterons is 75 centimeters, what is the intensity of the magnetic field in the cyclotron? 1 weber/m^2.
4. The time required for an ion to traverse a semicircle of radius R in a transverse magnetic field of intensity B is $\pi R/v$, and the velocity v is given by equation (1) (page 648). Show that this time is independent of the radius of the path.
5. The mass of an electron and a positron is converted into radiation in the form of two photons of equal energy. What is the wave-length of the radiation produced?
   2.4×10^{-12} meters (.024 angstroms).

EPILOGUE

Freedom in the search for truth has been suffering from such a threat as it has never experienced before. Ironically, this threat came partly from the land where "Lehrfreiheit" and "Lernfreiheit" were first generally recognized and practiced. So far has Germany strayed from her own former ideals that in 1936, at an anniversary of the University of Heidelberg, Dr. Bernhard Rust, the Nazi Minister of Science and Education, could say:

> The old idea based on the sovereign right of abstract intellectual endeavor has gone forever. The new science is quite the opposite of uncontrolled search for truth which has been the ideal heretofore. The true freedom of science is to support the State and share its destiny and to make the search for truth subservient to this aim.

The outgrowth of this policy is written large in the condition of Europe and Asia in 1943. The United Nations have assumed the task of restoring freedom wherever it has been lost. The first step is the destruction of the power of those who are responsible for this relapse into intellectual barbarism. The next will be to furnish leadership in a return to the scientific ideal of disinterested search for truth.

Physics, as the pioneer science, can well set the standard for such a return. Also, it is potentially better equipped than any other agency for turning technology from destructive back to constructive ends. But if that potentiality is to be realized, physicists must be much more than mere specialists in the most technical of all sciences. They must inculcate breadth of view of their field and through it a comprehension of the nature of the rest of the intellectual enterprise. In the hope of contributing to this end, the present treatise adds to the conventional presentation of physics an opportunity for the student to become familiar with the struggles involved in establishing physics as the predecessor and prototype of all the sciences.

APPENDIX

List of References

This is not a bibliography. It is merely a list of the books to which some direct reference is made in the text. The arrangement is alphabetical by authors. References to this list in the text are by the numbers appearing at the left. Only books appear here. References to periodicals are completed in footnotes wherever made.

---

1. Alembert, Jean de la Rond d', *Traité de Dynamique.* Paris: Asher, 1758.

2. Al Khazini, *The Book of the Balance of Wisdom.* 1137. (See *Journal American Oriental Society,* VI, 1–128.)

3. Arago, D. F. J., *Oeuvres Complètes.* Paris: Gide, 1862.

4. Archer, Gleason L., *History of Radio to 1926.* Chicago: American Historical Society Inc., 1938.

5. Archimedes, *The Works of Archimedes.* T. L. Heath, translator and editor; London: Cambridge University Press, 1897.

6. Aristotle, *De Caelo.* J. L. Stocks, translator; New York: Oxford University Press, 1922.

7. Aristotle, *Problems.* Loeb Classical Library. W. S. Hett, translator; Cambridge, Massachusetts: Harvard University Press, 1936–.

8. Aristotle, *Questiones Mechanicae.* Van Cappelle, editor; Amsterdam, 1812.

9. Aristotle, *Works of.* E. W. Webster, translator; New York: Oxford University Press, 1932.

10. Atwood, George, *Treatise on Rectilinear Motion and Rotation of Bodies.* Cambridge, 1784.

11. Bacon, Francis, *Novum Organum.* J. Devey, editor; New York: Collier, 1902.

12. Bacon, Roger, *Opera Inedita.* J. S. Brewer, editor; New York: Longmans, 1859.

13. Bacon, Roger, *Opus Majus.* R. B. Burke, translator; Philadelphia: University of Pennsylvania Press, 1928.

14. Ball, W. W. R., *An Essay on Newton's* PRINCIPIA. New York: Macmillan, 1893.

15. Ball, W. W. R., *A Short Account of the History of Mathematics.* New York: Macmillan, 1901.

16. Bartholinus, Erasmus, *Experimenta Crystalli Islandici,* Copenhagen, Sumptibus D. Paulli; Hafniae, 1669.

17. Benedetti, G. B., *De Mechanicis,* included as a section in his *Diversarum Speculationum Liber.* Taurini, 1599.

18. Bernouilli, Daniel, *Hydrodynamica sive viribus et motibus fluidorum commentarii* Argent, 1738.

19. Birch, Thomas, *The History of the Royal Society of London.* London: Millar, 1757.

20. Bishop, Morris, *Pascal the Life of Genius.* New York: Reynal and Hitchcock, 1936.

21. Black, Joseph, M.D., *Lectures on Elements of Chemistry.* J. Robinson, translator; London: Longman and Rees, 1803.

22. Browne, Sir Thomas, *Pseudodoxia Epidemica.* London: E. Dod, 1646.

23. Bullialdus, Ismaelis, *Astronomia Philolaica.* Paris: S. Piget, 1645.

24. Cabeo, Nicolo, *Philosophia Magnetica.* Cologne: Kinckium, 1629.

25. Cajori, F., *A History of Physics.* New York: Macmillan, 1929.

26. Cavendish, Henry, *Scientific Papers.* 2 vols.; London: Cambridge University Press, 1921.

27. Clarke, George A., *Clouds*, London, Constable & Co., 1920.

27a. Cohen, I. Bernard, *Roemer and the First Determination of the Velocity of Light.* The Burndy Library, Inc., 1942.

27b. Cohen I. Bernard, *Franklin's "Experiments and Observations on Electricity."* Cambridge, Massachusetts: Harvard University Press, 1941.

28. Cooper, Lane, *Aristotle, Galileo and the Tower of Pisa.* Ithaca, New York: Cornell University Press, 1935.

29. Coulomb, Charles Augustin, *Theorie des Machines Simples.* Paris: Bachelier, 1821.

30. Crew, Henry, *The Rise of Modern Physics.* Baltimore: Williams and Wilkins, 1935.

31. Crew, Henry, *The Wave Theory of Light.* New York: American Book Co., 1900.

32. Crowther, J. G., *British Scientists of the Nineteenth Century.* London: Kegan Paul, Trench, Treubner & Co., 1935.

33. Dampier-Whetham, W. C., *A History of Science and its Relations with Philosophy and Religion.* New York: Macmillan, 1929.

34. Darrow, K. K., *The Renaissance of Physics.* New York: The Macmillan Company, 1936.

35. de Broglie, Louis, *Matter and Light, The New Physics.* W. H. Johnston, translator; New York: W. W. Norton and Company, 1939.

36. Dodge, H. L., *Problems in Physics.* United States War Department, 1920.

37. Euler, Leonhard, *Theoria Motus Corporum Solidorum et Rigidorum.* Rostock, 1765.

38. Fahrenheit, G. D., *Experimente und Beobachtungen uber das Gefrieren des Wassers im Vacuum.* Ostwald, *Klassiker*, no. 57.

39. Faraday, Michael, *Experimental Researches in Electricity.* 3 vols.; London: Taylor & Francis, 1839–55.

40. Faraday, Michael, *Faraday's Diary.* 7 vols.; London: G. Bell & Sons, 1932.

41. Fletcher, Harvey, *Speech and Hearing.* New York: Van Nostrand, 1929.

42. Fourier, Jean B. J., *Analytical Theory of Heat.* Alexander Freeman, translator; London: Cambridge University Press, 1878.

43. Franklin, Benjamin, *Experiments and Observations in Electricity*, I. Bernard Cohen, editor. Harvard University Press, 1941.

44. Franklin, Benjamin, *Works of Benjamin Franklin.* Jared Sparks, editor; Chicago: Townsend MacCoun, 1882.

45. Fraunhofer, J., *Gesammelte Schriften.* E. Lommel, Munich, 1888.

46. Galilei, Galileo, *Dialogues Concerning Two New Sciences.* Henry Crew and Alfonso de Salvio, translators; Northwestern University Studies, 1933.

47. Gilbert, William, *De Magnete.* 1600. P. F. Mottelay, translator; New York: Wiley, 1904.

48. Gow, James, *A Short History of Greek Mathematics.* London: Cambridge University Press, 1884.

49. Grimaldi, Franciscus Maria., *Physico, mathesis de lumine coloribus et iride.* Bononiae, V. Benatii, 1645.

50. Harlow, Alvin F., *Old Wires and New Waves.* New York: D. Appleton-Century Co., 1936.

51. Hart, Ivor B., *The Mechanical Investigations of Leonardo da Vinci.* Chicago: The Open Court Publishing Company, 1925.

52. Heaviside, Oliver, *Electrical Papers*. New York: Macmillan, 1892.

53. Helmholtz, Hermann von, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. A. J. Ellis, translator; London: Longmans, Green, 1885.

54. Henry, Joseph, *Annual Report of the Smithsonian Institution*, 1856, pp. 221–34.

55. Hirn, G. A., *Exposition analytique et expérimentale de la théorie mécanique de chaleur*. Mallet-Bachelier, 1862.

56. Hooke, Robert, *De Potentia Restitutiva*, London: J. Martyn, 1678.

57. Hooke, Robert, *Micrographia*. London: J. Allestry, 1667.

58. Hoppe, Edmund, *Geschichte der Optik*. Leipzig: J. J. Weber, 1926.

59. Huygens, Christiaan, *Horologium Oscillatorium*. Ostwald, *Klassiker*, no. 192.

60. Huygens, Christiaan, *Oeuvres Complètes*. 18 vols.; Société Hollandaise des Sciences, 1888–.

61. Huygens, Christiaan, *Treatise on Light*. S. P. Thompson, translator; London; Macmillan, 1912.

62. Jones, H. Bence, *The Life and Letters of Faraday*. 2 vols.; New York: Longmans, Green, 1870.

63. Kelvin, Lord (Sir William Thomson), *Mathematical and Physical Papers*. London: Cambridge University Press, 1882–1911.

64. Kepler, Johann, *Ad Vitellionem Paralipomena*, Freft, 1604.

65. Kepler, Johann, *Dioptrice*, Francus, 1611.

66. Knudsen, Vern O., *Architectural Acoustics*. New York: Wiley & Sons, 1932.

67. Leibnitz, Gottfried W. von, *Opera Omnia*. L. Dutens, editor; Geneva: apud fratres de Tournes, 1768.

68. Lenard, Philip, *The Great Men of Science*. H. S. Hatfield, translator; New York: Macmillan, 1933.

69. Lockyer, Joseph Norman, *Studies in Spectrum Analysis*. London: Kegan Paul, 1878.

70. Lodge, Oliver, *Talks about Radio*. New York: Doran, 1925.

71. Lorentz, H. A., *Collected Papers*. 9 vols.; The Hague: Martinus Nijhoff, 1935–39.

72. Lucretius, *De. Rerum Natura*. R. C. Trevelyan, translator; London: Cambridge University Press, 1937.

73. Mach, Ernst, *Principles of Physical Optics*. J. S. Anderson and A. F. A. Young, translators; London: Methuen, 1926.

74. Mach, Ernst, *The Science of Mechanics*. T. J. McCormack, translator; Chicago: The Open Court Publishing Company, 1907.

75. McKie, D., and N. H. de V. Heathcote, *The Discovery of Specific and Latent Heats*. London: Edward Arnold & Co., 1935.

76. Magie, W. F., *The Second Law of Thermodynamics*. New York: Harper & Brothers, 1899.

77. Magie, W. F., *A Source Book in Physics*. New York: McGraw-Hill, 1935.

78. Major, R. H., *Select Letters of Christopher Columbus*. London, 1847–49.

79. Mariotte, Edme, *Histoire de l' Académie Royale des Sciences*. Paris. 1773.

80. Mariotte, Edme, *Oeuvres de M. Mariotte*. P. Vander Aa, 1717.

81. Maurolycus, Franciscus, *Photismi de Lumine*. 1611. Henry Crew, translator; New York: Macmillan, 1940.

82. Maxwell, James Clerk, *Theory of Heat*. New York: Longmans, Green, 1891.

83. Michelson, A. A., *Détermination Experimentale de la Valeur du Mètre*. Paris: Gauthier-Villars, 1894.

84. Miller, D. C., *Anecdotal History of the Science of Sound*. New York: Macmillan, 1935.

85. Miller, D. C., *The Science of Musical Sounds*. New York: Macmillan, 1916.

85a. Miller, D. C., *Sound Waves; Their Shape and Speed*. New York: The Macmillan Company, 1937.

85b. Millikan, Robert A., *Electrons, Protons, Photons, Neutrons, and Cosmic Rings.* Chicago: University of Chicago Press, 1935.

86. Mottelay, Paul F., *A Bibliographical History of Electricity and Magnetism.* London: Griffin, 1922.

87. Mott-Smith, Morton C., *The Story of Energy.* New York: D. Appleton-Century Co., 1934.

88. Needham, Joseph, and Walter Pagel, editors, *The Background to Modern Science.* New York: The Macmillan Company, 1938.

89. Newton, Isaac, *Optical Lectures.* London, 1728.

90. Newton, Isaac, *Opticks.* E. T. Whittaker, editor; New York: McGraw-Hill, 1931. Reprinted from the 4th edition, 1730.

91. Newton, Isaac, *Principia Mathematica.* Florian Cajori, editor; Berkeley, California: University of California Press, 1934.

92. Norman, Robert, *A Discourse of the Variation of the Cumpass or Magneticall Needle.* London, 1581.

93. Norman, Robert, *Newe Attractive,* London, 1581.

94. Ohm, Georg S., *Die Galvanische Kette.* Berlin: Riemann, 1827.

95. Ostwald, W., *Klassiker der Exakten Wissenschaften.* Leipzig: Englemann, 1888–.

96. Peregrine, Peter, *Epistle . . . to Sygerus of Foucancourt concerning the magnet.* S. P. Thompson, translator; London: Oxford, 1902.

97. Pliny (Plinius Secundus, Caius) *Natural History.* 2d cent. A.D. John Bostock and H. T. Riley, translators; 6 vols.; London: Henry G. Bohn, 1857.

98. Porta, Giambattista della, *Magiae Naturalis sive De Miraculis Rerum Naturalium.* Cancer, 1558.

99. Porta, Giambattista della, *Natural Magick.* Translated from the 1589 edition of *Magiae Naturalis.* J. Wright, 1669.

100. Powell, Baden, *Historical View of the Progress of the Physical and Mathematical Sciences.* Lardner's *Cabinet of Natural Philosophy,* vol. II. 1884.

101. Powell, Baden, *History of Natural Philosophy.* Oxford, 1834.

102. Poynting, J. H., and J. J. Thompson, *Heat.* London: C. Griffin & Co., 1904.

103. Preston, Thomas, *Theory of Light.* 4th edition; New York: Macmillan, 1912.

104. Prevost, Pierre, *Sur l'Équilibre du Feu.* Genève: J. J. Paschoud, 1792.

105. Priestley, Joseph, *History and Present State of Discoveries Relating to Vision, Light and Colours.* London: Johnson, 1772.

106. Randall, W. W., editor, *The Expansion of Gases by Heat.* New York: American Book Co., 1902.

107. Rankine, W. J. M., *Miscellaneous Scientific Papers.* London: C. Griffin & Co., 1860.

108. Robins, Benjamin, *New Principles of Gunnery.* J. Nourse, 1742.

109. Roentgen, W. K., *On a New Kind of Rays.* G. F. Barker, translator; New York: Harper and Brothers, 1899.

110. Sabine, Wallace C., *Collected Papers on Acoustics.* Cambridge, Massachusetts: Harvard University Press, 1922.

111. Sarton, G., *Introduction to the History of Science.* Baltimore: Williams & Wilkins, 1927–.

112. Scheele, Carl W., *Chemical Treatise on Air and Fire.* L. Dobbin, translator; London: Bell, 1931.

113. Schelling, F. W. T. von, *Vorlesungen uber die Methode des Akademischen Studiums.* 1802. *Sämtliche Werke.* Stuttgart, 1856.

114. Seneca, Lucius Annaeus, *Physical Science in the Time of Nero.* Translated from the *Quaestiones Naturales* by John Clarke; New York: Macmillan, 1910.

115. Singer, Charles, *Religion and Science.* Benn's Sixpenny Library no. 144. London: Ernest Benn, Ltd., 1928.

116. Singer, Charles, editor, *Studies in the History and Method of Science.* Oxford: Clarendon Press, 1917 and 1921.

117. Smith, Preserved, *A History of*

Modern Culture, vol. I. New York: Henry Holt, 1930.

118. Soemmering, S. T., *Uber einen elektrischen Telegraphen.* Munich, 1811.

119. Stallo, J. B., *General Principles of the Philosophy of Nature.* London: Chapman, 1848.

120. Stevin, Simon, *Oeuvres Mathematiques.* A. Girard, editor; Elzevier, 1634.

121. Taylor, L. W., *College Manual of Optics.* Boston: Ginn, 1926.

122. Thompson, S. P., *Dynamo-Electric Machinery.* 5th edition; 2 vols.; London: Spon, 1897.

123. Thompson, S. P., *Michael Faraday.* New York: Macmillan, 1898.

124. Torricelli, Evangelista, *Opere.* Florence, 1919.

125. Turner, Dorothy M., *Makers of Science; Electricity and Magnetism.* New York: Oxford University Press, 1927.

126. Tyndall, John, *Heat Considered as a Mode of Motion.* New York: Appleton, 1863.

127. Ubaldi, Guidi, *Mechanicorum Liber*, Pisauri, Hier Concordia, 1577.

128. Walker, Ralph, *Treatise on the Magnet.* London, 1798.

129. Wallis, John, *Opera Mathematica et Miscellanea. Oxonii*, 1693–99.

130. Wilde, Emil, *Geschichte der Optik*, 2 vols.; Berlin: Rucher und Püchler, 1838.

131. Wolf, Abraham, *A History of Science, Technology and Philosophy in the Eighteenth Century.* London: George Allen & Unwin, Ltd., 1938.

132. Wolf, Abraham, *A History of Science, Technology and Philosophy in the Sixteenth and Seventeenth Centuries.* London: George Allen & Unwin, Ltd., 1935.

133. Wood, Alexander, *The Physical Bases of Music.* London: Cambridge University Press, 1925.

134. Young, Thomas, *Course of Lectures on Natural Philosophy and the Mechanical Arts.* London: Cox, 1845.

135. Zahm, J. A., *Sound and Music.* Chicago: McClurg, 1892.

The Meter-Kilogram-Second System of Units

In 1935 the International Committee on Weights and Measures, the central authority on all questions of international scientific standards, legislated that

> the actual substitution of the absolute system of electrical units [the M.K.S. system] shall take place on January 1, 1940.

The war will delay the realization of this plan in Europe, but there is little reason why it should not be put into effect here. It is a highly desirable simplification which had been debated for more than a generation before finally receiving official sanction.

In accordance with that provision, the M.K.S. system of units has been incorporated into this text. It displaces the C.G.S. system, first officially sanctioned in 1881. In its application to mechanics the M.K.S. system constitutes a relatively minor departure from the older system. It involves the substitution of the meter for the centimeter as the basic unit of length and the kilogram for the gram as the basic unit of mass. These substitutions seem the more conservative in that the world's standard of length is the meter and that the world's standard of mass is the kilogram. Along with a number of other anomalies, the M.K.S. system clears up that of two of the world's standards not being used as the basic units.

But the real motivation for the change is the desire to clarify electrical units. Up to now the common or "practical" units have been in reality outside the basic scheme of fundamental units. The principal feature of the M.K.S. system is that the simple shifts from centimeter to meter and from gram to kilogram legitimatize the practical units and remove all occasion for further use of the electrostatic and electromagnetic units.

This should make a strong appeal to those who teach general physics. Though physics has cold-shouldered the practical units for two generations and more, it has never been able to decide which of the two other common sets of units it preferred. The effect on beginning students has been devastating. In a subject which would beset them with an unusual array of conceptual hazards even if there were only one system of units to deal with, they have had to learn three different units for almost every electrical entity, and to shift from one to another as the requirements of a particular problem or the whim of the teacher indicated. This difficulty disappears as the M.K.S. system comes in and the old electrostatic and electromagnetic systems go into the discard.

To facilitate the process of transition from the old to the new system a conversion table of the units most commonly used is appended. Inspection will show that for the most part M.K.S. units are of magnitudes more convenient to use than C.G.S. units, once one has learned to think in the new terms. Aside from length and mass, the absurdly small unit of force, the *dyne*, is replaced by the *newton*, 10^5 times as large. The still more minuscule *erg* is replaced by the *joule*, 10^7 times as large. The awkward *erg-per-second* is replaced by the *watt*, a most natural unit of power. With electrical units the advantages of the M.K.S. system are still more impressive. In a few cases the advantage lies mildly the other way. The density of water, instead of being numerically unity, is 1000 kgm/m^3, with that of other substances in proportion. Specific gravities, of course, remain unaffected. Resistivities are inconveniently small when the cross-sectional unit is a square meter. It is possible that as a practical matter constants of this variety may retain their old values. Even the C.G.S. system condescended to express resistivities in ohms instead of in electromagnetic units!

At the infrequent points where a choice between the so-called *rationalized* and *unrationalized* electrical units has been necessary, the rationalized units have been specified. This issue, a relatively minor one as far as general physics is concerned, has not yet been decided by the aforementioned international standardizing agency.

Table of a Few of the More Common Physical Entities, Comparing M.K.S. and C.G.S. Units

(The International Committee on Weights and Measures prescribed in 1935 that the M.K.S. system of units should displace the C.G.S. system as of January 1, 1940.)

Entity	M.K.S. Unit	C.G.S. Unit	Ratio M.K.S. Unit/ C.G.S. Unit
Length	Meter	Centimeter	10^{2}
Mass	Kilogram	Gram	10^{3}
Time	Second	Second	1
Density	Kilograms per cubic meter	Grams per cubic centimeter	10^{-3}
Velocity	Meters per second	Centimeters per second	10^{2}
Acceleration	Meters per second²	Centimeters per second²	10^{2}
Force	Newton	Dyne	10^{5}
Torque	Newton meter	Dyne centimeter	10^{7}
Moment of inertia	Kilogram meter²	Gram centimeter²	10^{7}
Work	Joule	Erg	10^{7}
Power	Watt	Erg per second	10^{7}
Quantity of heat	Kilogram-calorie	Gram-calorie	10^{3}
Luminous intensity	International candle	International candle	1
Luminous flux	Lumen	Lumen	1
Illumination	Lux	Phot	10^{-4}
Current	Ampere	Abampere	10^{-1}
Potential difference	Volt	Abvolt	10^{8}
Resistance	Ohm	Abohm	10^{9}
Resistivity	Ohm-meter	Abohm-centimeter	10^{11}
Quantity	Coulomb	Abcoulomb	10^{-1}
Capacitance	Farad	Abfarad	10^{-9}
Inductance	Henry	Abhenry	10^{9}
Field strength	Ampere turns per meter	Gauss	10^{4}
Magnetic flux	Weber	Maxwell	10^{8}
Flux density	Weber per square meter	Gauss	10^{4}

Certain Physical Constants in M.K.S. Units

(Values are given to four significant figures. Only three are required in the solutions of the problems at the end of Chapters 48–51.)

Electronic charge	$e = 1.601 \times 10^{-19}$ coulombs
Electronic ratio of charge to mass	$e/m = 1.76 \times 10^{11} \dfrac{\text{coulombs}}{\text{kgm}}$
Mass of electron	$m = 9.103 \times 10^{-31}$ kgm
Planck's constant	$h = 6.610 \times 10^{-34}$ joule-secs
Mass associated with unit atomic weight	$m_1 = 1.659 \times 10^{-27}$ kgm
Avogadro's number	$N = 6.028 \times 10^{23} \dfrac{\text{molecules}}{\text{mole}}$
Speed of light	$c = 2.998 \times 10^{8}$ meters/sec
Permittivity of space	$k_0 = 8.854 \times 10^{-12}$
Permeability of space	$\mu_0 = 1.257 \times 10^{-6}$

Periodic Table

Group	O	I	II	III	IV
Period 1		H 1 1.008			
2	He 2 4.003	Li 3 6.94	Be 4 9.02	B 5 10.82	C 6 12.01
3	Ne 10 20.18	Na 11 22.997	Mg 12 24.32	Al 13 26.97	Si 14 28.06
4 A	A 18 39.94	K 19 39.10	Ca 20 40.08	Sc 21 45.10	Ti 22 47.90
4 B		Cu 29 63.57	Zn 30 65.38	Ga 31 69.72	Ge 32 72.60
5 A	Kr 36 83.70	Rb 37 85.48	Sr 38 87.63	Y 39 88.92	Zr 40 91.22
5 B		Ag 47 107.88	Cd 48 112.41	In 49 114.76	Sn 50 118.70
6 A	Xe 54 131.30	Cs 55 132.91	Ba 56 137.36	57-71 Rare Earths	Hf 72 178.60
6 B		Au 79 197.20	Hg 80 200.61	Tl 81 204.39	Pb 82 207.21
7	Rn 86 222.00	? 87	Ra 88 226.05	Ac 89 228	Th 90 232.12

Rare Earths 57-71	La 57 138.92	Ce 58 140.13	Pr 59 140.92
	Gd 64 156.90	Tb 65 159.20	Dy 66 162.46

Periodic Table (continued)

V	VI	VII	VIII		
N 7 14.008	O 8 16.00	F 9 19.00	N 7 14.008		
P 15 30.98	S 16 32.06	Cl 17 35.46			
Va 23 50.95	Cr 24 52.01	Mn 25 54.93	Fe 26 55.85	Co 27 58.94	Ni 28 58.69
As 33 74.91	Se 34 78.96	Br 35 79.92			
Cb 41 92.91	Mo 42 95.95	Ma 43 ?	Ru 44 101.70	Rh 45 102.91	Pd 46 106.70
Sb 51 121.76	Te 52 127.61	I 53 126.92			
Ta 73 180.88	W 74 183.92	Re 75 186.31	Os 76 190.20	Ir 77 193.10	Pt 78 195.23
Bi 83 209.00	Po 84 210	? 85			
Pa 91 231	U 92 238.07				

Nd 60 144.27	Il 61 ?	Sm 62 150.43	Eu 63 152.00	
Ho 67 164.94	Er 68 167.20	Tm 69 169.40	Yb 70 173.04	Lu 71 174.99

Atomic Weights

Element	*Symbol*	*Atomic Number*	*Atomic Weight*	*Element*	*Symbol*	*Atomic Number*	*Atomic Weight*
Aluminum	Al	13	26.97	Molybdenum	Mo	42	95.95
Antimony	Sb	51	121.76	Neodymium	Nd	60	144.27
Argon	A	18	39.94	Neon	Ne	10	20.18
Arsenic	As	33	74.91	Nickel	Ni	28	58.69
Barium	Ba	56	137.36	Nitrogen	N	7	14.008
Beryllium	Be	4	9.02	Osmium	Os	76	190.20
Bismuth	Bi	83	209.00	Oxygen	O	8	16.00
Boron	B	5	10.82	Palladium	Pd	46	106.70
Bromine	Br	35	79.92	Phosphorus	P	15	30.98
Cadmium	Cd	48	112.41	Platinum	Pt	78	195.23
Calcium	Ca	20	40.08	Potassium	K	19	39.10
Carbon	C	6	12.01	Praseodymium	Pr	59	140.92
Cerium	Ce	58	140.13	Radium	Ra	88	226.05
Cesium	Cs	55	132.91	Radon	Rn	86	222.00
Chlorine	Cl	17	35.46	Rhenium	Re	75	186.31
Chromium	Cr	24	52.01	Rhodium	Rh	45	102.91
Cobalt	Co	27	58.94	Rubidium	Rb	37	85.48
Columbium	Cb	41	92.91	Ruthenium	Ru	44	101.70
Copper	Cu	29	63.57	Samarium	Sm	62	150.43
Dysprosium	Dy	66	162.46	Scandium	Sc	21	45.10
Erbium	Er	68	167.20	Selenium	Se	34	78.96
Europium	Eu	63	152.00	Silicon	Si	14	28.06
Fluorine	F	9	19.00	Silver	Ag	47	107.88
Gadolinium	Gd	64	156.90	Sodium	Na	11	22.997
Gallium	Ga	31	69.72	Strontium	Sr	38	87.63
Germanium	Ge	32	72.60	Sulfur	S	16	32.06
Gold	Au	79	197.20	Tantalum	Ta	73	180.88
Hafnium	Hf	72	178.60	Tellurium	Te	52	127.61
Helium	He	2	4.003	Terbium	Tb	65	159.20
Holmium	Ho	67	164.94	Thallium	Tl	81	204.39
Hydrogen	H	1	1.008	Thorium	Th	90	232.12
Indium	In	49	114.76	Thulium	Tm	69	169.40
Iodine	I	53	126.92	Tin	Sn	50	118.70
Iridium	Ir	77	193.10	Titanium	Ti	22	47.90
Iron	Fe	26	55.85	Tungsten	W	74	183.92
Krypton	Kr	36	83.70	Uranium	U	92	238.07
Lanthanum	La	57	138.92	Vanadium	Va	23	50.95
Lead	Pb	82	207.21	Xenon	Xe	54	131.30
Lithium	Li	3	6.94	Ytterbium	Yb	70	173.04
Lutecium	Lu	71	174.99	Yttrium	Y	39	88.92
Magnesium	Mg	12	24.32	Zinc	Zn	30	65.38
Manganese	Mn	25	54.93	Zirconium	Zr	40	91.22
Mercury	Hg	80	200.61				

Table of Elastic Moduli and Breaking Strengths

(Tabular values multiplied by 10^{10} give newtons/m^2)

Material	Young's modulus	Rigidity modulus	Volume modulus	Breaking strength
Aluminum	7	2.5	7	.010
Copper	10	4.2	12	.020
Cast Iron	11	5	10	.033
Steel (mild)	22	8	16	.050

Greek Alphabet

Greek letter	Greek name	English equivalent	Greek letter	Greek name	English equivalent
Α α	Alpha	a	Ν ν	Nu	n
Β β	Beta	b	Ξ ξ	Xi	x
Γ γ	Gamma	g	Ο ο	Omicron	ŏ
Δ δ	Delta	d	Π π	Pi	p
Ε ε	Epsilon	ĕ	Ρ ρ	Rho	r
Ζ ζ	Zeta	z	Σ σ	Sigma	s
Η η	Eta	ē	Τ τ	Tau	t
Θ θ	Theta	th	Υ υ	Upsilon	u
Ι ι	Iota	i	Φ φ	Phi	ph
Κ κ	Kappa	k	Χ χ	Chi	ch
Λ λ	Lambda	l	Ψ ψ	Psi	ps
Μ μ	Mu	m	Ω ω	Omega	ō

Natural Sines 0°–45°									Proportional Parts (Add)		
°	0′	10′	20′	30′	40′	50′		°	1 2 3	4 5 6	7 8 9
0	.0000	.0029	.0058	.0087	.0116	.0145	.0175	89	3 6 9	12 14 17	20 23 26
1	.0175	.0204	.0233	.0262	.0291	.0320	.0349	88	3 6 9	12 14 17	20 23 26
2	.0349	.0378	.0407	.0436	.0465	.0494	.0523	87	3 6 9	12 14 17	20 23 26
3	.0523	.0552	.0581	.0610	.0640	.0669	.0698	86	3 6 9	12 14 17	20 23 26
4	.0698	.0727	.0756	.0785	.0814	.0843	.0872	85	3 6 9	12 14 17	20 23 26
5	.0872	.0901	.0929	.0958	.0987	.1016	.1045	84	3 6 9	12 14 17	20 23 26
6	.1045	.1074	.1103	.1132	.1161	.1190	.1219	83	3 6 9	12 14 17	20 23 26
7	.1219	.1248	.1276	.1305	.1334	.1363	.1392	82	3 6 9	12 14 17	20 23 26
8	.1392	.1421	.1449	.1478	.1507	.1536	.1564	81	3 6 9	12 14 17	20 23 26
9	.1564	.1593	.1622	.1650	.1679	.1708	.1736	80	3 6 9	12 14 17	20 23 26
10	.1736	.1765	.1794	.1822	.1851	.1880	.1908	79	3 6 9	12 14 17	20 23 26
11	.1908	.1937	.1965	.1994	.2022	.2051	.2079	78	3 6 8	11 14 17	20 22 25
12	.2079	.2108	.2136	.2164	.2193	.2221	.2250	77	3 6 8	11 14 17	20 22 25
13	.2250	.2278	.2306	.2334	.2363	.2391	.2419	76	3 6 8	11 14 17	20 22 25
14	.2419	.2447	.2476	.2504	.2532	.2560	.2588	75	3 6 8	11 14 17	20 22 25
15	.2588	.2616	.2644	.2672	.2700	.2728	.2756	74	3 6 8	11 14 17	20 22 25
16	.2756	.2784	.2812	.2840	.2868	.2896	.2924	73	3 6 8	11 14 17	20 22 25
17	.2924	.2952	.2979	.3007	.3035	.3062	.3090	72	3 6 8	11 14 17	20 22 25
18	.3090	.3118	.3145	.3173	.3201	.3228	.3256	71	3 6 8	11 14 17	20 22 25
19	.3256	.3283	.3311	.3338	.3365	.3393	.3420	70	3 5 8	11 14 16	19 22 24
20	.3420	.3448	.3475	.3502	.3529	.3557	.3584	69	3 5 8	11 14 16	19 22 24
21	.3584	.3611	.3638	.3665	.3692	.3719	.3746	68	3 5 8	11 14 16	19 22 24
22	.3746	.3773	.3800	.3827	.3854	.3881	.3907	67	3 5 8	11 14 16	19 22 24
23	.3907	.3934	.3961	.3987	.4014	.4041	.4067	66	3 5 8	11 14 16	19 22 24
24	.4067	.4094	.4120	.4147	.4173	.4200	.4226	65	3 5 8	11 14 16	19 22 24
25	.4226	.4253	.4279	.4305	.4331	.4358	.4384	64	3 5 8	10 13 16	18 21 23
26	.4384	.4410	.4436	.4462	.4488	.4514	.4540	63	3 5 8	10 13 16	18 21 23
27	.4540	.4566	.4592	.4617	.4643	.4669	.4695	62	3 5 8	10 13 16	18 21 23
28	.4695	.4720	.4746	.4772	.4797	.4823	.4848	61	3 5 8	10 13 16	18 21 23
29	.4848	.4874	.4899	.4924	.4950	.4975	.5000	60	3 5 8	10 13 15	18 20 23
30	.5000	.5025	.5050	.5075	.5100	.5125	.5150	59	3 5 8	10 13 15	18 20 23
31	.5150	.5175	.5200	.5225	.5250	.5275	.5299	58	3 5 8	10 13 15	18 20 23
32	.5299	.5324	.5348	.5373	.5398	.5422	.5446	57	3 5 8	10 13 15	18 20 23
33	.5446	.5471	.5495	.5519	.5544	.5568	.5592	56	3 5 7	10 12 14	17 19 22
34	.5592	.5616	.5640	.5664	.5688	.5712	.5736	55	3 5 7	10 12 14	17 19 22
35	.5736	.5760	.5783	.5807	.5831	.5854	.5878	54	2 5 7	10 12 14	17 19 22
36	.5878	.5901	.5925	.5948	.5972	.5995	.6018	53	2 5 7	9 12 14	16 18 21
37	.6018	.6041	.6065	.6088	.6111	.6134	.6157	52	2 5 7	9 12 14	16 18 21
38	.6157	.6180	.6202	.6225	.6248	.6271	.6293	51	2 5 7	9 12 14	16 18 21
39	.6293	.6316	.6338	.6361	.6383	.6406	.6428	50	2 4 7	9 11 13	15 18 20
40	.6428	.6450	.6472	.6494	.6517	.6539	.6561	49	2 4 7	9 11 13	15 18 20
41	.6561	.6583	.6604	.6626	.6648	.6670	.6691	48	2 4 7	9 11 13	15 18 20
42	.6691	.6713	.6734	.6756	.6777	.6799	.6820	47	2 4 6	8 11 13	15 17 19
43	.6820	.6841	.6862	.6884	.6905	.6926	.6947	46	2 4 6	8 11 13	15 17 19
44	.6947	.6967	.6988	.7009	.7030	.7050	.7071	45	2 4 6	8 11 13	15 17 19
°		50′	40′	30′	20′	10′	0′	°	1 2 3	4 5 6	7 8 9
Natural Cosines 45°–90°									Proportional Parts (Subtract)		

Natural Sines 45°–90°									Proportional Parts (Add)								
°	0′	10′	20′	30′	40′	50′		°	1	2	3	4	5	6	7	8	9
45	.7071	.7092	.7112	.7133	.7153	.7173	.7193	44	2	4	6	8	10	12	14	16	18
46	.7193	.7214	.7234	.7254	.7274	.7294	.7314	43	2	4	6	8	10	12	14	16	18
47	.7314	.7333	.7353	.7373	.7392	.7412	.7431	42	2	4	6	8	10	12	14	15	18
48	.7431	.7451	.7470	.7490	.7509	.7528	.7547	41	2	4	6	8	10	11	13	15	17
49	.7547	.7566	.7585	.7604	.7623	.7642	.7660	40	2	4	6	8	9	11	13	15	17
50	.7660	.7679	.7698	.7716	.7735	.7753	.7771	39	2	4	6	7	9	11	13	15	17
51	.7771	.7790	.7808	.7826	.7844	.7862	.7880	38	2	4	5	7	9	11	13	15	16
52	.7880	.7898	.7916	.7934	.7951	.7969	.7986	37	2	3	5	7	9	10	12	14	16
53	.7986	.8004	.8021	.8039	.8056	.8073	.8090	36	2	3	5	7	9	10	12	14	16
54	.8090	.8107	.8124	.8141	.8158	.8175	.8192	35	2	3	5	7	8	10	12	14	15
55	.8192	.8208	.8225	.8241	.8258	.8274	.8290	34	2	3	5	7	8	10	11	13	15
56	.8290	.8307	.8323	.8339	.8355	.8371	.8387	33	2	3	5	6	8	10	11	13	15
57	.8387	.8403	.8418	.8434	.8450	.8465	.8480	32	2	3	5	6	8	9	11	12	14
58	.8480	.8496	.8511	.8526	.8542	.8557	.8572	31	2	3	5	6	8	9	11	12	14
59	.8572	.8587	.8601	.8616	.8631	.8646	.8660	30	1	3	4	6	7	9	10	12	13
60	.8660	.8675	.8689	.8704	.8718	.8732	.8746	29	1	3	4	6	7	9	10	11	13
61	.8746	.8760	.8774	.8788	.8802	.8816	.8829	28	1	3	4	6	7	8	10	11	12
62	.8829	.8843	.8857	.8870	.8884	.8897	.8910	27	1	3	4	5	7	8	9	11	12
63	.8910	.8923	.8936	.8949	.8962	.8975	.8988	26	1	3	4	5	6	8	9	10	12
64	.8988	.9001	.9013	.9026	.9038	.9051	.9063	25	1	3	4	5	6	7	9	10	11
65	.9063	.9075	.9088	.9100	.9112	.9124	.9135	24	1	2	4	5	6	7	9	10	11
66	.9135	.9147	.9159	.9171	.9182	.9194	.9205	23	1	2	4	5	6	7	8	9	11
67	.9205	.9216	.9228	.9239	.9250	.9261	.9272	22	1	2	3	4	6	7	8	9	10
68	.9272	.9283	.9293	.9304	.9315	.9325	.9336	21	1	2	3	4	5	6	7	9	10
69	.9336	.9346	.9356	.9367	.9377	.9387	.9397	20	1	2	3	4	5	6	7	8	9
70	.9397	.9407	.9417	.9426	.9436	.9446	.9455	19	1	2	3	4	5	6	7	8	9
71	.9455	.9465	.9474	.9483	.9492	.9502	.9511	18	1	2	3	4	5	6	7	7	8
72	.9511	.9520	.9528	.9537	.9546	.9555	.9563	17	1	2	3	3	4	5	6	7	8
73	.9563	.9572	.9580	.9588	.9596	.9605	.9613	16	1	2	2	3	4	5	6	7	7
74	.9613	.9621	.9628	.9636	.9644	.9652	.9659	15	1	2	2	3	4	5	5	6	7
75	.9659	.9667	.9674	.9681	.9689	.9696	.9703	14	1	1	2	3	4	4	5	6	7
76	.9703	.9710	.9717	.9724	.9730	.9737	.9744	13	1	1	2	3	4	4	5	6	6
77	.9744	.9750	.9757	.9763	.9769	.9775	.9781	12	1	1	2	3	3	4	4	5	6
78	.9781	.9787	.9793	.9799	.9805	.9811	.9816	11	1	1	2	2	3	4	4	5	5
79	.9816	.9822	.9827	.9833	.9838	.9843	.9848	10	1	1	2	2	3	3	4	4	5
80	.9848	.9853	.9858	.9863	.9868	.9872	.9877	9	0	1	1	2	2	3	3	4	4
81	.9877	.9881	.9886	.9890	.9894	.9899	.9903	8	0	1	1	2	2	3	3	4	4
82	.9903	.9907	.9911	.9914	.9918	.9922	.9925	7	0	1	1	1	2	2	3	3	3
83	.9925	.9929	.9932	.9936	.9939	.9942	.9945	6	0	1	1	1	2	2	2	3	3
84	.9945	.9948	.9951	.9954	.9957	.9959	.9962	5	0	1	1	1	1	2	2	2	3
85	.9962	.9964	.9967	.9969	.9971	.9974	.9976	4	0	1	1	1	1	2	2	2	2
86	.9976	.9978	.9980	.9981	.9983	.9985	.9986	3	0	0	1	1	1	1	2	2	2
87	.9986	.9988	.9989	.9990	.9992	.9993	.9994	2	0	0	1	1	1	1	1	1	2
88	.9994	.9995	.9996	.9997	.9997	.9998	.9998	1	0	0	0	0	0	1	1	1	1
89	.9998	.9999	.9999	1.0000	1.0000	1.0000	1.0000	0	0	0	0	0	0	1	1	1	1
°		50′	40′	30′	20′	10′	0′	°	1	2	3	4	5	6	7	8	9
Natural Cosines 0°–45°									Proportional Parts (Subtract)								

Logarithmic Sines 0°–45°									Proportional Parts (Add)		
°	0′	10′	20′	30′	40′	50′		°	1 2 3	4 5 6	7 8 9
0	$-\infty$	$\bar{3}$.4637	.7648	.9408	*.0658*	*.1627*	*.2419*	89	Differences	3011–792	
1	$\bar{2}$.2419	.3088	.3668	.4179	.4637	.5050	.5428	88		669–378	
2	.5428	.5776	.6097	.6397	.6677	.6940	.7188	87		348–248	
3	.7188	.7423	.7645	.7857	.8059	.8251	.8436	86		235–185	
4	.8436	.8613	.8783	.8946	.9104	.9256	.9403	85		177–147	
5	$\bar{2}$.9403	.9545	.9682	.9816	.9945	*.0070*	*.0192*	84		142–122	
6	$\bar{1}$.0192	.0311	.0426	.0539	.0648	.0755	.0859	83		119–104	
7	.0859	.0961	.1060	.1157	.1252	.1345	.1436	82		102–91	
8	.1436	.1525	.1612	.1697	.1781	.1863	.1943	81		89–80	
9	.1943	.2022	.2100	.2176	.2251	.2324	.2397	80		79–73	
10	$\bar{1}$.2397	.2468	.2538	.2606	.2674	.2740	.2806	79	7 14 20	27 34 41	48 54 61
11	.2806	.2870	.2934	.2997	.3058	.3119	.3179	78	6 12 19	25 31 37	43 50 56
12	.3179	.3238	.3296	.3353	.3410	.3466	.3521	77	6 11 17	23 29 34	40 46 51
13	.3521	.3575	.3629	.3682	.3734	.3786	.3837	76	5 11 16	21 27 32	37 42 48
14	.3837	.3887	.3937	.3986	.4035	.4083	.4130	75	5 10 15	20 25 29	34 39 44
15	$\bar{1}$.4130	.4177	.4223	.4269	.4314	.4359	.4403	74	5 9 14	18 23 27	32 37 41
16	.4403	.4447	.4491	.4533	.4576	.4618	.4659	73	4 9 13	17 21 26	30 34 39
17	.4659	.4700	.4741	.4781	.4821	.4861	.4900	72	4 8 12	16 20 24	28 32 36
18	.4900	.4939	.4977	.5015	.5052	.5090	.5126	71	4 8 11	15 19 23	27 30 34
19	.5126	.5163	.5199	.5235	.5270	.5306	.5341	70	4 7 10	14 18 22	25 29 32
20	$\bar{1}$.5341	.5375	.5409	.5443	.5477	.5510	.5543	69	3 7 10	14 17 20	24 27 31
21	.5543	.5576	.5609	.5641	.5673	.5704	.5736	68	3 6 10	13 16 19	22 26 29
22	.5736	.5767	.5798	.5828	.5859	.5889	.5919	67	3 6 9	12 15 18	21 24 28
23	.5919	.5948	.5978	.6007	.6036	.6065	.6093	66	3 6 9	12 15 17	20 23 26
24	.6093	.6121	.6149	.6177	.6205	.6232	.6259	65	3 6 8	11 14 16	19 22 25
25	$\bar{1}$.6259	.6286	.6313	.6340	.6366	.6392	.6418	64	3 5 8	11 13 16	19 21 24
26	.6418	.6444	.6470	.6495	.6521	.6546	.6570	63	3 5 8	10 13 15	18 20 23
27	.6570	.6595	.6620	.6644	.6668	.6692	.6716	62	2 5 7	10 12 15	17 19 22
28	.6716	.6740	.6763	.6787	.6810	.6833	.6856	61	2 5 7	9 12 14	16 19 21
29	.6856	.6878	.6901	.6923	.6946	.6968	.6990	60	2 4 7	9 11 13	16 18 20
30	$\bar{1}$.6990	.7012	.7033	.7055	.7076	.7097	.7118	59	2 4 6	9 11 13	15 17 19
31	.7118	.7139	.7160	.7181	.7201	.7222	.7242	58	2 4 6	8 10 12	14 16 19
32	.7242	.7262	.7282	.7302	.7322	.7342	.7361	57	2 4 6	8 10 12	14 16 18
33	.7361	.7380	.7400	.7419	.7438	.7457	.7476	56	2 4 6	8 10 11	13 15 17
34	.7476	.7494	.7513	.7531	.7550	.7568	.7586	55	2 4 6	7 9 11	13 15 16
35	$\bar{1}$.7586	.7604	.7622	.7640	.7657	.7675	.7692	54	2 4 5	7 9 11	12 14 16
36	.7692	.7710	.7727	.7744	.7761	.7778	.7795	53	2 3 5	7 9 10	12 14 15
37	.7795	.7811	.7828	.7844	.7861	.7877	.7893	52	2 3 5	7 8 10	11 13 15
38	.7893	.7910	.7926	.7941	.7957	.7973	.7989	51	2 3 5	6 8 10	11 13 14
39	.7989	.8004	.8020	.8035	.8050	.8066	.8081	50	2 3 5	6 8 9	11 12 14
40	$\bar{1}$.8081	.8096	.8111	.8125	.8140	.8155	.8169	49	1 3 4	6 7 9	10 12 13
41	.8169	.8184	.8198	.8213	.8227	.8241	.8255	48	1 3 4	6 7 9	10 11 13
42	.8255	.8269	.8283	.8297	.8311	.8324	.8338	47	1 3 4	6 7 8	10 11 13
43	.8338	.8351	.8365	.8378	.8391	.8405	.8418	46	1 3 4	5 7 8	9 11 12
44	.8418	.8431	.8444	.8457	.8469	.8482	.8495	45	1 3 4	5 6 8	9 10 12
°		50′	40′	30′	20′	10′	0′	°	1 2 3	4 5 6	7 8 9
Logarithmic Cosines 45°–90°									Proportional Parts (Subtract)		

Logarithmic Sines 45°–90°									Proportional Parts (Add)		
°	0′	10′	20′	30′	40′	50′		°	1 2 3	4 5 6	7 8 9
45	$\bar{1}$.8495	.8507	.8520	.8532	.8545	.8557	.8569	44	1 2 4	5 6 7	9 10 11
46	.8569	.8582	.8594	.8606	.8618	.8629	.8641	43	1 2 4	5 6 7	8 10 11
47	.8641	.8653	.8665	.8676	.8688	.8699	.8711	42	1 2 4	5 6 7	8 9 11
48	.8711	.8722	.8733	.8745	.8756	.8767	.8778	41	1 2 3	4 6 7	8 9 10
49	.8778	.8789	.8800	.8810	.8821	.8832	.8843	40	1 2 3	4 5 7	8 9 10
50	$\bar{1}$.8843	.8853	.8864	.8874	.8884	.8895	.8905	39	1 2 3	4 5 6	7 9 10
51	.8905	.8915	.8925	.8935	.8945	.8955	.8965	38	1 2 3	4 5 6	7 8 9
52	.8965	.8975	.8985	.8995	.9004	.9014	.9023	37	1 2 3	4 5 6	7 8 9
53	.9023	.9033	.9042	.9052	.9061	.9070	.9080	36	1 2 3	4 5 6	7 8 9
54	.9080	.9089	.9098	.9107	.9116	.9125	.9134	35	1 2 3	4 4 5	6 7 8
55	$\bar{1}$.9134	.9142	.9151	.9160	.9169	.9177	.9186	34	1 2 3	3 4 5	6 7 8
56	.9186	.9194	.9203	.9211	.9219	.9228	.9236	33	1 2 2	3 4 5	6 7 7
57	.9236	.9244	.9252	.9260	.9268	.9276	.9284	32	1 2 2	3 4 5	6 6 7
58	.9284	.9292	.9300	.9308	.9315	.9323	.9331	31	1 2 2	3 4 5	6 6 7
59	.9331	.9338	.9346	.9353	.9361	.9368	.9375	30	1 2 2	3 4 5	5 6 7
60	$\bar{1}$.9375	.9383	.9390	.9397	.9404	.9411	.9418	29	1 1 2	3 4 4	5 6 6
61	.9418	.9425	.9432	.9439	.9446	.9453	.9459	28	1 1 2	3 3 4	5 5 6
62	.9459	.9466	.9473	.9479	.9486	.9492	.9499	27	1 1 2	3 3 4	5 5 6
63	.9499	.9505	.9512	.9518	.9524	.9530	.9537	26	1 1 2	3 3 4	4 5 6
64	.9537	.9543	.9549	.9555	.9561	.9567	.9573	25	1 1 2	2 3 4	4 5 5
65	$\bar{1}$.9573	.9579	.9584	.9590	.9596	.9602	.9607	24	1 1 2	2 3 3	4 5 5
66	.9607	.9613	.9618	.9624	.9629	.9635	.9640	23	1 1 2	2 3 3	4 4 5
67	.9640	.9646	.9651	.9656	.9661	.9667	.9672	22	1 1 2	2 3 3	4 4 5
68	.9672	.9677	.9682	.9687	.9692	.9697	.9702	21	1 1 2	2 3 3	4 4 5
69	.9702	.9706	.9711	.9716	.9721	.9725	.9730	20	0 1 1	2 2 3	3 4 4
70	$\bar{1}$.9730	.9734	.9739	.9743	.9748	.9752	.9757	19	0 1 1	2 2 3	3 4 4
71	.9757	.9761	.9765	.9770	.9774	.9778	.9782	18	0 1 1	2 2 2	3 3 4
72	.9782	.9786	.9790	.9794	.9798	.9802	.9806	17	0 1 1	2 2 2	3 3 4
73	.9806	.9810	.9814	.9817	.9821	.9825	.9828	16	0 1 1	1 2 2	3 3 3
74	.9828	.9832	.9836	.9839	.9843	.9846	.9849	15	0 1 1	1 2 2	2 3 3
75	$\bar{1}$.9849	.9853	.9856	.9859	.9863	.9866	.9869	14	0 1 1	1 2 2	2 3 3
76	.9869	.9872	.9875	.9878	.9881	.9884	.9887	13	0 1 1	1 2 2	2 2 3
77	.9887	.9890	.9893	.9896	.9899	.9901	.9904	12	0 1 1	1 1 2	2 2 3
78	.9904	.9907	.9909	.9912	.9914	.9917	.9919	11	0 1 1	1 1 2	2 2 3
79	.9919	.9922	.9924	.9927	.9929	.9931	.9934	10	0 1 1	1 1 2	2 2 3
80	$\bar{1}$.9934	.9936	.9938	.9940	.9942	.9944	.9946	9	Differences	2	
81	.9946	.9948	.9950	.9952	.9954	.9956	.9958	8		2	
82	.9958	.9959	.9961	.9963	.9964	.9966	.9968	7		1–2	
83	.9968	.9969	.9971	.9972	.9973	.9975	.9976	6		1–2	
84	.9976	.9977	.9979	.9980	.9981	.9982	.9983	5		1–2	
85	$\bar{1}$.9983	.9985	.9986	.9987	.9988	.9989	.9989	4		0–2	
86	.9989	.9990	.9991	.9992	.9993	.9993	.9994	3		0–1	
87	.9994	.9995	.9995	.9996	.9996	.9997	.9997	2		0–1	
88	.9997	.9998	.9998	.9999	.9999	.9999	.9999	1		0–1	
89	.9999	.0000	.0000	.0000	.0000	.0000	.0000	0		0–1	
°		50′	40′	30′	20′	10′	0′	°	1 2 3	4 5 6	7 8 9
Logarithmic Cosines 0°–45°									Proportional Parts (Subtract)		

NATURAL TANGENTS 0°–45° | **PROPORTIONAL PARTS (Add)**

°	0′	10′	20′	30′	40′	50′		°	1	2	3	4	5	6	7	8	9
0	.0000	.0029	.0058	.0087	.0116	.0145	.0175	89	3	6	9	12	15	17	20	23	26
1	.0175	.0204	.0233	.0262	.0291	.0320	.0349	88	3	6	9	12	15	17	20	23	26
2	.0349	.0378	.0407	.0437	.0466	.0495	.0524	87	3	6	9	12	15	17	20	23	26
3	.0524	.0553	.0582	.0612	.0641	.0670	.0699	86	3	6	9	12	15	17	20	23	26
4	.0699	.0729	.0758	.0787	.0816	.0846	.0875	85	3	6	9	12	15	18	21	24	26
5	.0875	.0904	.0934	.0963	.0992	.1022	.1051	84	3	6	9	12	15	18	21	24	26
6	.1051	.1080	.1110	.1139	.1169	.1198	.1228	83	3	6	9	12	15	18	21	24	27
7	.1228	.1257	.1287	.1317	.1346	.1376	.1405	82	3	6	9	12	15	18	21	24	27
8	.1405	.1435	.1465	.1495	.1524	.1554	.1584	81	3	6	9	12	15	18	21	24	27
9	.1584	.1614	.1644	.1673	.1703	.1733	.1763	80	3	6	9	12	15	18	21	24	27
10	.1763	.1793	.1823	.1853	.1883	.1914	.1944	79	3	6	9	12	15	18	21	24	27
11	.1944	.1974	.2004	.2035	.2065	.2095	.2126	78	3	6	9	12	15	18	21	24	27
12	.2126	.2156	.2186	.2217	.2247	.2278	.2309	77	3	6	9	12	15	18	22	25	28
13	.2309	.2339	.2370	.2401	.2432	.2462	.2493	76	3	6	9	12	15	18	22	25	28
14	.2493	.2524	.2555	.2586	.2617	.2648	.2679	75	3	6	9	12	16	19	22	25	28
15	.2679	.2711	.2742	.2773	.2805	.2836	.2867	74	3	6	9	12	16	19	22	25	28
16	.2867	.2899	.2931	.2962	.2994	.3026	.3057	73	3	6	10	13	16	19	22	26	29
17	.3057	.3089	.3121	.3153	.3185	.3217	.3249	72	3	6	10	13	16	19	22	26	29
18	.3249	.3281	.3314	.3346	.3378	.3411	.3443	71	3	6	10	13	16	19	23	26	29
19	.3443	.3476	.3508	.3541	.3574	.3607	.3640	70	3	7	10	13	16	20	23	26	30
20	.3640	.3673	.3706	.3739	.3772	.3805	.3839	69	3	7	10	13	17	20	23	26	30
21	.3839	.3872	.3906	.3939	.3973	.4006	.4040	68	3	7	10	13	17	20	23	27	30
22	.4040	.4074	.4108	.4142	.4176	.4210	.4245	67	3	7	10	14	17	21	24	27	31
23	.4245	.4279	.4314	.4348	.4383	.4417	.4452	66	3	7	10	14	17	21	24	28	31
24	.4452	.4487	.4522	.4557	.4592	.4628	.4663	65	4	7	11	14	18	21	25	28	32
25	.4663	.4699	.4734	.4770	.4806	.4841	.4877	64	4	7	11	14	18	21	25	29	32
26	.4877	.4913	.4950	.4986	.5022	.5059	.5095	63	4	7	11	15	18	22	25	29	33
27	.5095	.5132	.5169	.5206	.5243	.5280	.5317	62	4	7	11	15	19	22	26	30	33
28	.5317	.5354	.5392	.5430	.5467	.5505	.5543	61	4	8	11	15	19	23	26	30	34
29	.5543	.5581	.5619	.5658	.5696	.5735	.5774	60	4	8	12	15	19	23	27	31	35
30	.5774	.5812	.5851	.5890	.5930	.5969	.6009	59	4	8	12	16	20	23	27	31	35
31	.6009	.6048	.6088	.6128	.6168	.6208	.6249	58	4	8	12	16	20	24	28	32	36
32	.6249	.6289	.6330	.6371	.6412	.6453	.6494	57	4	8	12	16	20	24	28	33	37
33	.6494	.6536	.6577	.6619	.6661	.6703	.6745	56	4	8	13	17	21	25	29	34	38
34	.6745	.6787	.6830	.6873	.6916	.6959	.7002	55	4	9	13	17	21	26	30	34	39
35	.7002	.7046	.7089	.7133	.7177	.7221	.7265	54	4	9	13	18	22	26	31	35	40
36	.7265	.7310	.7355	.7400	.7445	.7490	.7536	53	5	9	14	18	23	27	32	36	41
37	.7536	.7581	.7627	.7673	.7720	.7766	.7813	52	5	9	14	18	23	28	32	37	42
38	.7813	.7860	.7907	.7954	.8002	.8050	.8098	51	5	10	14	19	24	29	33	38	43
39	.8098	.8146	.8195	.8243	.8292	.8342	.8391	50	5	10	15	20	24	29	34	39	44
40	.8391	.8441	.8491	.8541	.8591	.8642	.8693	49	5	10	15	20	25	30	35	40	45
41	.8693	.8744	.8796	.8847	.8899	.8952	.9004	48	5	10	16	21	26	31	36	42	47
42	.9004	.9057	.9110	.9163	.9217	.9271	.9325	47	5	11	16	21	27	32	37	43	48
43	.9325	.9380	.9435	.9490	.9545	.9601	.9657	46	6	11	17	22	28	33	39	44	50
44	.9657	.9713	.9770	.9827	.9884	.9942	1.0000	45	6	11	17	23	29	34	40	46	51
°		50′	40′	30′	20′	10′	0′	°	1	2	3	4	5	6	7	8	9

NATURAL COTANGENTS 45°–90° | **PROPORTIONAL PARTS (Subtract)**

Natural Tangents 45°–90°								
°	0′	10′	20′	30′	40′	50′		°
		Dif.	Dif.	Dif.	Dif.	Dif.	Dif.	
45	1.0000	58 .0058	59 .0117	59 .0176	59 .0235	60 .0295	60 .0355	44
46	.0355	61 .0416	61 .0477	61 .0538	61 .0599	62 .0661	62 .0723	43
47	.0724	62 .0786	64 .0850	63 .0913	64 .0977	64 .1041	65 .1106	42
48	.1106	65 .1171	66 .1237	66 .1303	66 .1369	67 .1436	68 .1504	41
49	.1504	68 .1572	68 .1640	69 .1709	69 .1778	69 .1847	71 .1918	40
50	1.1918	70 .1988	71 .2059	72 .2131	72 .2203	73 .2276	73 .2349	39
51	.2349	74 .2423	74 .2497	75 .2572	75 .2647	76 .2723	76 .2799	38
52	.2799	77 .2876	78 .2954	78 .3032	79 .3111	79 .3190	80 .3270	37
53	.3270	81 .3351	81 .3432	82 .3514	83 .3597	83 .3680	84 .3764	36
54	.3764	84 .3848	85 .3933	86 .4019	87 .4106	87 .4193	88 .4281	35
55	1.4281	89 .4370	90 .4460	90 .4550	91 .4641	92 .4733	93 .4826	34
56	.4826	93 .4919	94 .5013	95 .5108	96 .5204	97 .5301	98 .5399	33
57	.5399	98 .5497	100 .5597	100 .5697	101 .5798	102 .5900	103 .6003	32
58	.6003	104 .6107	105 .6212	107 .6319	107 .6426	108 .6534	109 .6643	31
59	.6643	110 .6753	111 .6864	113 .6977	113 .7090	115 .7205	116 .7321	30
60	1.7321	116 .7437	119 .7556	119 .7675	121 .7796	121 .7917	123 .8040	29
61	.8040	125 .8165	126 .8291	127 .8418	128 .8546	130 .8676	131 .8807	28
62	.8807	133 .8940	134 .9074	136 .9210	137 .9347	139 .9486	140 .9626	27
63	.9626	142 .9768	144 .9912	145 *.0057*	147 *.0204*	149 *.0353*	150 *.0503*	26
64	2.0503	152 .0655	154 .0809	156 .0965	158 .1123	160 .1283	162 .1445	25
65	2.1445	164 .1609	166 .1775	168 .1943	170 .2113	173 .2286	174 .2460	24
66	.2460	177 .2637	180 .2817	181 .2998	185 .3183	186 .3369	190 .3559	23
67	.3559	191 .3750	195 .3945	197 .4142	200 .4342	203 .4545	206 .4751	22
68	.4751	209 .4960	212 .5172	214 .5386	219 .5605	221 .5826	225 .6051	21
69	.6051	228 .6279	232 .6511	235 .6746	239 .6985	243 .7228	247 .7475	20
70	2.7475	250 .7725	255 .7980	259 .8239	263 .8502	268 .8770	272 .9042	19
71	.9042	277 .9319	281 .9600	287 .9887	291 *.0178*	297 *.0475*	302 *.0777*	18
72	3.0777	307 .1084	313 .1397	319 .1716	325 .2041	330 .2371	338 .2709	17
73	.2709	343 .3052	350 .3402	357 .3759	365 .4124	371 .4495	379 .4874	16
74	.4874	387 .5261	395 .5656	403 .6059	411 .6470	421 .6891	430 .7321	15
75	3.7321	439 .7760	448 .8208	459 .8667	469 .9136	481 .9617	491 *.0108*	14
76	4.0108	503 .0611	515 .1126	527 .1653	540 .2193	554 .2747	568 .3315	13
77	.3315	582 .3897	597 .4494	613 .5107	629 .5736	646 .6382	664 .7046	12
78	.7046	683 .7729	701 .8430	722 .9152	742 .9894	764 *.0658*	788 *.1446*	11
79	5.1446	811 .2257	836 .3093	862 .3955	890 .4845	919 .5764	949 .6713	10
80	5.6713	5.7694	5.8708	5.9758	6.0844	6.1970	6.3138	9
81	6.3138	6.4348	6.5606	6.6912	6.8269	6.9682	7.1154	8
82	7.1154	7.2687	7.4287	7.5958	7.7704	7.9530	8.1443	7
83	8.1443	8.3450	8.5555	8.7769	9.0098	9.2553	9.5144	6
84	9.5144	9.7882	10.078	10.385	10.712	11.059	11.430	5
85	11.430	11.826	12.251	12.706	13.197	13.727	14.301	4
86	14.301	14.924	15.605	16.350	17.169	18.075	19.081	3
87	19.081	20.206	21.470	22.904	24.542	26.432	28.636	2
88	28.636	31.242	34.368	38.188	42.964	49.104	57.290	1
89	57.290	68.750	85.940	114.59	171.89	343.77	∞	0
°		50′	40′	30′	20′	10′	0′	°
Natural Cotangents 0°–45°								

Logarithmic Tangents 0°–45°									Proportional Parts (Add)								
°	0′	10′	20′	30′	40′	50′		°	1	2	3	4	5	6	7	8	9
0	$-\infty$	$\bar{3}$.4637	.7648	.9409	*.0658*	*.1627*	*.2419*	89				3011–792					
1	$\bar{2}$.2419	.3089	.3669	.4181	.4638	.5053	.5431	88				670–378					
2	.5431	.5779	.6101	.6401	.6682	.6945	.7194	87				348–249					
3	.7194	.7429	.7652	.7865	.8067	.8261	.8446	86				235–185					
4	.8446	.8624	.8795	.8960	.9118	.9272	.9420	85				170–148					
5	$\bar{2}$.9420	.9563	.9701	.9836	.9966	*.0093*	*.0216*	84	Differences			143–124					
6	$\bar{1}$.0216	.0336	.0453	.0567	.0678	.0786	.0891	83				120–105					
7	.0891	.0995	.1096	.1194	.1291	.1385	.1478	82				104–93					
8	.1478	.1569	.1658	.1745	.1831	.1915	.1997	81				91–82					
9	.1997	.2078	.2158	.2236	.2313	.2389	.2463	80				81–74					
10	$\bar{1}$.2463	.2536	.2609	.2680	.2750	.2819	.2887	79	7	14	21	28	36	43	50	57	64
11	.2887	.2954	.3020	.3085	.3149	.3212	.3275	78	7	13	20	26	33	39	46	52	59
12	.3275	.3336	.3397	.3458	.3517	.3576	.3634	77	6	12	18	24	30	36	42	48	54
13	.3634	.3691	.3748	.3804	.3859	.3914	.3968	76	6	11	17	22	28	33	39	44	50
14	.3968	.4021	.4074	.4127	.4178	.4230	.4281	75	5	10	16	21	26	31	36	42	47
15	$\bar{1}$.4281	.4331	.4381	.4430	.4479	.4527	.4575	74	5	10	15	20	25	29	34	39	44
16	.4575	.4622	.4669	.4716	.4762	.4808	.4853	73	5	9	14	19	24	28	33	38	42
17	.4853	.4898	.4943	.4987	.5031	.5075	.5118	72	4	9	13	18	22	27	31	35	40
18	.5118	.5161	.5203	.5245	.5287	.5329	.5370	71	4	8	13	17	21	25	29	34	38
19	.5370	.5411	.5451	.5491	.5531	.5571	.5611	70	4	8	12	16	20	24	28	32	36
20	$\bar{1}$.5611	.5650	.5689	.5727	.5766	.5804	.5842	69	4	8	12	15	19	23	27	31	35
21	.5842	.5879	.5917	.5954	.5991	.6028	.6064	68	4	7	11	15	18	22	26	30	33
22	.6064	.6100	.6136	.6172	.6208	.6243	.6279	67	4	7	11	14	18	22	25	29	32
23	.6279	.6314	.6348	.6383	.6417	.6452	.6486	66	3	7	10	14	17	21	24	28	31
24	.6486	.6520	.6553	.6587	.6620	.6654	.6687	65	3	7	10	13	17	20	23	27	30
25	$\bar{1}$.6687	.6720	.6752	.6785	.6817	.6850	.6882	64	3	7	10	13	16	20	23	26	29
26	.6882	.6914	.6946	.6977	.7009	.7040	.7072	63	3	6	10	13	16	19	22	26	29
27	.7072	.7103	.7134	.7165	.7196	.7226	.7257	62	3	6	9	12	15	18	21	25	28
28	.7257	.7287	.7317	.7348	.7378	.7408	.7438	61	3	6	9	12	15	18	21	24	27
29	.7438	.7467	.7497	.7526	.7556	.7585	.7614	60	3	6	9	12	15	18	21	23	26
30	$\bar{1}$.7614	.7644	.7673	.7701	.7730	.7759	.7788	59	3	6	9	12	14	17	20	23	26
31	.7788	.7816	.7845	.7873	.7902	.7930	.7958	58	3	6	9	11	14	17	20	23	25
32	.7958	.7986	.8014	.8042	.8070	.8097	.8125	57	3	6	8	11	14	17	20	22	25
33	.8125	.8153	.8180	.8208	.8235	.8263	.8290	56	3	6	8	11	14	17	19	22	25
34	.8290	.8317	.8344	.8371	.8398	.8425	.8452	55	3	5	8	11	13	16	19	22	24
35	$\bar{1}$.8452	.8479	.8506	.8533	.8559	.8586	.8613	54	3	5	8	11	13	16	19	21	24
36	.8613	.8639	.8666	.8692	.8718	.8745	.8771	53	3	5	8	11	13	16	18	21	24
37	.8771	.8797	.8824	.8850	.8876	.8902	.8928	52	3	5	8	10	13	16	18	21	24
38	.8928	.8954	.8980	.9006	.9032	.9058	.9084	51	3	5	8	10	13	16	18	21	23
39	.9084	.9110	.9135	.9161	.9187	.9212	.9238	50	3	5	8	10	13	15	18	21	23
40	$\bar{1}$.9238	.9264	.9289	.9315	.9341	.9366	.9392	49	3	5	8	10	13	15	18	21	23
41	.9392	.9417	.9443	.9468	.9494	.9519	.9544	48	3	5	8	10	13	15	18	20	23
42	.9544	.9570	.9595	.9621	.9646	.9671	.9697	47	3	5	8	10	13	15	18	20	23
43	.9697	.9722	.9747	.9772	.9798	.9823	.9848	46	3	5	8	10	13	15	18	20	23
44	.9848	.9874	.9899	.9924	.9949	.9975	.0000	45	3	5	8	10	13	15	18	20	23
°		50′	40′	30′	20′	10′	0′	°	1	2	3	4	5	6	7	8	9
Logarithmic Cotangents 45°–90°									Proportional Parts (Subtract)								

Logarithmic Tangents 45°–90°									Proportional Parts (Add)		
°	0′	10′	20′	30′	40′	50′		°	1 2 3	4 5 6	7 8 9
45	0.0000	.0025	.0051	.0076	.0101	.0126	.0152	44	3 5 8	10 13 15	18 20 23
46	.0152	.0177	.0202	.0228	.0253	.0278	.0303	43	3 5 8	10 13 15	18 20 23
47	.0303	.0329	.0354	.0379	.0405	.0430	.0456	42	3 5 8	10 13 15	18 20 23
48	.0456	.0481	.0506	.0532	.0557	.0583	.0608	41	3 5 8	10 13 15	18 20 23
49	.0608	.0634	.0659	.0685	.0711	.0736	.0762	40	3 5 8	10 13 15	18 20 23
50	0.0762	.0788	.0813	.0839	.0865	.0890	.0916	39	3 5 8	10 13 15	18 20 23
51	.0916	.0942	.0968	.0994	.1020	.1046	.1072	38	3 5 8	10 13 16	18 21 23
52	.1072	.1098	.1124	.1150	.1176	.1203	.1229	37	3 5 8	10 13 16	18 21 23
53	.1229	.1255	.1282	.1308	.1334	.1361	.1387	36	3 5 8	11 13 16	18 21 24
54	.1387	.1414	.1441	.1467	.1494	.1521	.1548	35	3 5 8	11 13 16	19 21 24
55	0.1548	.1575	.1602	.1629	.1656	.1683	.1710	34	3 5 8	11 14 16	19 22 24
56	.1710	.1737	.1765	.1792	.1820	.1847	.1875	33	3 6 8	11 14 17	19 22 25
57	.1875	.1903	.1930	.1958	.1986	.2014	.2042	32	3 6 8	11 14 17	20 22 25
58	.2042	.2070	.2098	.2127	.2155	.2184	.2212	31	3 6 9	11 14 17	20 23 25
59	.2212	.2241	.2270	.2299	.2327	.2356	.2386	30	3 6 9	12 15 17	20 23 26
60	0.2386	.2415	.2444	.2474	.2503	.2533	.2562	29	3 6 9	12 15 18	21 23 26
61	.2562	.2592	.2622	.2652	.2683	.2713	.2743	28	3 6 9	12 15 18	21 24 27
62	.2743	.2774	.2804	.2835	.2866	.2897	.2928	27	3 6 9	12 15 19	22 25 28
63	.2928	.2960	.2991	.3023	.3054	.3086	.3118	26	3 6 10	13 16 19	22 25 29
64	.3118	.3150	.3183	.3215	.3248	.3280	.3313	25	3 7 10	13 16 20	23 26 29
65	0.3313	.3346	.3380	.3413	.3447	.3480	.3514	24	3 7 10	13 17 20	23 27 30
66	.3514	.3548	.3583	.3617	.3652	.3686	.3721	23	3 7 10	14 17 21	24 28 31
67	.3721	.3757	.3792	.3828	.3864	.3900	.3936	22	4 7 11	14 18 22	25 29 32
68	.3936	.3972	.4009	.4046	.4083	.4121	.4158	21	4 7 11	15 19 22	26 30 33
69	.4158	.4196	.4234	.4273	.4311	.4350	.4389	20	4 8 12	15 19 23	27 31 35
70	0.4389	.4429	.4469	.4509	.4549	.4589	.4630	19	4 8 12	16 20 24	28 32 36
71	.4630	.4671	.4713	.4755	.4797	.4839	.4882	18	4 8 13	17 21 25	29 34 38
72	.4882	.4925	.4969	.5013	.5057	.5102	.5147	17	4 9 13	18 22 26	31 35 40
73	.5147	.5192	.5238	.5284	.5331	.5378	.5425	16	5 9 14	19 23 28	32 37 42
74	.5425	.5473	.5521	.5570	.5619	.5669	.5719	15	5 10 15	20 25 29	34 39 44
75	0.5719	.5770	.5822	.5873	.5926	.5979	.6032	14	5 10 16	21 26 31	37 42 47
76	.6032	.6086	.6141	.6196	.6252	.6309	.6366	13	6 11 17	22 28 33	39 45 50
77	.6366	.6424	.6483	.6542	.6603	.6664	.6725	12	6 12 18	24 30 36	42 48 54
78	.6725	.6788	.6851	.6915	.6981	.7047	.7114	11	7 13 20	26 33 39	46 52 59
79	.7113	.7181	.7250	.7320	.7391	.7464	.7537	10	7 14 21	28 35 42	49 57 64
80	0.7537	.7611	.7687	.7764	.7842	.7922	.8003	9	Differences	74–81	
81	.8003	.8085	.8169	.8255	.8342	.8431	.8522	8		82–91	
82	.8522	.8615	.8709	.8806	.8904	.9005	.9109	7		93–104	
83	.9109	.9214	.9322	.9433	.9547	.9664	.9784	6		105–120	
84	.9784	.9907	*.0034*	*.0164*	*.0299*	*.0437*	*.0580*	5		123–143	
85	1.0580	.0728	.0882	.1040	.1205	.1376	.1554	4		148–178	
86	.1554	.1739	.1933	.2135	.2348	.2571	.2806	3		185–235	
87	.2806	.3055	.3318	.3599	.3899	.4221	.4569	2		249–348	
88	.4569	.4947	.5362	.5819	.6331	.6911	.7581	1		378–670	
89	.7581	.8373	.9342	2.0591	.2352	.5363	∞	0		792–3011	
°		50′	40′	30′	20′	10′	0′	°	1 2 3	4 5 6	7 8 9
Logarithmic Cotangents 0°–45°									Proportional Parts (Subtract)		

	LOGARITHMS										PROPORTIONAL PARTS								
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	4	8	12	17	21	25	29	33	37
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	11	15	19	23	26	30	34
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	10	14	17	21	24	28	31
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	6	10	13	16	19	23	26	29
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	18	21	24	27
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	8	11	14	17	20	22	25
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	5	8	11	13	16	18	21	24
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2	5	7	10	12	15	17	20	22
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2	5	7	9	12	14	16	19	21
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2	4	7	9	11	13	16	18	20
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	2	4	6	8	10	12	14	15	17
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	8	9
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	8	9
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	5	6	7	8
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	4	5	6	7	8
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	1	2	3	3	4	5	6	7	8
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1	2	3	3	4	5	6	7	8
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	1	2	2	3	4	5	6	7	7
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1	2	2	3	4	5	6	6	7
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1	2	2	3	4	5	6	6	7
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

	Logarithms										Proportional Parts								
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1	2	2	3	4	5	5	6	7
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	1	2	2	3	4	5	5	6	7
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1	2	2	3	4	5	5	6	7
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	1	1	2	3	4	4	5	6	7
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	1	1	2	3	4	4	5	6	7
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	1	1	2	3	4	4	5	6	6
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1	1	2	3	4	4	5	6	6
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1	1	2	3	3	4	5	6	6
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	1	1	2	3	3	4	5	5	6
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	1	1	2	3	3	4	5	5	6
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	1	1	2	3	3	4	5	5	6
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1	1	2	3	3	4	5	5	6
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1	1	2	3	3	4	5	5	6
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	1	1	2	3	3	4	4	5	6
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	1	1	2	2	3	4	4	5	6
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	1	1	2	2	3	4	4	5	6
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1	1	2	2	3	4	4	5	5
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1	1	2	2	3	4	4	5	5
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1	1	2	2	3	4	4	5	5
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1	1	2	2	3	4	4	5	5
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1	1	2	2	3	3	4	5	5
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1	1	2	2	3	3	4	5	5
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	1	1	2	2	3	3	4	4	5
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	1	1	2	2	3	3	4	4	5
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1	1	2	2	3	3	4	4	5
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079	1	1	2	2	3	3	4	4	5
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1	1	2	2	3	3	4	4	5
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	1	1	2	2	3	3	4	4	5
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1	1	2	2	3	3	4	4	5
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1	1	2	2	3	3	4	4	5
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1	1	2	2	3	3	4	4	5
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	1	1	2	2	3	3	4	4	5
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440	0	1	1	2	2	3	3	4	4
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	0	1	1	2	2	3	3	4	4
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0	1	1	2	2	3	3	4	4
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0	1	1	2	2	3	3	4	4
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633	0	1	1	2	2	3	3	4	4
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0	1	1	2	2	3	3	4	4
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727	0	1	1	2	2	3	3	4	4
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773	0	1	1	2	2	3	3	4	4
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0	1	1	2	2	3	3	4	4
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	0	1	1	2	2	3	3	4	4
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0	1	1	2	2	3	3	4	4
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	0	1	1	2	2	3	3	4	4
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996	0	1	1	2	2	3	3	3	4
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

EXPONENTIAL TABLE

x	e^{-x}	*$\text{Log}_{10} e^{-x}$	x	e^{-x}	*$\text{Log}_{10} e^{-x}$	x	e^{-x}	*$\text{Log}_{10} e^{-x}$
.00	1.0000	0.0000	**1.7**	.1827	.2617	**4.8**	.00823	.9154
.05	.9512	.9783	**1.8**	.1653	.2183	**5.0**	$.0_26738$	.8285
.10	.9048	.9566	**1.9**	.1496	.1749	**5.2**	$.0_25517$	.7417
.15	.8607	.9349	**2.0**	.1353	.1313	**5.4**	$.0_24517$	.6548
.20	.8187	.9131	**2.1**	.1225	.0881	**5.6**	$.0_23697$	.5679
.25	.7788	.8914	**2.2**	.1108	.0445	**5.8**	$.0_23028$	.4811
.30	.7408	.8697	**2.3**	.1003	.0013	**6.0**	$.0_22479$	.3942
.35	.7047	.8480	**2.4**	.09072	.9577	**6.2**	$.0_22030$	.3074
.40	.6703	.8263	**2.5**	.08208	.9142	**6.4**	$.0_21662$	.2205
.45	.6376	.8046	**2.6**	.07427	.8708	**6.6**	$.0_21354$	.1317
.50	.6065	.7828	**2.7**	.06721	.8274	**6.8**	$.0_21114$	.0468
.55	.5770	.7612	**2.8**	.06081	.7840	**7.0**	$.0_39118$	.9599
.60	.5488	.7394	**2.9**	.05502	.7405	**7.2**	$.0_37466$	.8731
.65	.5221	.7178	**3.0**	.04979	.6971	**7.4**	$.0_36112$	.7862
.70	.4966	.6960	**3.1**	.04505	.6537	**7.6**	$.0_35005$	.6994
.75	.4724	.6743	**3.2**	.04076	.6102	**7.8**	$.0_34097$	.6125
.80	.4493	.6525	**3.3**	.03688	.5668	**8.0**	$.0_33354$	.5256
.85	.4274	.6308	**3.4**	.03337	.5334	**8.2**	$.0_32747$	.4388
.90	.4066	.6092	**3.5**	.03020	.4800	**8.4**	$.0_32249$	.3519
.95	.3867	.5874	**3.6**	.02732	.4365	**8.6**	$.0_31841$	.2651
1.00	.3679	.5657	**3.7**	.02472	.3931	**8.8**	$.0_31507$	.1782
1.10	.3329	.5223	**3.8**	.02237	.3497	**9.0**	$.0_31234$	.0913
1.20	.3012	.4789	**3.9**	.02024	.3062	**9.2**	$.0_31010$	.0045
1.30	.2725	.4354	**4.0**	.01832	.2629	**9.4**	$.0_48272$	.9176
1.40	.2466	.3920	**4.2**	.01500	.1761	**9.6**	$.0_46773$	.8308
1.50	.2231	.3485	**4.4**	.01228	.0892	**9.8**	$.0_45545$	.7439
1.60	.2019	.3051	**4.6**	.01005	.0022	**10.0**	$.0_44541$	.6571

* Mantissa only.

Index of Authors

Index of Subjects