TEXT-BOOK OF GENERAL PHYSICS

JOSEPH S. AMES

Physics.

TEXT-BOOK

OF

GENERAL PHYSICS

FOR

HIGH SCHOOLS AND COLLEGES

BY

JOSEPH S. AMES, Ph.D.

PROFESSOR OF PHYSICS AND DIRECTOR OF THE PHYSICAL LABORATORY
IN THE JOHNS HOPKINS UNIVERSITY

NEW YORK ·:· CINCINNATI ·:· CHICAGO
AMERICAN BOOK COMPANY

AMES'S TEXT-BOOK OF GEN. PHYSICS.

W. P.

PREFACE

When this book was begun it was with the intention of having it serve as a revision of the *Theory of Physics*, which was written seven years ago. It was soon discovered, however, that this plan was not possible; and so the new book was written without any reference to the former one. The ideas that guided the preparation of both are the same, however. I still believe that the most important element in a course of instruction in Physics is "a text-book which states the theory of the subject in a clear and logical manner so that recitations can be held on it"; and my attempt has been "to give a concise statement of the experimental facts on which the science of Physics is based, and to present with these statements the accepted theories which correlate or 'explain' them."

This text-book is divided into the following sections: Mechanics and Properties of Matter, Heat, Vibrations and Waves, Sound, Light, Magnetism and Electricity. In Mechanics special emphasis is placed upon the fundamental ideas of time, space, and force; and motion of rotation is given as much prominence as motion of translation. In the chapters on Heat, correct definitions of temperature and heat quantities are given, and the energy relations are discussed in full. Various types of vibrations and waves are next described and analyzed, and numerous extracts are given from the original memoirs of Huygens, Young, and Fresnel. In this section of the book the general phenomena of interference and diffraction are discussed. In Sound attention is called specially to the explanation of the characteristics of a musical note and of consonance. The phenomena of Light are considered as due to waves in the ether; and a prominent feature of the treatment is the stress placed upon the idea of resolving power. In the subjects of Magnetism and Electricity, the question of the measurement of the quantities is borne constantly in mind; and in the description of the varied phenomena no statements are made that are contrary to the accepted or probable theories as to the nature of electrical charges.

Numerous references are made in the course of each chapter and at the end of the sections to the historical development of Physics, and attention is called to a few books of reference. There is no experiment or observation, which has an important bearing on our knowledge or theories of Physics, that is not mentioned and explained; and the few great Principles of Nature are given the prominence that they deserve. Numerous illustrations accompany the text; and they have one and all been prepared with the express purpose of helping the student to form a clear idea of the subjects presented.

I wish to thank many friends for their kindness in offering suggestions as to text-books and modes of teaching Physics, and in particular Professor Crew of Northwestern University, Evanston, whose criticisms and advice have been most helpful.

J. S. AMES.

THE JOHNS HOPKINS UNIVERSITY,
BALTIMORE.

CONTENTS

GENERAL PHYSICS

INTRODUCTION

Matter and Physical Phenomena.—Through our various senses, such as those of sight, touch, and hearing, we are constantly receiving sensations which we interpret objectively; that is, we locate the cause of a sensation in a definite portion of space. We picture to ourselves the existence there of a something, which we call "matter"; and to a limited portion of space which contains matter we give the name "physical body." Thus, when one sees and feels a block of wood, lifts it, moves it, tries to compress or bend it, various sensations are perceived, which are all associated with the idea of matter.

Matter may be divided into two great classes: that which is living, such as plants and animals; and that which is not, such as pieces of wood and glass, water, and air. "Physics" is, broadly speaking, the science concerned with this second division of matter, which may be called "ordinary" matter; and phenomena occurring in connection with matter of this kind are called "physical phenomena." Many of these occur also in all forms of living matter; but associated with them there are other phenomena which have not as yet been proved to be physical.

Methods of Physics. — The scientific study of a subject involves three distinct ideas: the discovery, the investigation, and the explanation of phenomena. The first two require no discussion here; but what is meant by the last may not be at once evident. "To explain a phenomenon" is to determine its exact connection with other phenomena which are known, to describe it in terms of simpler ones, and in this manner to reduce the number of fundamental ideas as far as possible. Thus, the phenomenon of lightning is "explained" when it is shown that it is exactly similar to electrical sparks which can be produced by ordinary electrical machines; and the motions of the planets around the sun are "explained" when it is proved that they all obey certain mathematical relations known as Kepler's Laws, and that these are consequences of a generalization called Newton's Law of Gravitation — a statement involving only such ideas as force, inertia, and distance.

In seeking for explanations of phenomena, we assume, either directly or indirectly, that there is a definite connection between consecutive events, of such a nature that, if we are able to reproduce *exactly* a definite condition, the same effect will follow, regardless of the epoch of time or the location in space. Thus, if a stone is hanging from a string at a definite distance above the ground, and the string is cut, we can predict, from previous observations, exactly what will happen. We are justified in this belief as to the connection between "cause and effect" by all of our experiences and observations. If at any time there should be a deviation from what we might expect as the consequence of certain conditions, we should not regard it as an exception or variation, but should look immediately for the presence of some unnoted condition which might modify the result.

The first step in explaining a phenomenon is to make an hypothesis connecting it with other phenomena, and then to see by observation and experiment whether this hypothesis is satisfactory. To be so, it must be possible; it must be in harmony with the phenomena in question; it should be simple and yet include as many phenomena as possible; and it should admit of verification, or disproof, by direct observation or experiment. Thus the wave theory of light was an hypothesis advanced by Thomas Young to explain among other things the phenomena of the colors of soap bubbles; and all the consequences of this theory have been shown to be in accord with experimental observations. It has therefore in a sense ceased to be an hypothesis or a theory, and is a statement of fact, so far as we have any means of judging.

The Ether. — A careful study of the phenomena of light led philosophers many years ago to the belief that there is present in space another medium for phenomena than that furnished by ordinary matter. As we shall see later, there is every reason for believing that throughout the vast regions of space in the solar system, and beyond, ordinary matter is absent, except in certain cases as thin atmospheres around the planets and stars; it is also known that the sensation which we call "light" is due to the fact that *waves* enter the eye; therefore, when one sees a star, one has evidence that waves have left the star and have crossed space void of ordinary matter. There must be present, then, in this space a medium that can carry waves: it is called "the ether" (sometimes "æther"), and its properties will be discussed more fully in what follows. Similarly, in order to explain many electrical and magnetic phenomena, it was necessary to assume the existence of a medium different from ordinary matter, and one of the great achievements of the last century was the proof that this medium was the ether. There is, moreover, at present no evidence of the existence in our universe of any other media than ordinary matter and the ether.

Physics and its Branches. — The object of Physics may therefore be defined to be the attempt to determine the exact connections between phenomena, both in ordinary matter and in the ether, and to express these relations with as few hypotheses as possible concerning the nature and properties of matter and the ether.

From time to time certain branches of Physics, which deal with special groups of phenomena and are therefore to a greater or less degree independent, have been separated from the main subject; and out of them distinct sciences have been formed: such are Astronomy, Chemistry, Geology, Mineralogy, etc. As these, in turn, have grown, they have in many cases again approached the parent subject; and such specialized divisions of science as Astrophysics, Physical Chemistry, Geological Physics, etc., are now recognized. Again, many practical commercial applications of physical principles have been made; and their importance has led to the formation of distinct subjects known as the "engineering sciences"; such are Electrical, Mechanical, Steam, Hydraulic, etc., Engineering.

Physics, itself, has been subdivided in many ways, according to the point of view of the writer: Physics of Matter and Physics of the Ether; Mathematical Physics and Experimental Physics; etc. But none of these are satisfactory, for it is impossible to make any such sharply defined divisions. For purely practical purposes, however, some method of subdivision is advisable; and the one followed in this book is to consider as distinct subjects: Mechanics, Properties of Matter, Heat, Sound, Light, Magnetism and Electricity. It should not be thought, however, that these branches are independent of one another; for, on the contrary, they are most intimately connected, and it is impossible to discuss any one by itself. Each year new connections are being established.

Physical Quantities. — A physical quantity is something that we can imagine as capable of change in amount, something to which we can assign a numerical value. Some quantities can be *measured;* others cannot. To measure a quantity, another similar one must first be chosen as a standard or unit; and then the number of times this is contained in the original quantity is its measure. Thus, a length can be measured in terms of an inch, a yard, a mile, a centimetre, etc., depending upon the choice of unit; an electric current can be measured; the power of a steam engine can be measured; etc. It is possible to understand the meaning of a zero value of any measurable quantity; further, two or more measurable quantities of the same kind, *e.g.* two lengths, can be added. On the other hand, many quantities cannot be measured, although it is possible to give them numerical values. Thus, as we shall see, the temperature of a body cannot be measured in the sense described above, although it may have a definite value upon a definite scale, *e.g.* the scale of Fahrenheit. Similarly, the pitch of a musical note may have assigned to it a definite symbol or number; but it cannot be measured: there is no unit of pitch. (We can, of course, count the number of vibrations of the sounding body that causes the musical note; but this is not measuring the shrillness of the sound.) Strictly speaking, intervals of time cannot be measured, as will be shown shortly.

Fundamental Physical Quantities. — To most physical quantities exact definitions can be given; but there are a few for which this is impossible. Our knowledge of them is intuitive; there are no simpler ideas in terms of which we can describe them. The question as to the exact number of these belongs to the province of Mental Philosophy, and is not yet definitely established; but those of main importance in Physics are the following:

1. *The fundamental ideas in regard to space*, viz.: a straight line, a polygon, a solid figure bounded by plane faces. (A curved line admits of definition.) It is impossible to define what is meant by "length"; and the idea of two equal lengths admits of no ambiguity; neither does that of two equal areas; nor that of two equal volumes. We can choose a unit length, construct from it a square as a unit area and a cube as a unit volume, and then measure lengths, areas, and volumes, assuming the possibility of superposition.

2. *The idea of time*, or rather *duration of time.* — We have a definite conception of what is meant by two equal intervals of time; and yet we cannot express it in words. Certain physical phenomena appear to us to be periodic, that is to repeat themselves at exactly equal intervals of time, *e.g.* the vibrations of a pendulum or of the balance wheel of a watch; but we cannot prove this. There is every reason, however, for believing that these motions are exactly periodic, for at any instant the external conditions affecting the motion are exactly the same as they were at an interval of time before, so far as we can tell; and, further, if we compare a great many motions that satisfy our mental requirements as being periodic, they are always found to be periodic with reference to one another.

In order to give a number to an instant of time, one must choose some periodic motion such as described above, *e.g.* a certain pendulum vibrating under definite conditions, and some arbitrary epoch of time from which to count the number of vibrations of the vibrating instrument: the number of these vibrations between the epoch and the instant for which a number is desired is this number. (It is evident that we are not *measuring* a quantity, for we are not determining how many times one quantity is contained in another.)

This idea of equal intervals of time, combined with that of equal lengths along a straight line, makes up the conception of uniform velocity, which is therefore intuitive.

3. *The idea of effort as perceived by our muscular sense.* — It is by this sense that we obtain our fundamental ideas of the properties of matter. If a heavy body is held in the hand, or carried on the back, we are conscious of a definite sensation, which cannot be described or defined. It is called "force" sensation; or, we say we "feel a force." The sensation is said to be due in this case to the "weight" of the body; or, in other words, "weight" is that property of matter of which we become conscious through our muscular sense when we

support a body free from the earth. Similarly, if a spring is kept stretched by the hands, if a bow is maintained in a bent condition, etc., we are conscious of the same force sensation. This is said to be due to the "elasticity" of the spring, the bow, etc. Again, if we throw or catch a ball, if we set a barrel rolling or stop it, or, in general, if we alter the motion of any body, we feel a force. It is said to be due to the "inertia" of matter; or, "inertia" is that property of matter of which we become conscious through our muscular sense when we change the motion of a body. These properties of matter, viz., weight, elasticity and inertia, owing to which we feel forces, will be discussed later in detail.

Force. — We may express these facts in regard to force sensation differently. By means of our muscles we can support a body free from the earth, change its shape or size, or alter its motion; and if we do so, we are conscious of a definite sensation. The same conditions can be brought about by other means than our muscles. Thus, a body can be held free from the earth by resting on a flexible board, which it will therefore bend, or by hanging from a spring, which it will therefore stretch. A compressed spring, if allowed to relax, may be made to give speed to a bullet; a bent bow, if loosed, gives speed to an arrow. If a moving body strikes another, the motion of both is changed; etc. For simplicity, in describing these various phenomena, we use the word "force," which is borrowed from everyday language. We say, when a heavy body hangs from the stretched spring, that the latter "exerts a force" on the hanging body, and that the body "exerts a force" on the spring; further, we say that this last action is due to the force on the body exerted by the earth — in other words, "weight" is a force due to the earth's action. Similarly in the other cases, a force is said to act on the body which undergoes the change. We speak of forces due to elasticity, to impact, etc. This does not imply the *existence* of anything, but is merely a *description of a condition;* we mean that these phenomena are associated in our minds with the idea of effort.

Through our muscular sense we are conscious of differences in the intensity of the force sensation, and so speak of large and small forces, although our sensations cannot help us to measure them. We must define "force" in terms of physical quantities before we can do this. For purposes of definition any one of the three types of force effects may be chosen as the fundamental idea. The accepted mode of definition and measurement will be discussed later. See Chapter II.

Properties of Ordinary Matter. — Our fundamental ideas in regard to matter are obtained through our muscular sense; and they have been referred to already in speaking of the conception of force. They will now be described more in detail.

1. *Gravity.* — If a body is kept from falling toward the earth by holding it in the hand, a force is felt; and if it is held by a cord or a spring, the latter exerts a force on it sufficient to neutralize the action of the earth: the body is said to have "weight." We distinguish between heavy and light bodies; and the method of assigning proper numerical values will be described later. (See page 61.) Similarly, there is, so far as we know, a force of attraction between any two portions of matter, however large or small and at whatever distance apart. We mean by this that, unless prevented, the two bodies will approach each other with a continually increasing speed, and to keep them from doing this, a force is required. There is a force between the earth and the moon, the sun and the planets, two pendulums hanging near each other, etc. This fact will be discussed more fully in the chapter on Gravitation.

2. *Inertia.* — If the motion of a body is changed in any way by means of our muscles, we are conscious of the sensation of force; and the name "inertia" is given to that property of the body owing to which this is true. The following are illustrations: throwing a ball or stopping it,

setting a barrel rolling, slowing up a grindstone. If we push a box, we can tell by our sensation whether it is full or empty ; similarly, we can distinguish between a full and an empty barrel, the sensation being the more intense in the case of the full box or barrel. Further, when the change in the motion is sudden, the sensation is different from what it is when the change is gradual : it varies with the *rate of the change*, being more intense for a more sudden change. This change of motion may of course be produced under the "action" of any force, *e.g.* the firing of a cannon ball, the impact of one billiard ball upon another, etc. ; and the numerical values of the force and the rate of change of the motion are connected, as will be shown later. It is plain, though, from general considerations, that to produce a sudden change of motion requires a great force. For example, in beating a carpet, the particles of dust are driven out, because, as the carpet is struck by the stick and moves forward, there is not sufficient force of "friction" between the dust and the carpet to produce the necessary change in motion of the particles to enable them to stay in the carpet. Similarly, if a horse with a rider stops suddenly there may not be sufficient force holding the latter in the saddle to produce the necessary rate of change in motion, so as to stop him too, and he is thrown.

In the case of the barrel or grindstone, the question is somewhat different ; for it is not difficult to see that their inertia depends not alone upon the rate of change of motion and the matter considered as a whole, but also upon where the matter is located. Thus, the inertia of a wheel is greater if the matter is concentrated in the rim, as in the flywheel of an engine, than if it is distributed uniformly as in a grindstone. It will be shown later that the inertia of a body rotating round an axis varies as the distance of the matter from the axis is changed, other things being unaltered.

3. *Properties of size and shape.* — One of the most evident characteristics of matter is that it has certain properties of shape and size. Thus, the name "solid" is given to a portion of matter which, like a stone, a cork, a piece of putty, etc., has a shape and size independent of how it is held or supported. The name "fluid" is given to a portion of matter which "flows," that is, yields to any force, however small, which acts in such a manner as to tend to make it change its shape. There are two divisions of fluids, "liquids" and "gases." Liquids, such as water, molasses, oil, etc., will, if left to themselves, assume the shape of a sphere, *e.g.* raindrops, shot formed from molten lead, etc.; but if placed in hollow solid vessels (here on the surface of the earth), they take the shape of the vessel, flowing under the influence of gravity. Their volume, however, remains unchanged as they are poured from one vessel into another, regardless of their change in shape; and so, in general, unless the liquid entirely fills a closed vessel, there will be a horizontal upper bounding surface to the liquid. The possibility of forming drops or of exhibiting this so-called "free surface" is the criterion of a liquid as distinct from a gas. The latter, when inclosed in a vessel, *e.g.* a bottle or a soap bubble, is distributed uniformly throughout it, and has therefore its shape and size.

It is often difficult to say whether a piece of matter should be called a solid or a liquid; for it may have the properties of both as defined above, depending upon the time taken for the various changes. Thus a piece of pitch is brittle to a sudden blow and apparently a solid in all respects; yet, if put in a funnel, it will, after the lapse of months or years, take the shape of the latter and gradually *flow* out of the opening. Again, if by suitable air pumps the gas inside a hollow closed solid vessel, like a glass bulb, be exhausted as far as possible, the matter which remains behind has properties quite distinct from those of a gas, and its condition has been called by Sir William Crookes the "Fourth State of Matter."

The shape and size of all kinds of matter may, however, be changed by suitable means. If a cork is held in the hand, it is not difficult to compress it; a rubber cord can be elongated; a brass wire can be twisted; a rubber ball full of air can be squeezed, thus compressing the air; even water (and all liquids) can be compressed by suitable means. If these changes in size and shape are produced by the hands, one is conscious of a force; but, of course, they may be produced by other forces. The amount of the force required is, however, different for different bodies, and suitable names have been given the corresponding properties. A few of the more common of these may be described. Thus, a body is said to have "elasticity" or is called "elastic" if, after a deformation of size or shape under the action of a force, it returns to its previous condition when the force is removed. Glass, ivory, steel, water, all gases, are elastic; putty, lead, etc., are inelastic or "plastic," because when once deformed they remain so. A "rigid" body is one to change whose *shape* a great force is required, whereas one whose *volume* is easily changed is said to be "compressible." A solid is "ductile" if it can be drawn into wires, *e.g.* iron, brass, etc.; and if it can be hammered out into a thin film it is "malleable," *e.g.* gold, silver, aluminium. A fluid is "viscous" if it flows slowly, *e.g.* molasses, pitch.

Other Properties of Matter. — To describe fully the condition of a body it is not sufficient, however, to say that it is a solid or a liquid or a gas. In the first place, as has just been shown, this is often impossible; and in any case it is necessary to state exactly all its physical properties. Thus a body is under a certain "pressure"; it has a certain "density"; it is at a certain "temperature"; it has certain "electrical" or "magnetic" properties; etc. (These words will be defined later, each in its proper place; but they are in such everyday use that their meaning is probably more or less definite to every one.) In certain cases there are exact relations between the numerical values of these properties for a given body; for instance, if p, d, T, are symbols for the numerical values of the pressure, density, and "absolute" temperature of a gas, the relation

$$p = RdT$$

is found to hold, where R is a factor of proportionality different for different gases. Such a relation is called an "equation of condition." Thus, if the temperature is changed, one or more of the other properties are affected. Further, if any of these properties are altered, so are also, in general, the various "properties of size and shape," *e.g.* elasticity, rigidity, viscosity, etc.

If different kinds of matter are compared, it is observed that in some cases a body, when considered as made up of small but finite portions, is alike in all its parts: it is said to be "homogeneous," *e.g.* water, air, a piece of copper, etc.; whereas, in other cases, the various parts may differ from each other, *e.g.* an iron rod one end of which is in a flame, the trunk of a tree, a sponge, etc. (A body like a sponge or a piece of sandstone, which has cavities in it, is called "porous"; and all substances except "glazed" ones are so to a greater or less extent.) A body whose parts differ either in their chemical nature or in their physical condition is said to be "heterogeneous"; and if the division into parts is carried far enough, all bodies are heterogeneous. Thus, a piece of steel is apparently homogeneous; but when examined under a microscope, it is found to be far from this. Similarly, glass is most heterogeneous.

Again, taking any pure substance, *e.g.* water, we can imagine it divided and subdivided, the parts still being water; but it is shown in Chemistry that at last a portion of water will be obtained which is so small that when it is subdivided its parts have properties different from water. This last portion of a substance, which still retains the properties of the whole, is called a "molecule" of that substance. Molecules may be separated into parts in many ways, some of which will be discussed later. Matter, then, is to be regarded as discontinuous, being built up of molecules, which in turn have parts. The nature of these "ultimate" particles is a question which will be referred to again.

Properties of the Ether. — To analyze and describe the properties of the ether is not a simple problem, because it does not affect our senses directly; our knowledge of it is entirely indirect. The velocity of waves in the ether — "the velocity of light" — is known to be about 3×10^{10} centimetres, or about 186,600 miles, per second; and this fact proves that the ether has inertia, for it is seen from this that time is required to produce the motions which constitute the advance of the waves. Further, the ether must permit "elastic deformations" in a general sense, that is, changes of its condition under the action of suitable forces, such that when these forces are removed it returns to its previous state; otherwise waves could not be produced in it. We have no knowledge as to whether it is continuous or not; but, since waves producing the effect of light are propagated through air, water, glass, and all bodies to a greater or less extent, the ether must be present in them all. It can best be regarded as a universal medium permeating all space known to us, in which are immersed at certain points the minute ultimate portions of matter referred to above. There must be an extremely intimate connection between these particles and the ether, as will be shown later in discussing radiation and absorption. There is no evidence of the existence of gravitation between different portions of the ether; in fact, it is not improbable that the phenomena of gravitation, as seen with ordinary matter, depend upon the presence of the ether. All observations and experiments point to the idea that the ether is at rest; and that, as the earth moves in space or a raindrop falls, it moves through the ether with as much ease as a hand with fingers separated can move through air.

Fundamental Ideas in Physics. — We are accustomed to think of the properties of matter that are revealed to us by our muscular sense as fundamental ideas in nature, and to attempt to describe the properties of the ether in terms of them. It is impossible, however, to think that the properties of weight, of inertia, of elasticity, etc., are all independent of one another, and uninfluenced by the ether. In fact, as will be shown later, we can explain the elasticity of a gas as a result of the inertia of its parts. It is the aim of students at the present time to show that gravity and all the properties of size and shape can be explained in terms of the inertia of matter and of the properties of the ether. Our main reason in holding so firmly to the idea of the fundamental nature of inertia is that it corresponds to one of our senses; but we are not justified in believing that in a universe as large and as varied as ours, one of man's senses should be the fundamental fact. As will be seen later, we can connect the inertia of matter with certain phenomena in the ether, which, for lack of a better name, are called "electrical." If we can explain these in terms of the motion of portions of the ether, we shall have reduced all natural phenomena — in matter that is not living — to motions in the medium called the ether.

Physical Standards. — As has been explained in previous sections, the fundamental ideas of matter are based upon that of force; and therefore the fundamental concepts of nature are space, duration of time, and force. To measure a force involves, as will be shown later, the selection of a standard piece of matter; to measure space, that of a standard straight line; to give a number to an interval of time, that of a standard "time-keeping" apparatus. Different observers throughout the world wish to use the same standards; and those must therefore be selected which can be reproduced, so that copies can be distributed.

There is no ambiguity in regard to the meaning of the words "equal lengths" or "equal intervals of time"; and so, if a standard length is chosen or if a standard time-keeping apparatus is adopted, we know what is meant by making another rod of an equal length, or by selecting another time keeper whose period is the same. Consequently, we can make bars or rods whose lengths equal the standard length as closely as is necessary; and we can make clocks that have the same period as the standard time keeper. But what is meant by making a copy of the standard piece of matter? Matter has so many properties that, unless some one is chosen as a basis of comparison, the words "equal pieces of matter" have no meaning. This question will be discussed in full in a later chapter; for the present we shall speak of standards of length and time only.

Length. — The standard length that has been adopted by most civilized countries and by all scientific laboratories is the distance in a straight line between the end edges of a certain platinum bar that is kept in Paris, when this bar is at the temperature of melting ice; *i.e.* at 0° Centigrade. (This bar was constructed under the direction of Borda, the French physicist, and was officially adopted by France in 1799. It is known as the *Mètre des Archives.*) This length is called a "metre," for which the symbol is (m.); 1000 metres is called a "kilometre" (Km.); one hundredth of a metre, a "centimetre" (cm.); one thousandth of a metre, a "millimetre" (mm.); one thousandth of a millimetre, a "micron" (μ); and a thousandth of a micron has the symbol $\mu\mu$. Founded on any one of these units of lengths there can be defined evidently a unit area and a unit volume.

The original metre bar of Borda was constructed with the intention of having its length equal as nearly as possible to one ten-millionth of the earth's quadrant, *i.e.* the distance on the surface of the earth from its north pole to the equator, along the meridian circle passing through Paris; but certain numerical and experimental errors were made, so that there is no such exact relation between the metre and the earth. According to recent measurements, the length of the earth's quadrant is about 10,000,880 metres. Copies of the metre bar have been made, so designed that they may be used with greater ease and accuracy than the original; their slight discrepancies from the latter have been measured; and they are now used as the standards for comparison purposes.

The unit length that is used in Great Britain and in the United States is the "Standard Yard," which is defined by Act of Parliament as follows: "The straight line or distance between the centres of the transverse lines in the two gold plugs in the bronze bar deposited in the Office of the Exchequer shall be the general standard yard at 62° F. and if lost it shall be replaced by means of its copies." This bar is now kept in the Standards Office of the Board of Trade at Westminster. The relation between the metre and the yard is

$$1 \text{ m.} = 1.093633 \text{ yd.} = 39.37079 \text{ in.}$$

$$1 \text{ yd.} = 0.9143935 \text{ m.}$$

Time. — The standard time-keeping apparatus is the earth itself as it makes its revolutions around the sun. The interval of time of one complete revolution is called a "year." The period of one rotation on its axis is called a "day" — a "sidereal day," to be exact, because the direction of some star is taken as a means of determining when the period is complete. The "solar day" is the interval of time that elapses between two consecutive "high noons," *i.e.* instants of passage of the sun across the meridian, at any one point on the surface of the earth. The solar day varies in length during the year; but its mean value, *i.e.* the average of the solar days for a year, is found to remain constant from year to year, so far as our observations can determine; and this interval of time is called a "mean solar day." A pendulum making vibrations at such a rate that it completes $60 \times 60 \times 24$, or 86,400, periods in a solar day, is said to beat "mean solar seconds," for which the symbol is "sec.": 60 of these seconds is called a "mean solar minute" (min.); 60 such minutes, a "mean solar hour"; and 24 such hours make up, therefore, a mean solar day.

A Few Mathematical Principles. — The first branch of Physics to be considered in detail will be Mechanics, which is the science of the inertia of matter; but first it is advisable to discuss and explain a few purely mathematical ideas, *avoiding, however, all complicated points*.

1. *The limiting value of a ratio.* — Picture a body falling freely toward the earth and consider what is meant by the question "how fast is it falling at any instant?" The body is moving faster and faster as time goes on; but if at any instant we consider the motion during a small interval of time, *e.g.* one thousandth of a second, the change in the speed in this time is not great; and so if we represent this interval by Δt,* and the distance passed over in this time by Δx, the distance traversed "per unit of time," is $\frac{\Delta x}{\Delta t}$. As the interval of time considered is taken smaller and smaller, the distance traversed in that time also becomes less and less, and the motion changes less; *i.e.* it is more uniform. The limiting value of $\frac{\Delta x}{\Delta t}$ as Δt is considered smaller and smaller, is called the "rate" of falling at the instant considered. If this rate were to remain unchanged, its value would equal that of the distance traversed in a unit of time. Thus, we speak of a train as moving at any instant at the rate of sixty miles per hour, meaning that, if the motion could continue unchanged as it is at that instant, the train would actually move sixty miles during the next hour.

Again, consider any two quantities which are so related that if one is altered the other is also. Call their numerical values x and y. Then for any definite value of x the "rate of change of y with reference to x" is defined to be the

* It is customary to use the symbol "Δx" (read: "Delta x") to mean either a small variation in x, or a small amount of x.

limiting value of the ratio $\frac{\Delta y}{\Delta x}$, as Δx, a small variation in the value of x from that specified, is taken smaller and smaller.

2. *The numerical value of a plane angle.* — A plane angle is the difference in direction between two straight lines that lie in a plane. The scientific mode of assigning a number to one is as follows: Prolong the lines, if necessary, until they meet; with this point of intersection as a centre and with any line as a radius, describe a circle; let L be the length of the arc of this circle intercepted between the two lines and let R be the length of the radius of the circle; then by definition the ratio $\frac{L}{R}$ is the numerical value of the angle. Thus, in the figure, $\overline{AB}$ and $\overline{CD}$ are the two lines lying in a plane; O is their point of intersection; $\overline{OP}$ is the radius of the circle, *i.e.* R; $\overline{PQ}$ is the intercepted arc, *i.e.* L; hence the angle has the value $\frac{PQ}{OP}$.

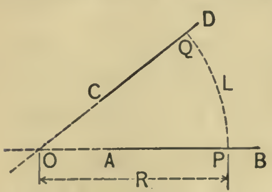


FIG. 1. — Numerical value of the angle is $\frac{L}{R}$.

This definition of the value for an angle is adopted because it is known from geometry that the ratio of the intercepted arc to the radius of its circle is the same for all values of the radius; and therefore it is not necessary to specify the latter. If N is the numerical value of the angle, this relation can be expressed $N = \frac{L}{R}$, or $L = RN$; in words, the length of the intercepted arc equals the product of the length of the corresponding radius and the value of the angle. A unit angle is called a "radian"; it is an angle such that the lengths of the intercepted arc and the radius are equal.

In ordinary language angles are expressed in "degrees," "minutes," and "seconds"; there being 60 seconds in a minute, 60 minutes in a degree, and 90 degrees in a right angle.

It is easy to find the relation between a radian and a degree; because, if R is the radius of a circle, $\frac{\pi R}{2}$ is the length of the arc intercepted by two lines making a right angle, where $\pi = 3.1416$ approximately; and therefore $\frac{\pi}{2}$, or 1.5708, is the value of a right angle in terms of radians. Hence,

$$1.5708 \text{ radians} = 90^\circ,$$
$$\text{or} \quad 1 \text{ radian} = 57^\circ.2958$$
$$= 57^\circ\ 17'\ 45''.$$

An angle has a *sign* as well as a numerical value. Thus, if $\overline{OA}$ be chosen as a fixed direction, there is a difference between the angles (AOB) and (AOC), although they are numerically equal. The latter corresponds to a rotation like that of the hands of a watch; the former to a contrary rotation. One angle, it is immaterial which, is called plus (+); the other, minus (−).

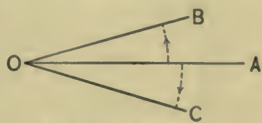


FIG. 2.—Angle (AOB) = −angle (AOC).

When an angle becomes very small, the ratio of the value of its sine to its own value approaches unity. For, referring to the cut, the value of the angle between $\overline{OA}$ and $\overline{OB}$ is the ratio of the arc $\overline{AB}$ to the radius; the value of the sine of this angle is defined to be the ratio of $\overline{CB}$ to the radius; therefore, the ratio of the angle to its sine is the ratio of the arc $\overline{AB}$ to $\overline{CB}$. This equals the ratio of the arc $\overline{BAD}$ to its chord $\overline{BCD}$; and, as the angle is made smaller and smaller, this ratio approaches unity, because in the *limit* an arc and its chord are equal.

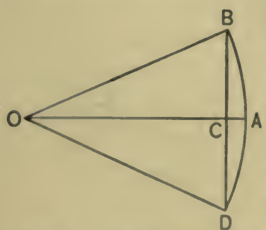


FIG. 3.—In the limit when the angle (AOB) is infinitesimal, $\sin (AOB) = (AOB)$.

3. *Vectors and vector quantities.*—A vector is a limited portion of a straight line in a definite direction. Thus the

straight lines $\overline{AB}$ and $\overline{CD}$ are vectors; their lengths are the distances between A and B and between C and D; and their directions are indicated by the arrows. Three ideas are involved: the direction of the line, the *sense* of this direction (*i.e.* a distinction is made between a line drawn to the right and one drawn to the left, etc.), and the length of the line. The *position* of the line is immaterial; so two vectors of the same length and in the same direction, wherever placed, are equal. A vector, then, is a straight line traced by a point moving *from one position to another*, as is indicated by the use of an "arrow" in the line.

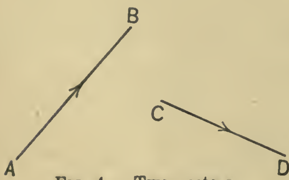
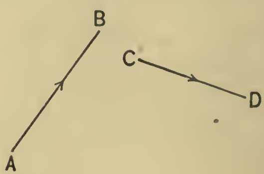


Fig. 4. — Two vectors.

The process of "addition of vectors" is defined as follows: move one vector parallel to itself until one of its ends meets that end of the other which causes the arrows to indicate continuous advance from the free end of one vector to that of the other, and then join the former free end to the latter by a straight line. The "sum" is therefore a vector. Thus $\overline{AB}$ and $\overline{CD}$ may be added in two ways: (1) move $\overline{CD}$ parallel to itself until C coincides with B, — the arrows now indicate continuous advance from A to D, — and join these points by a



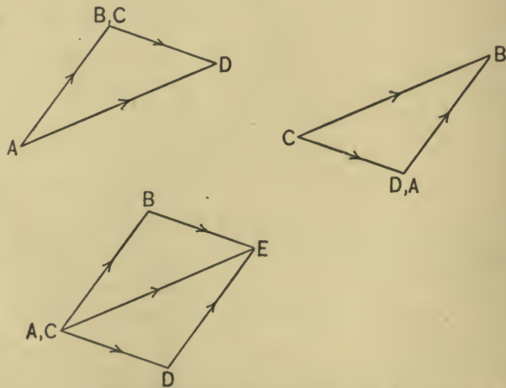


Fig. 5. — Three methods for the addition of the vectors $\overline{AB}$ and $\overline{CD}$.

straight line, thus forming the vector $\overline{AD}$; (2) move $\overline{CD}$ parallel to itself until D coincides with A, — the arrows now indicate continuous advance from C to B, — and join these points by a straight line, thus forming the vector $\overline{CB}$. It is evident from geometry that these two vectors are identical, having the same length and the same direction and sense. (If a parallelogram is formed, having the two vectors as adjacent sides, both starting from the same point, the diagonal is their sum.)

This process is called "geometrical addition"; and it can obviously be extended to three and more vectors.

The simplest case is evidently that when the two vectors are in the same straight line: if they are in the same sense, the numerical value of the sum is the ordinary arithmetical sum; while, if they are in opposite senses, it is their arithmetical difference. If, then, two vectors are in the same line and in the same sense, both may be called positive; but if they have opposite senses, we should call one positive (+) and the other negative (−); and their geometrical sum equals in numerical value the *algebraic* sum in both cases, and has the direction of the two vectors. Its *sense* of direction in the former case is that of both vectors; in the latter, that of the greater.

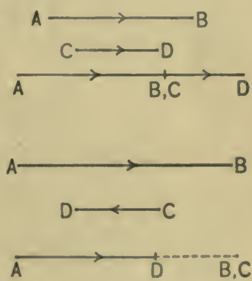


FIG. 6. — Addition of two parallel vectors.

Looking at this process of geometrical addition in a converse manner, it may be said that the vector $\overline{AD}$ is the geometrical sum of $\overline{AB}$ and $\overline{CD}$, where $\overline{AB}$ and $\overline{CD}$ are any two vectors such that, when added, their initial and final points are A and D: the vector $\overline{AD}$ is said to be "resolved into components." The case when the two components are at right angles is the most important.

Let $\overline{AD}$ be any vector and $\overline{OP}$ any straight line; drop perpendiculars $\overline{AA'}$ and $\overline{DD'}$ upon $\overline{OP}$; $\overline{A'D'}$ is called the "projection of $\overline{AD}$ upon $\overline{OP}$." Draw through A a line parallel to $\overline{OP}$; it intersects $\overline{DD'}$ in B. Then the vector $\overline{AD}$

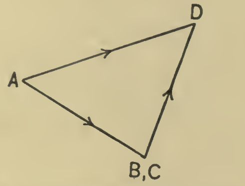


Fig. 7. — Resolution of the vector $\overline{AD}$ into components.

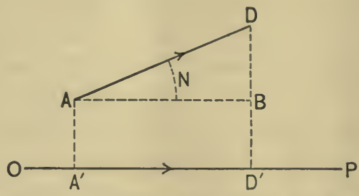


Fig. 8. — Projection of the vector $\overline{AD}$ upon the line OP.

equals the geometrical sum of the vectors $\overline{AB}$ and $\overline{BD}$. Let the lengths of $\overline{AB}$, $\overline{BD}$, and $\overline{AD}$ be b, v, and h; then, by geometry, $h^2 = b^2 + v^2$; and, if N is the angle (BAD), by the definitions of trigonometry:

$$\frac{v}{h} = \text{sine } N, \quad \frac{b}{h} = \text{cosine } N, \quad \frac{v}{b} = \text{tangent } N,$$

or, as ordinarily written,

$$v = h \sin N, \; b = h \cos N, \; v = b \tan N.$$

So the projection of $\overline{AD}$ on $\overline{OP}$ equals the product of $\overline{AD}$ and cosine N.*

The vector $\overline{AB}$ is called "the component in the direction $\overline{OP}$ of the vector $\overline{AD}$." (If $\overline{AB}$ and $\overline{BD}$ were not perpendicular, *i.e.* if (ABD) were not a right angle, the latter vector might be so resolved as to have a component in the direction $\overline{OP}$; and in that case the former would not be the only component of the vector $\overline{AD}$ in this direction.) But, as just shown, $\overline{AB} = \overline{AD} \cos N$. The general rule, then, for

* In a similar manner if perpendicular lines are dropped upon a plane from the points forming the contour of any limited surface, the area inclosed by the feet of these lines is called the projection on this plane of the limited surface. If this surface is plane and has the area A, if the projected area is A_1, and if the angle between the two planes (*i.e.* between lines perpendicular to them) is N, it is seen that $A_1 = A \cos N$.

obtaining the numerical value of the component in a particular direction of a given vector is to *multiply the numerical value of the vector by the cosine of the angle between it and the specified direction.*

It is obvious from geometry that the component of any vector $\overline{AC}$ along any line $\overline{OP}$, *i.e.* $\overline{A_1C_1}$, is the algebraic sum of the components along this same line of any two vectors whose geometrical sum is $\overline{AC}$, *e.g.* $\overline{AB}$ and $\overline{BC}$; viz., $\overline{A_1C_1} = \overline{A_1B_1} + \overline{B_1C_1}$.

The geometrical addition of the two vectors $\overline{AB}$ and $\overline{BC}$ to produce the vector $\overline{AC}$ may be expressed in words as follows: In order to obtain the vector $\overline{AC}$ by joining a vector to $\overline{AB}$, it is necessary to add the vector $\overline{BC}$. Therefore, $\overline{BC}$ is the "geometrical difference" between the vector $\overline{AC}$ and the vector $\overline{AB}$. Thus, the rule for subtracting one vector from another is evident; the difference is such a vector that, if added to the former, it gives the latter.

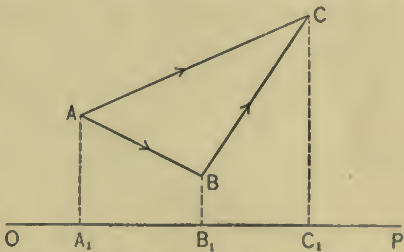


Fig. 9.—The projection of the vector $\overline{AC}$ upon $\overline{OP}$ equals the sum of the projections of the two components $\overline{AB}$ and $\overline{BC}$.

There are many physical quantities which require for their description a numerical value and a direction, *e.g.* velocity, force; they can be represented graphically by a vector, and are called "vector quantities."

4. *Average or mean values.* — If there is a set of similar quantities, a_1, a_2, $\cdots$, a_n, their "arithmetical mean" is defined to be $\frac{a_1 + a_2 + \cdots + a_n}{n}$. It often happens, however, that, owing to the physical or mathematical conditions, the various quantities are not of the same importance. For instance, suppose the mean age of a class of students is desired: let the age

of 2 members be 16, of 5 be 17, of 10 be 18, of 1 be 19; the ages are then 16, 17, 18, and 19, but the *importance* of 17 in the average is 5 times that of 19. In this case evidently the proper mode of finding the mean value is to multiply 16 by 2, 17 by 5, etc.; add these products and divide by the number in the class, *i.e.* $2 + 5 + 10 + 1$. So in the general case, if m_1 is the *importance* of a_1, m_2 that of a_2, etc., the mean value is

$$\frac{m_1a_1 + m_2a_2 + \cdots + m_na_n}{m_1 + m_2 + \cdots + m_n}.$$

If a quantity assumes different values in succession, but in a continuous manner, *e.g.* the speed of a falling body, it is often necessary to find its mean value for a definite interval of time. This can be done by a graphical method, using the theory of limits. Let a_1 be the value of the quantity at the instant of time T_1; a_2 its value at T_2; etc. Choose the intervals of time $T_2 - T_1$, $T_3 - T_2$, etc., equal. The *importance* in the average of a given value, *e.g.* a_m, is the length of time it lasts. The actual quantity during the interval $T_2 - T_1$ is changing from a_1 to a_2; during the interval $T_3 - T_2$ it is changing from a_2 to a_3; etc. As an approximation, let us assume that the quantity keeps its value a_1 during the interval $T_2 - T_1$, then suddenly assumes the value a_2, which it keeps during the interval $T_3 - T_2$, etc. By the above definition, the mean value of a is, therefore,

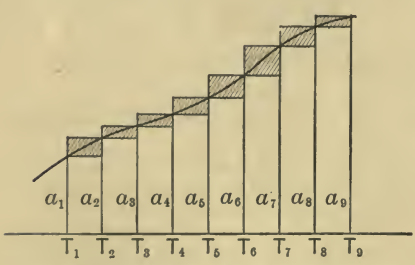


FIG. 10. — Graphical method for determining the mean value of a series of quantities.

$$\frac{(T_2 - T_1)a_1 + (T_3 - T_2)a_2 + \cdots + (T_n - T_{n-1})a_{n-1}}{(T_2 - T_1) + (T_3 - T_2) + \cdots + (T_n - T_{n-1})}.$$

The value of the denominator is evidently $T_n - T_1$, or the total interval of time. The numerator can be represented graphically. On a horizontal line mark points T_1, T_2, ⋯, T_n at equal intervals apart, and at these points erect vertical lines of length a_1, a_2, ⋯, a_n. Construct rectangles having a_1 and $(T_2 - T_1)$, a_2, and $(T_3 - T_2)$, etc., as sides, as shown by the unshaded rectangles in the cut. The numerator of the above fraction is the value of the sum of the areas of these parallelograms. This sum divided by the total interval of time is the *approximate* mean of a_1, a_2, etc. Another approximate mean may be obtained by assuming that the value of the quantity during the interval $T_2 - T_1$ is a_2; during $T_3 - T_2$ is a_3; etc. The mean is then

$$\frac{(T_2 - T_1)a_2 + (T_3 - T_2)a_3 + \cdots + (T_n - T_{n-1})a_n}{(T_2 - T_1) + (T_3 - T_2) + \cdots + (T_n - T_{n-1})}.$$

The denominator is, as before, the total interval of time; and the numerator is the sum of the areas of the rectangles whose sides are $(T_2 - T_1)$ and a_2, $(T_3 - T_2)$ and a_3, etc., as shown in the cut by the rectangles formed by adding the shaded portions to the former ones.

These two numerators are evidently not equal; but, as the intervals of time $T_2 - T_1$, $T_3 - T_2$, etc., are taken smaller and smaller, the two sets of rectangles approach the same limit; viz., the area included between the base line, the two vertical lines of lengths a_1 and a_n, and the line (curved or straight) that passes through the ends of the vertical lines erected at the points T_1, T_2, etc., as these in the limit become consecutive. So the true mean value of the quantity a is the "area of the curve" as just described, divided by the value of the total interval of time.

The simplest case is when the quantity is varying uniformly with the time; *i.e.* $\frac{a_2 - a_1}{T_2 - T_1} = \frac{a_3 - a_2}{T_3 - T_2} = \text{etc.} = \frac{a_n - a_{n-1}}{T_n - T_{n-1}}$ whatever T_1, T_2, etc., and T_n are; for the line described above as the locus of the ends of the lines a_1, a_2, etc., and a_n, is evidently a straight line. Its "area" is known from geometry to be

$$\tfrac{1}{2}(a_n + a_1) \times (T_n - T_1).$$

Hence the mean value of the quantity is this divided by $(T_n - T_1)$, the total interval of time; *i.e.* $\frac{1}{2}(a_n + a_1)$ or *the arithmetical mean of the initial and final values.*

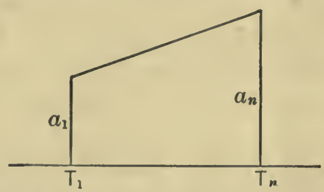


Fig. 11. — Special case, when the quantity whose mean is desired is varying at a uniform rate.

Another mode of taking means has its origin as follows: Let a_1, a_2, a_3, etc., form a "geometric series"; *i.e.* $a_2 = ra_1$, $a_3 = ra_2 = r^2a_1$, etc. It is seen that $a_2 = \sqrt{a_1a_3}$. So, in general, the "geometric mean" of two similar quantities a and b is defined to be $\sqrt{ab}$.

MECHANICS AND PROPERTIES OF MATTER

INTRODUCTION

We have recognized three so-called fundamental properties of matter: inertia, weight, and the one which includes the varied characteristics of size and shape. Each of these will now be considered in greater or less detail.

As has been said before, the science of Mechanics is that branch of Physics which deals with the inertia of matter. It is often divided into two parts, "Kinematics" and "Kinetics": the former is the science of motion considered apart from matter; that is, it treats of possible *motion;* the latter is strictly the science of the inertia of matter. If there is no change of any kind in the motion, what we call "rest" being a special case of this, the science is called "Statics"; while if the motion is changing, the science is called "Dynamics." These two sciences are branches, then, of Kinetics. Statics can, however, be considered as a special limiting case of Dynamics; and this plan is adopted in the present book; so Mechanics will be treated under the two divisions, Kinematics and Dynamics. In the former, the question as to the different possible kinds of motion will be discussed; in the latter, the physical conditions under which these types of motion occur. Kinematics is a geometrical science; dynamics, a physical one.

Weight will be discussed under the more general head of Gravitation. This will be followed by several chapters on the properties of solids, liquids, and gases.

CHAPTER I

KINEMATICS

General Description. — Kinematics has been defined as the science of motion apart from matter; that is, it is concerned with the study of the possible motions of the *geometrical* quantities: a point, a plane figure, and a solid figure. For the sake of illustration, many material bodies will be referred to; but all the statements and theorems are meant to apply to *figures*, not to *bodies*, unless the contrary is expressly noted.

If the motion of any actual body is observed (for instance, a stick thrown at random in the air, a moving baseball, the wheel of a moving wagon), it is seen that there are two types of motion involved: the object moves as a whole, and it also turns. These motions are independent of each other; one may occur without the other. In the up and down motion of an elevator, in the motion of a railway car on a straight track, etc., there is no turning; in the motion of the fly-wheel of a stationary engine, in the opening or closing of a door, etc., we may say that the motion is one of turning only.

The name "translation" is given to that kind of motion during which all lines in the figure remain parallel to their original positions; further, all points of the figure move through paths that are geometrically identical. Thus, to describe completely any case of translation, all that is necessary is to describe the motion of any one point of the moving figure.

The name "rotation" is given to that kind of motion during which each point of the figure moves in a circle; *e.g.* a door as it opens or closes. The planes of these circles are parallel, and their centres all lie on a straight line which is called the "axis." If a plane section, perpendicular to the axis, is taken through the rotating figure, all the lines of the figure in this plane have identically the same angular motions; otherwise the figure would break up into parts. To describe, therefore, motion of rotation at any instant, we must know two things: the position and motion of the axis and the angular motion of any line *fixed in the figure* with reference to any line *fixed in space*, provided the two lines lie in the same plane perpendicular to the axis. Thus, consider the rotation of a figure like that of a grindstone; its axis is the central line of the axle. In the cut, which represents a cross section by a plane perpendicular to the axis, let $\overline{PQ}$ be a line fixed in the moving figure and $\overline{AB}$ be a line fixed in space; the position and motion of the figure at any instant are given by a knowledge of the angle between these lines and of its changes in value. It should be noted that this is a *special* case of rotation, because the axis does not change its position, as it does in general.

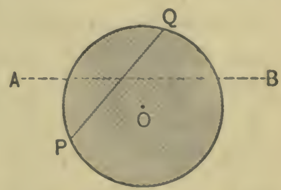


FIG. 12.—Rotation of figure around a fixed axis, O, is defined by the angle between $\overline{AB}$, a line fixed in space, and PQ, a line fixed in the figure.

To describe the most complicated motion, therefore, we must consider it resolved into two parts, a translation and a rotation, and must discuss each separately.

Translation

In translation, as has been already explained, it is necessary to describe the motion of a point only. The simplest case of this is motion along a straight line; but the most general case, that of motion along a curved line, is not difficult. To describe this motion the first thing that it is necessary to know is the position of the point at any instant with reference to some fixed figure.

Linear Displacement.—Let the path of the point with reference to some fixed figure be represented by the curve in the cut. Let *O* be its position at any instant, and *P* that at a later time. The vector $\overline{OP}$ is called the "linear displacement" of the point *with reference to the fixed figure* during this interval of time. This same vector might be the displacement for any motion that passed through *O* and *P*; or, in other words, a point may pass from *O* to *P* by various paths. The displacement is then a vector quantity and may be resolved into components in as many ways as we choose; conversely, two or more displacements may be compounded by geometrical addition. The importance of mentioning the "fixed figure of reference" may be seen from an illustration: if a stone is dropped from the top of the mast of a moving steamer, it will fall at its foot; the displacement with reference to the steamer is a vertical line, while with reference to the earth it is an oblique one, being the geometrical sum of the vertical line and the displacement of the steamer.

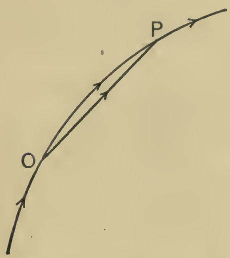


FIG. 13.—The vector $\overline{OP}$ is the linear displacement of *P* with reference to *O*.

Linear Velocity.—If the interval of time taken for the displacement is extremely small, *P* is very close to *O*; and, in the limit, the displacement $\overline{OP}$ coincides with the actual path along the curve, and has, in fact, the direction of the tangent to the curve at the point *O*. If, as the displacement becomes very small, its length is represented by Δx, and the corresponding interval of time by Δt, the ratio $\frac{\Delta x}{\Delta t}$ in the limit is called the "linear velocity" at the point *O* with reference to the fixed figure; that is, it is the "rate of change" of the displacement. It is evidently a vector quantity for it is defined by its numerical value, which is that of $\frac{\Delta x}{\Delta t}$ in the limit, and by the direction and "sense" of the displacement in the limit; viz., its direction is that of the tangent at O drawn from O to P, when P is close to O. The numerical value of the linear velocity is called the "linear speed"; so that the velocity at any point is characterized by the value of the speed and by the direction and "sense" of the tangent to the path at that point.

If the motion is uniform along a straight line, that is, if the velocity is constant (both in amount and in direction), the speed is equal numerically to the distance traversed in a unit of time; and if the motion is not uniform, the speed at any instant is the distance which the point would travel during the next unit of time if the motion were to remain uniform. The unit of linear speed on the C. G. S. system is the speed of "1 cm. in 1 sec."; and the unit of linear velocity in a definite direction is the unit of speed in that direction. (This C. G. S. unit speed has not received a name; in fact, the only system of units in which there is a unit speed which has received a name is that based on the nautical mile — 6080 ft. — as the unit length, and the hour as the unit time: the speed "one nautical mile in one hour" is called a "Knot." The expression "16 knots per hour" is therefore incorrect; for a knot is a speed, not a length.)

Since the linear velocity is a vector quantity, it can be resolved into components; and, conversely, two or more linear velocities may be compounded by geometrical addition. These statements are illustrated by many familiar facts: if a man walks across a moving railway carriage, his velocity with reference to the ground is compounded of that of the train and of that which he would have

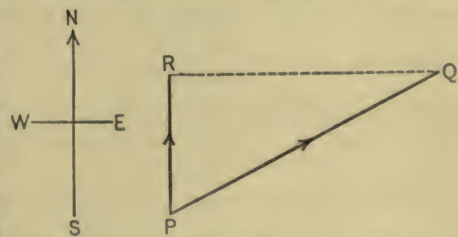


FIG. 14. — The component of $\overline{PQ}$ in the northern direction is $\overline{PR}$.

if the train were at rest; if a boat is rowed across a river, the actual velocity with reference to the earth is the geometrical sum of that of the water of the river and of that due to the oars; the velocity of a raindrop with reference to the window pane of a moving carriage as it strikes it is the geometrical sum of a velocity equal but opposite to that of the carriage and of its own downward velocity at that instant; if a man walks in a northeast direction with a speed of s cm. per second, his velocity may be represented by a vector $\overline{PQ}$ whose length is proportional to s and whose direction is northeast; and his velocity in a northern direction is given by the component, $\overline{PR}$, in the direction north and south, whose numerical value is $s \cos 45°$, or, more properly, $s \cos \frac{\pi}{4}$.

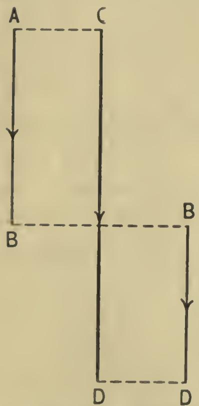


Fig. 15.—Rectilinear motion: AB and CD are the velocities at different instants; BD is their difference.

Similarly, one velocity may be subtracted from another, the difference being also a velocity. We will consider two illustrations: a body falling freely toward the earth and an extremely small particle moving in a circle with a constant speed. In the first case, the velocity at any instant is represented by a vertical vector $\overline{AB}$ and at some later instant by another vertical vector $\overline{CD}$ of greater length, because as the body falls, its speed increases. Call the length of $\overline{AB}$ s_1, and of $\overline{CD}$ s_2. The change in velocity is the difference between these vectors; that is, it is a vertical vector $\overline{BD}$ of length equal to $s_2 - s_1$.

In the second case in which the particle is moving in a circle let the constant speed be s; and let the direction of motion be that indicated by the arrows. When the particle is at the point A, its velocity has the direction of the tangent and the numerical value s; it can therefore be represented by the vector $\overline{PQ}$ which is parallel to the tangent at A, and has a length proportional to s. Similarly, when the particle is at the point B, its velocity can be represented by the vector $\overline{PS}$ which is parallel to the tangent at B and whose length is equal to that of $\overline{PQ}$ (since the speed does not alter). The change in the velocity in the time taken for the particle to move from A to B is the difference between the vectors $\overline{PS}$ and $\overline{PQ}$; that is, it is the vector $\overline{QS}$.

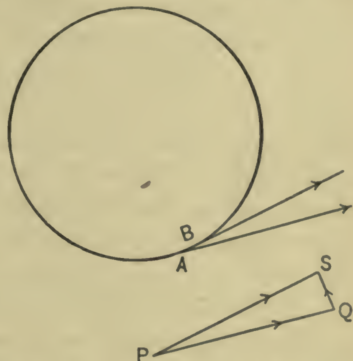


FIG. 16. — Uniform motion in a circle. $\overline{PQ}$ and $\overline{PS}$ are the velocities of the point at A and B.

Linear Acceleration. — To return to the original problem, that of describing the general case of the motion of a point in a curved path, we have defined the displacement and the velocity at any point, the latter being the rate of change of the former. The velocity may change, however, both in direction and in speed; and its rate of change at any instant is called the "linear acceleration" at that instant; that is, if Δv is the change in the velocity during the time Δt, the limiting value of $\frac{\Delta v}{\Delta t}$ is the acceleration. Moreover, since the change in the velocity, Δv, is a vector quantity, so is also the acceleration. Its numerical value is that of $\frac{\Delta v}{\Delta t}$ in the limit; and its direction is that of Δv in the limit.

We may consider separately two cases; in one of which the direction remains constant but the speed changes, while in the other the speed remains constant but the direction changes. As an illustration of the former we may take the motion of a falling body; and of the latter, the uniform motion of a particle in a circle. These two cases have already been partially discussed.

In the former motion let the change in speed from s_1 to s_2 take place in the interval of time $T_2 - T_1$; then the acceleration has the numerical value $\frac{s_2 - s_1}{T_2 - T_1}$ when the interval $T_2 - T_1$ is taken infinitely small, and its direction is vertically down. If the acceleration is constant, it is, therefore, the change in the speed in a unit of time.

In the latter case, that of uniform motion in a circle, let the interval of time during which the particle moves from A to B, and the velocity accordingly changes from $\overline{PQ}$ to $\overline{PS}$, be taken extremely small, so that the length of the arc $\overline{AB}$ becomes minute also; then, if this interval of time is called Δt, the acceleration at the point A is the limiting value of $\left(\frac{\text{vector } \overline{QS}}{\Delta t}\right)$. Call the lengths of the various straight lines in the diagram by the letters marking their terminal points: by geometry the triangles (SPQ) and (BOA) are similar, hence, $\frac{\overline{QS}}{\overline{PQ}} = \frac{\overline{AB}}{\overline{OA}}$; but $\overline{PQ}$ equals s, the value of the constant speed; $\overline{OA}$ equals the radius, r, of the circle; and in the limit, when Δt is infinitely small, the length of the *chord* $\overline{AB}$ equals the length of the *arc* $\overline{AB}$, the value of which is $s \cdot \Delta t$, because s is the distance traversed in one second, and hence the distance $s \cdot \Delta t$ is traversed in Δt seconds. It follows, then, that the *numerical value* of the acceleration at the point A, viz. $\lim\left(\frac{\overline{QS}}{\Delta t}\right) = \lim\left(\frac{\overline{PQ} \times \overline{AB}}{\Delta t \times \overline{OA}}\right) = \frac{s \times s\Delta t}{\Delta t \times r} = \frac{s^2}{r}$.

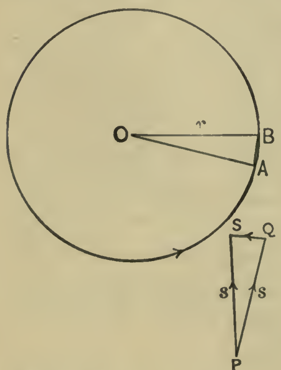


FIG. 17. — Uniform motion in a circle: the acceleration is the limiting value of $\frac{QS}{\Delta t}$.

The *direction* of the acceleration is that of the vector $\overline{QS}$ in the limit, when B approaches A, and therefore S approaches Q. Since $\overline{PQ}$ and $\overline{PS}$ are of equal length, in the limit $\overline{QS}$ is perpendicular to $\overline{PQ}$. But $\overline{PQ}$ is parallel to the tangent at A; and $\overline{QS}$ is therefore parallel to the radius $\overline{OA}$ of the circle in which the point is moving, and is directed toward the centre. In conclusion, then, when a point is moving in a circle with constant speed, its acceleration at any point is along the radius drawn from the point toward the centre and has the numerical value $\frac{s^2}{r}$, where s is the value of the constant speed and r is the length of the radius. (It is thus seen that, calling Δv the infinitesimal change in the velocity in the infinitesimal interval of time Δt, Δv is a vector at right angles to the one representing the velocity at any point; and that their geometrical sum is a vector, lying in their plane, whose numerical value in the limit is the same as that of the original velocity but whose direction is different; in fact, its direction is that of the tangent at a point of the circle infinitely near the original point. In words, by continually adding to the velocity an infinitesimal velocity perpendicular to it, the speed does not change, while the direction does.)

Since an acceleration is a vector quantity, we may resolve one in any manner; or, conversely, we may add two or more accelerations geometrically. Illustrations are of common occurrence. Experiments show, as will be explained later, that a body falling freely near the earth acquires an acceleration whose numerical value is about 980 on the C. G. S. system (*i.e.* in one second its speed increases by the amount 980 cm. per second); this is ordinarily written g, and is the same for all bodies, whatever their size, shape, or mass, provided they fall in a vacuum, but differs slightly at different latitudes on the earth's surface. If, then, a body is moving down an "inclined plane" (that is, a surface inclined to the horizontal plane) which is so smooth that friction can be neglected, the acceleration down the plane is the component parallel to the plane of the vertical acceleration g, if we *assume* that accelerations in nature can be "resolved." Let the latter be given by the vector $\overline{AB}$; its component parallel to the plane is the vector $\overline{AC}$, where $\overline{BC}$ is perpendicular to the plane. If the angle between the plane and the earth is called N, it is evident that the angle $(ABC) = N$; and the length of the vector $\overline{AC}$ equals $g \sin N$. (As N is made small, this component acceleration becomes small; and it may become so small that it can be measured easily, whereas g itself is so large that any direct measurement of it is extremely difficult. This experiment was first performed by Galileo.)

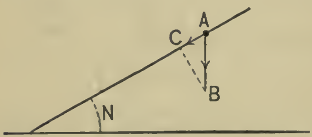


FIG. 18. — The component along an inclined plane of a vertical vector $\overline{AB}$ is $\overline{AB} \sin N$.

Again, when a small piece of iron is brought near a magnet, each moves toward the other, if it is free to do so, and acquires an acceleration; and if the piece of iron is allowed to fall freely past the end of the magnet, its acceleration at any instant is the geometrical sum of that which it would have if the magnet were not there, viz. g, and of that which it would have if there were no gravity. (The fact that in nature the accelerations produced by forces can be added geometrically involves the definite *assumption* that forces act independently of one another.)

The acceleration of a moving point is not necessarily constant either in direction or amount; but no name has been given the "rate of change of linear acceleration," and, in fact, it is a quantity of no importance physically, as we shall see later.

Special Cases. — There are several special cases of translation which deserve detailed description owing to their great importance in problems of Physics.

1. *A point moves in a straight line with a constant acceleration.*

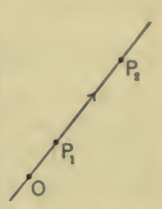


FIG. 19. — Rectilinear motion.

Since the acceleration is constant, its numerical value equals that of the change in speed in one second; and therefore, if the value of the speed at any point P_1 and instant of time T_1 is s_1, that at the point P_2 which is reached at the instant T_2 is given by the following equation, in which a is the value of the acceleration:

$$s_2 - s_1 = a(T_2 - T_1),$$

or

$$s_2 = s_1 + a(T_2 - T_1). \qquad (1)$$

(It should be noted that, if the sense of the direction of a is that of the velocity, the speed increases in value with the time, *e.g.* a falling body; whereas, if the sense of the two directions are *opposite*, a and s have different signs, and the formula becomes $s_2 = s_1 - a(T_2 - T_1)$ if a is here the *numerical* value of the acceleration; therefore the speed decreases, *e.g.* a body thrown vertically upward or a body thrown along a sheet of ice.)

If the speed were constant, the distance traversed in t seconds would be the product of this by the value of the speed; but the latter is varying from instant to instant. Since the speed is, however, by assumption changing at a uniform rate, its *mean value* (see page 32) between the instants T_2 and T_1 is the average of s_2 and s_1, *i.e.* $\frac{s_2 + s_1}{2}$; and the distance passed over in the interval of time $(T_2 - T_1)$ is

$$\frac{(s_2 + s_1)}{2} \times (T_2 - T_1) = s_1(T_2 - T_1) + \tfrac{1}{2} a(T_2 - T_1)^2.$$

If any point O in the line of motion is chosen as a point of reference or "origin," and if the distances $\overline{OP}_1$ and $\overline{OP}_2$ are called L_1 and L_2, the displacement in the interval of time $T_2 - T_1$ is $L_2 - L_1$. That is,

$$L_2 - L_1 = s_1(T_2 - T_1) + \tfrac{1}{2} a(T_2 - T_1)^2,$$

or
$$L_2 = L_1 + s_1(T_2 - T_1) + \tfrac{1}{2} a(T_2 - T_1)^2. \qquad (2)$$

(If the acceleration is an opposite sense to the speed, a must be given a negative value.) Thus, if the acceleration is known, and if the position and speed at any one instant are given, they can be predicted for any future instant. Conversely, if any motion is found to obey either of these laws (for one is a consequence of the other), it is known that the acceleration is a constant.

These two formulæ assume their simplest form when we agree to measure time and distance from the instant and position in which the moving point is at rest. For instance, let the point be at rest at P_1, *i.e.* $s_1 = 0$; then we will choose O to coincide with P_1, *i.e.* $L_1 = 0$; and also choose this instant as the one from which to measure time. Hence the formulæ become

$$s_2 = aT_2,$$

$$L_2 = \tfrac{1}{2} aT_2^2.$$

These are due to Galileo, and it was by showing that when a body moved down an inclined plane the displacement varied as the square of the time taken, that he convinced himself of the constancy of the acceleration parallel to the plane. Since this is constant, so is that for a body falling freely.

In the general formula (2) it is seen that the displacement, $L_2 - L_1$, is made up of two parts: $s_1(T_2 - T_1)$ is the distance the point would have gone if there had been no acceleration; $\frac{1}{2} a(T_2 - T_1)^2$ is therefore the additional displacement owing to the acceleration.

Another general formula may be obtained from equations (1) and (2) by eliminating $(T_2 - T_1)$ from them: substitute in (2) the value of $(T_2 - T_1)$ obtained from (1,) viz. $\frac{(s_2 - s_1)}{a}$. This gives

$$a(L_2 - L_1) = \tfrac{1}{2}(s_2^2 - s_1^2). \qquad (3)$$

Comparing equations (1) and (3), it is seen that the former defines a as a function of the interval of *time*, viz., $a = \frac{s_2 - s_1}{T_2 - T_1}$, while the latter expresses it as a function of the *displacement*, viz., $a = \frac{1}{2}\frac{s_2^2 - s_1^2}{L_2 - L_1}$.

2. *A point moves with a constant acceleration in one direction and with a constant velocity in a direction at right angles to this.*

Let the constant acceleration have the numerical value a; and let distances in its direction, measured from the point at which the speed in this direction is zero, be called y. Then, by counting time from the instant when the point is in the position for which $y = 0$, we have the relation $y = \frac{1}{2} at^2$ for any later time t.

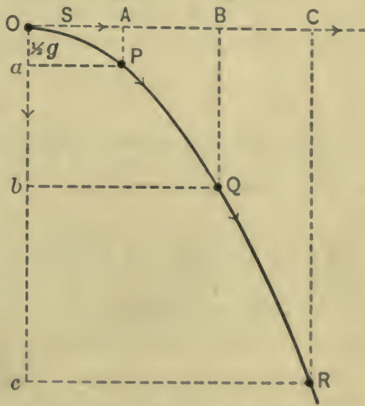


FIG. 20. — Parabolic motion: composition of a uniform horizontal motion with a uniformly accelerated vertical motion.

Similarly, let the constant velocity have the numerical value s; and let distances in its direction, measured from the point at which $y = 0$, be called x. Then, after the interval t, $x = st$. The resultant displacement is the geometrical sum of x and y.

The path of the moving point may be found by actual geometrical construction, giving to t different values; or, more simply, by the elimination of t from the two equations. This gives $y = \frac{1}{2}\frac{a}{s^2}x^2$; and, if this relation between x and y is plotted, it will give the path of the point. The curve is evidently a parabola, having as its axis the direction of the acceleration.

We need not, however, consider the motion as beginning at the point where $y = 0$. Let two lines be chosen at right angles to each other, $\overline{OD}$

in the direction of the constant acceleration, $\overline{OB}$ perpendicular to it. Let a point be projected obliquely to $\overline{OB}$ with the velocity V and making the angle N with it. This velocity can be resolved into two, $V \cos N$ along $\overline{OB}$, $V \sin N$ along $\overline{OD}$, but in the opposite *sense*. Therefore, assuming that the two motions at right angles to each other are independent, the point will maintain through its motion a constant velocity $V \cos N$ parallel to $\overline{OB}$; and, if the constant acceleration along $\overline{OD}$ is a, the point will move in the direction opposite to $\overline{OD}$ for an interval of time t, where $ta = V \sin N$. From this point on, the conditions are exactly as in the previous problem; so the point describes a parabola. But since the path is now a parabola, by symmetry, it must have been so before this point was reached. In the diagram, this point which marks the instant when the motion along $\overline{OD}$ in the opposite sense ceases and that along $\overline{OD}$ begins, is shown by C. The time taken for the moving point to reach it has just been shown to be $t = \frac{V \sin N}{a}$; and the displacement in this time along $\overline{OB}$ is the product of t and $V \cos N$, or $\frac{V^2 \sin N \cos N}{a}$. This is the distance $\overline{OA}$ on the diagram. The line $\overline{AC}$ has the length $\frac{1}{2} at^2$ or $\frac{1}{2} \frac{V^2 \sin^2 N}{a^2}$. The moving point in its further motion will cross the line $\overline{OB}$ at a point B, where $\overline{OB} = 2\,\overline{OA}$ or $\overline{OB} = \frac{2\,V^2 \sin N \cos N}{a}$, which may be written $\frac{V^2 \sin 2\,N}{a}$.

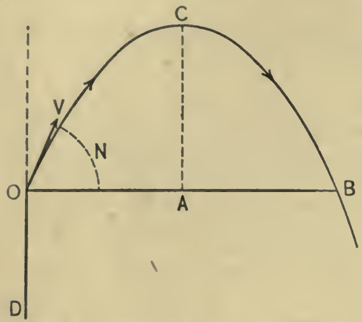


Fig. 21. — Motion of a projectile: $\overline{OV}$ is the direction of the initial velocity; $\overline{OB}$ is the horizontal trajectory.

For given values of V and a, this distance $\overline{OB}$ has its greatest value when $\sin 2\,N = 1$; that is, when $2\,N = \frac{\pi}{2}$, or when $N = \frac{\pi}{4}$ or 45°. In this case the line $\overline{AC}$ has the length $\frac{V^2}{4\,a}$. (This motion is illustrated by the path of a projectile or ball thrown obliquely upward. The distance $\overline{OB}$ is in this case the horizontal trajectory. Of course, the resistance of the air seriously influences these results. It is evident, too, that the horizontal trajectory is greater if the point of projection O is above the earth's surface, and the moving body is free to fall down to the surface.)

This problem of a projectile was first discussed by Galileo, 1638.

3. *A point is moving in a circle with constant speed.*—It has been shown already that in this case the acceleration is toward the centre along the radius drawn to the moving point at any instant, and has the numerical value $\frac{s^2}{r}$; where s is the value of the constant speed and r is the length of the radius.

This result can be expressed differently. Since r is the radius of the circle, the length of the circumference is $2\pi r$; so, if T is the time taken to complete one revolution, the speed $s = \frac{2\pi r}{T}$. Hence the acceleration may be written $\frac{4\pi^2 r}{T^2}$.

Or, if N is the number of complete revolutions in a unit of time, $N = \frac{1}{T}$, and the acceleration may be written $4\pi^2 r N^2$.

Again, since one complete revolution corresponds to an angle whose value is 2π, the angle turned through in a unit of time by the radius drawn to the moving point is $2\pi N$; so, writing this quantity h, $N = \frac{h}{2\pi}$, and the acceleration becomes rh^2. h is called the "angular speed": it evidently equals $\frac{s}{r}$; because s is the arc described in a unit of time by the radius whose length is r, and therefore $\frac{s}{r}$ is the value of the angle described in this time.

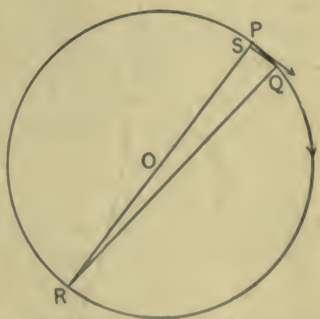


FIG. 22. — Uniform motion in a circle considered as a composition of uniform tangential motion and constant radial acceleration.

This problem may be solved in a different way which is due to Newton. Let the circle of motion be that shown in the cut.

Since the point moves in a circle, and since at any position, P, its velocity is along the tangent, a velocity must be added along the radius toward the centre in order to change the *direction* of motion. Draw a diameter PR

through P. Call the numerical value of this acceleration a, and that of the constant speed s. In an infinitesimal interval of time, t, the distance the point moves in the direction of the acceleration is $\frac{1}{2}at^2$; but the actual displacement is the chord of the arc whose length is st. Let this chord be given by $\overline{PQ}$ in the cut; then its projection on the diameter through P, $\overline{PS} = \frac{1}{2}at^2$; and in the limit $\overline{PQ} = st$, since the arc and chord coincide. But by geometry $\overline{PQ}^2 = \overline{PS} \times \overline{PR}$; or, calling the radius of the circle r, $s^2t^2 = \frac{1}{2}at^2 \times 2r$. That is, $a = \frac{s^2}{r}$.

4. *A point moves to and fro along a straight line, being the projection on any diameter of a point moving in a circle with constant speed.* (*Simple Harmonic Motion.*)

In the cut let P be the point moving in a circle with constant speed; let $\overline{AB}$ be any diameter; then Q is the projection of P on this; and, starting from any instant, as P describes a complete circumference, Q moves to and fro along the diameter, returning to its original position and *condition;* that is, at the end of this time, which is called a "period," the point Q is not alone in its original position, but is also moving in the same direction with the same speed and acceleration that it had at the beginning of the period. The motion of Q is called "simple harmonic."

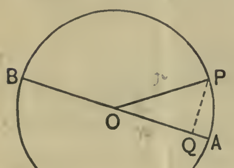


FIG. 23. — Harmonic motion: P moves uniformly in a circle, Q is its projection on any diameter.

The acceleration of Q is evidently the component of that of P resolved parallel to $\overline{AB}$; because Q has identically the same motion as that of P resolved in this direction. The acceleration of P is toward O, and has the numerical value rh^2, if r is the radius of the circle and h is the angular speed; its component parallel to $\overline{AB}$ is therefore $rh^2 \cos(POQ)$. If the displacement of Q with reference to O is called x, and if it is defined to be *positive* when drawn to the right, $\cos(POQ) = \frac{x}{r}$; and hence the numerical value of the acceleration of Q is h^2x, its direction being toward O, which is called the "origin." If we call all vectors positive when drawn toward the right, the acceleration is given, both in amount and in direction, by $-h^2x$, the minus sign indicating that when Q is on the right of O, *i.e.* when x has a positive value, the acceleration is negative, and therefore is toward the left, and when Q is on the left of O, *i.e.* when x has a negative value, the acceleration is positive and therefore is toward the right.

Since h is the constant angular speed of P, the period, or the time taken for a complete revolution, is $\frac{2\pi}{h}$. One half the extent of the swing of the vibrating point Q, *i.e.* the maximum value of the displacement x, is called the "amplitude" of the vibration. It therefore equals the length of the radius.

If there are two points moving around the same circle with the same speed, there will be two projected points in harmonic motion with the same period and the same amplitude; yet their motion is not the same at any instant, for their displacements, etc., are different, — one lags behind the other; they are said to differ in "phase." Thus, if one point is passing through the origin toward the right when the other is at the right-hand end of its swing, the difference in phase is a quarter of a period, from the standpoint of time; or, since one period corresponds to an angle 2π, a quarter of a period corresponds to an angle $\frac{\pi}{2}$, and so we call it a "difference of phase of $\frac{\pi}{2}$." Again, if one point is moving through the origin toward the right when the other is moving through the origin toward the left, the difference in phase is one half a period, or π.

Motions similar in every respect to that of the vibrating point Q are extremely common: any point of a stretched string vibrating in a simple manner, the end of a tuning fork, a body hanging from a spiral spring and set vibrating, etc. In all these cases the acceleration at any instant is found to obey the relation $-c^2x$, where x is the displacement at that instant of the vibrating point from its position when at rest, and c^2 is a constant quantity depending upon the nature of the vibrating system (c^2 is written for the constant quantity instead of c, for purposes of simplicity, as will appear immediately. Of course, C or any other symbol would serve equally as well, so far as its meaning is concerned).

We may, then, discuss the motion of a point in a straight line whose acceleration at any instant is $-c^2x$, where x is the displacement of the point from its origin and c^2 is a constant quantity. Such motion is "simple harmonic." Obviously, by comparison with the motion of the projected point Q in the problem just described, this harmonic motion is periodic, with the period $\frac{2\pi}{c}$. The amplitude is the maximum value of the displacement. Two harmonic motions of the same period may differ in phase.

The motion of the projected point Q in the original problem may be described in a different way: the value of its displacement may be stated instead of that of its acceleration.

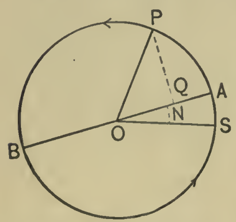


Fig. 24. — Harmonic motion: if $\overline{OQ} = x$, $\overline{OA} = R$, angular speed of $P = h$, $x = R\cos(ht - N)$.

As before, let P be a point moving in a circle of radius $\overline{OP}$ with a constant angular speed h, and let the motion of its projection Q on any diameter be considered. Call the displacement $\overline{OQ}$, x; the length of the radius, R; let time be counted from the instant at which the moving point P was at the point S, and call the angle (SOA), N. Then by trigonometry

$$x = R\cos(AOP) = R\cos(SOP - SOA),$$
$$= R\cos(SOP - N).$$

But h is the angular speed, and time is counted from the instant when P was at S; therefore the angle (SOP) equals ht, if t is the interval of time taken for the point to move from S to P. Hence

$$x = R\cos(ht - N).$$

In this expression, x is the displacement; the amplitude is R; the period is $\frac{2\pi}{h}$; and $(ht - N)$ is called "the phase." So two similar motions having the same periods, *i.e.* the same value of h, may differ in phase (they may have different values of N); and the "difference in phase" is $(ht - N_2) - (ht - N_1) = N_2 - N_1$. In words, the position in the circle of the moving points when one begins to count time is different in the two cases.

In general, therefore, motion of a point in a straight line whose displacement is given by the formula

$$x = A\cos(ct - N)$$

is harmonic; and in this expression the amplitude equals A, the period equals $\frac{2\pi}{c}$, and the phase equals $(ct - N)$.

Harmonic motions may be compounded by geometrical addition; and several special cases of importance physically will be discussed later.

The velocity of the point Q equals the component of the velocity of P resolved parallel to $\overline{BA}$. The linear velocity of P is along the tangent and has the value Rh; its component along $\overline{BA}$ is $-Rh\sin(AOP)$ or $-Rh\sin(ht - N)$, if velocities toward the right are called positive. So, in general, the velocity of the harmonic motion $x = A\cos(ct - N)$ is $-Ac\sin(ct - N)$; or, calling the period T, since $T = \frac{2\pi}{c}$, this equals $-A\frac{2\pi}{T}\sin(ct - N)$. Similarly, the acceleration of P is Rh^2 along $\overline{PO}$; then that of Q is $-Rh^2\cos(AOP)$ or $-Rh^2\cos(ht - N)$. This, therefore, equals $-h^2x$, as we saw before.

Rotation

As already stated, the name "rotation" is given the motion of a geometrical figure when, at any instant, each point of the figure moves in a circle ; the planes of these circles are all parallel, and their centres lie on a line called the "axis." It was shown further how to describe the angular motion around a fixed axis by means of two lines, one fixed in space, the other in the moving figure, but both lying in a plane perpendicular to the axis.

Angular Displacement. — In the cut, as before, let $\overline{AB}$ be a line fixed in space and $\overline{PQ}$ a line fixed in the rotating figure ; let O be the point where the axis cuts the plane. As the figure turns around the axis, the angle between $\overline{AB}$ and $\overline{PQ}$ varies: It is called the "angular displacement" at any instant of $\overline{PQ}$ with reference to $\overline{AB}$. To describe this displacement more fully, however, we must know the position of the axis and the "sense" in which the body is turning ; that is, whether, looking at the lines in the cut, $\overline{PQ}$ is rotating like the hands of a watch or in the opposite sense. These three ideas — the numerical value of the angle, the position of the axis, and the sense of the rotation — can all be represented by a vector placed so as to coincide with the axis ; for its position indicates that of the axis, its length can represent the value of the angle through which the figure turns, and its direction can be made by some agreement to indicate the sense of the rotation. The connection between the direction of the vector and the kind of rotation which is usually adopted is as follows : if an ordinary right-handed screw partially in a board is placed so as to coincide with the

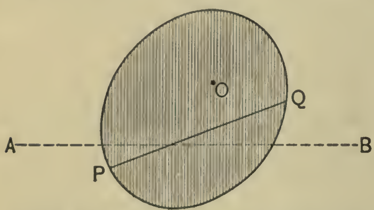


Fig. 25. — Angular displacement: O is trace of axis, $\overline{AB}$ is a line fixed in space, PQ is a line fixed in the figure.

vector and is turned in the sense of the rotation, the direction which it moves lengthwise into or out of the board is that given the vector. Thus, in the cut, if $\overline{AB}$ is the vector, and if an observer looks at a rotating figure in the direction from A toward B, the rotation is like that of the hands of a clock. If the arrow were in the opposite direction, the vector would indicate rotation in the opposite direction. This connection between the directions of the vector and the rotation is called the "right-handed screw relation." A vector located in a definite position in this manner is called a "rotor." It may be proved without difficulty that one can add two rotors geometrically: (1) if they lie in the same line, when the addition is therefore algebraic, because rotation in one sense is positive (+), and in the opposite is negative (−), and the two rotors are lines either in the same sense or in opposite senses; (2) if their axes lie in the same plane, *i.e.* if they meet at a point, and if the angular displacements to which they are proportional are infinitely small. (For this proof reference may be made to any treatise on Mechanics, such as that of Ziwet or Williamson.)

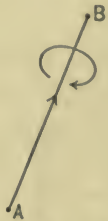


FIG. 26. — Right-handed screw relation between a vector $\overline{AB}$ and rotation.

Angular Velocity. — The rate of change of the angular displacement around a given axis is called the "angular velocity around that axis." The numerical value of this velocity is called the "angular speed"; so that in order to describe an angular velocity three things must be specified: the position of the axis, the sense of the rotation, the angular speed. If the angular motion is uniform, the value of the angular speed is that of the angular displacement in a unit of time. In

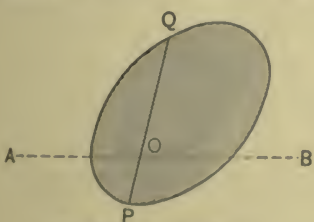


FIG. 27. — Angular Speed.

the cut let, as before, $\overline{AB}$ and $\overline{PQ}$ be the two fixed lines of reference, one in space and one in the figure; but in this case let them be drawn through the axis at O. Q is then any fixed point in the figure; call its distance from the axis, *i.e.* $\overline{OQ}$, r. If the axis is fixed in space, the length of the arc described by Q is evidently equal, by the definition of the value of an angle, to the product of r and the value of the angular displacement. It follows, then, that the *linear* speed of Q at any instant, *i.e.* the rate at which Q passes along the arc of its circle, equals the product of r by the value of the *angular* speed of that instant, *i.e.* the rate at which the radius describes the angle.

To return to the idea of angular velocity, it is evident that it can be represented by a rotor; and since an angular velocity is the rate of change of an angle, *i.e.* is the limit of the ratio of the value of a small angle to that of a corresponding small interval of time, it is proportional to an infinitely small angle, and therefore two rotors representing angular velocities may be added geometrically, if the two axes lie in the same plane.

Illustrations of angular velocities are common; a few may be described as follows: that of a flywheel is given by a line having a definite sense and length, coinciding with the central line of the axle; that of a door or gate by a limited portion of a vertical line drawn so as to coincide with the central line of the hinges; that of a cylindrical barrel rolling down an inclined plane by a line coinciding with the line of contact between the cylinder and the plane — in this case the axis is moving parallel to itself down the plane. As a hoop rolls along a floor, the rotor giving its angular velocity is a horizontal line *perpendicular* to its plane. If the hoop were at rest in an upright position, a sidewise push at the top would give it a rotation around a horizontal axis *in its own plane;* therefore, if a sidewise push is given a rolling hoop at its top, the rotor of the resulting motion is the geometrical sum of the two separate rotors, and is in a horizontal plane but in a different direction from either — this explains why pushing a rolling hoop sidewise at the top changes the direction of its path. (The independent action of two or more forces is assumed again.)

Angular Acceleration. — The rate of change of the angular velocity is called the angular acceleration. There are two special cases: in one, the direction of the axis remains fixed in space, while the angular speed varies, *e.g.* a door when it is opened or closed, a grindstone when it is set in motion or is gradually stopped, etc.; in the other, the angular speed remains constant and the direction of the axis changes, *e.g.* a spinning top whose axis is not vertical, a rolling hoop turning a corner, etc. In the general case, of course, both the angular speed and the direction of the axis change. In the case of rotation around a fixed axis there is evidently a simple connection between the angular acceleration and the linear acceleration of any point of the figure. If A is the angular acceleration and a the linear acceleration of a point at a distance r from the axis, it follows at once from the definition of the value of an angle that $a = rA$. (See page 47, where a similar formula for the velocities is proved.)

There is also harmonic motion of rotation around a fixed axis, analogous to that of translation; it is illustrated by vibrations of the balance wheel of a watch, by those of a clock's pendulum, etc. It is defined as follows: if the value of the angular displacement in a particular sense is called N, harmonic motion of rotation is such that the angular acceleration equals $-c^2N$, where c^2 is a constant, depending upon the vibrating system. The amplitude is the maximum value of N; the period may be proved to be $\frac{2\pi}{c}$; and two harmonic vibrations may differ in phase. It may be shown, further, that this definition is equivalent to saying that the angular displacement equals $A \cos (ct - M)$, where A is the amplitude; $\frac{2\pi}{c}$, the period; and $(ct - M)$, the phase.

General Remarks. — The displacement of a point in any direction is independent of a displacement in a direction at right angles to this, *e.g.* a man walking northward does not move to the east or to the west; and, since it is possible to draw from any point three lines that are mutually perpendicular, like the three lines meeting at the corner of a room, a point, and therefore a solid figure, may have three independent directions of translation. Similarly, rotation around any axis is independent of rotation around an axis perpendicular to it; and therefore a solid figure has three independent modes of rotation. So, in general, a solid figure may be displaced in any one of six independent ways — three of translation and three of rotation; it is said to have six "degrees of freedom." Freedom of motion may of course be hampered by various constraints; thus a figure in a straight groove has only one degree of freedom of translation, a figure turning on a fixed axis has only one degree of freedom of rotation.

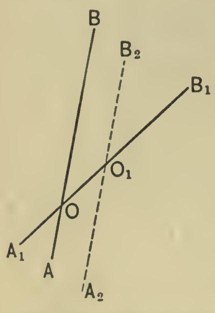


FIG. 28.—The rotation of $\overline{AB}$ about an axis at O into the position $\overline{A_1B_1}$ is equivalent to a translation from $\overline{AB}$ to $\overline{A_2B_2}$ and a rotation about an axis at O_1.

It is always possible to produce a given displacement of a figure in several ways. Thus, if a figure like a wire, $\overline{AB}$ in the cut, is displaced by a rotation around O into the position $\overline{A_1B_1}$, this same final position may be obtained by a translation from $\overline{AB}$ to $\overline{A_2B_2}$ and by a rotation around O_1 from $\overline{A_2B_2}$ to $\overline{A_1B_1}$. It is often a matter of importance to select the simplest mode of displacement to meet the requirements of a given problem.

There are several simple theorems which may be stated without proof, although such proofs are not difficult. The displacement of a plane figure in a plane into any other position can always be produced by a single rotation around an axis perpendicular to the plane; any change of position of a solid figure, one of whose points is fixed, may be produced by a single rotation around an axis through the fixed point; *any* change of position of a solid figure may be produced by a translation in a definite direction and a rotation around this direction as an axis, *i.e.* by a "screw motion."

Analogy between Translation and Rotation. — It may be useful to arrange in parallel columns the properties of translation and rotation that correspond to one another:

Translation	*Rotation*
A point moving in a straight line.	A figure turning on a fixed axis.
a. line of motion	*a.* axis of rotation
b. displacement	*b.* angular displacement
c. linear speed	*c.* angular speed
d. linear acceleration	*d.* angular acceleration
e. harmonic motion	*e.* harmonic motion
acceleration $= -c^2x$,	angular acceleration $= -c^2N$,
or displacement	or angular displacement
$x = A\cos(ct - C)$,	$N = A\cos(ct - C)$,
$\text{period} = \frac{2\pi}{c}$.	$\text{period} = \frac{2\pi}{c}$.
A point moving in a curved line; *i.e.* the direction of its motion is changing.	A figure, the direction of whose axis of rotation is changing.
Linear speed is altered by adding a velocity in the same direction.	Angular speed is altered by adding an angular velocity around the same axis.
Direction of motion is altered by adding a velocity in a different direction.	The direction of the axis is altered by adding an angular velocity around a different axis, the two lying in the same plane.

CHAPTER II

DYNAMICS

Introduction. — In describing our fundamental ideas of nature, emphasis was laid upon the conditions under which we feel the force sensation. Among the most important of these are the following: when we support a body free from the earth, when we change the size or shape of a body, when we alter the motion of a body. These conditions may be brought about by the action of material bodies, instead of by our muscles; and, when this is the case, we say "a force is acting," one body "is exerting a force on another," etc. We do not, however, mean to imply the existence of a *thing*, but of a condition. We must now devise some method of measuring forces; and we shall begin by discussing certain illustrations.

Consider a vertical wire whose upper end is clamped to some support and at whose lower end hangs a heavy body. The wire exerts an upward force on the body, and in the process it is stretched, its molecules are displaced from their ordinary position; the wire also exerts a downward force on its support; this, in turn, exerts an upward force on the wire, and in doing so it is bent and its molecules are slightly displaced; the support must rest upon the earth in some manner, and here again enter two forces. If the wire is broken, the heavy body falls with an acceleration toward the earth, thus showing that there is a downward force on the heavy body — even when it is hanging from the wire — due to the earth. In the case of the stretched wire, the bent support, etc., we see that the force is associated with the deformation of a body, that is, with the displacement of its molecules; in the case of the heavy body and the earth there is—to our eyes—no connecting mechanism and no deformation of matter; in both cases the presence and "action" of a material body is essential for the production of the force. Similarly, if two moving bodies strike each other, the motion of each is changed; so each exerts a force on the other.

We cannot directly investigate the motion of molecules, nor can we understand or even describe the action between the falling body and the earth, partly owing to the great size of the latter. For these reasons we shall first discuss, as an illustration of forces, the accelerations of material bodies of ordinary size, when these are produced by the interaction of the bodies themselves. If two billiard balls meet, the velocities of both change, *i.e.* they are accelerated; if a magnet and a small piece of iron are suspended by strings at the same level, each moves toward the other with an acceleration; if a man stands on a box which rests on a smooth floor, and jumps off sidewise, the box moves in a direction opposite to that in which the man jumps; etc. It is a familiar fact, too, that if one of the moving bodies is much heavier than the other, its acceleration is much less; and, as the bodies are varied, there is apparently a connection between some property of the body and its acceleration.

Fundamental Principles

1. **Principle of Inertia.**—In none of the cases described above is there an acceleration of one of the bodies without there being at the same time an acceleration of the other. So we make, as the fundamental assumption in regard to forces, that *the acceleration of a body depends upon its position with reference to neighboring bodies and upon their velocities.* (To measure these velocities and accelerations, some suitable geometrical figure of reference must be selected.) There is no way of *proving* this assumption or the following ones; but all observations are in accord with them.

2. **Principle of Independence of Action of Forces.**— Again, we shall assume that, *when a body is under the influence of several forces, the action of each one is independent of the actions of the others.*

3. **Principle of Action and Reaction. Definition of "Mass."** — Then, if we have an isolated system of two bodies, each will have a linear acceleration; and, in order to speak definitely, we shall consider the bodies as being so small that they occupy *points*. Such bodies are called "particles." We assume that *their accelerations are in the straight line joining the particles, but in opposite directions, and that the ratio of these accelerations is a constant quantity.* We can, therefore, assign a number to each body such that, if m_1 and m_2 are these numbers, and a_1 and a_2 are the accelerations, $m_1a_1 = -m_2a_2$. Similarly, if we have a third particle, we can assign a number to it by allowing it to "act" upon the first particle, using m_1 as its number, or upon the second one, using m_2 for it. Experiments prove that the numbers thus obtained for the third particle are the same. Therefore, if we adopt an arbitrary number for any one particle, the numbers obtained for all other particles are definite. These numbers are the values of what is called the "mass" of a body. (The system of masses in ordinary use will be described presently.)

4. **Definition of "Force."** — When a particle of mass m has an acceleration a, the product ma is a vector with a definite value and direction; and it is *defined to be the value of* what we have called *the force.* Thus, in the case of the two particles, there are two equal and opposite forces; and we say that the force of a particle A upon a particle B is equal numerically but opposite in direction to that of B upon A, or that "action and reaction are equal and opposite."

Since, then, $F = ma$, calling the value of force F, $a = \frac{F}{m}$, or when a given force acts upon a particle, the acceleration is in the direction of the force and its numerical value varies inversely as the mass of the particle. That is, if m is large, a is small; and conversely; so m measures the *inertia* of the particle with reference to translation.

5. **Centre of Mass.** — In practice we cannot obtain particles, for all material bodies occupy finite volumes; and so we cannot apply the definition of a force directly. Further, under the action of a force a body, as a rule, has both linear and angular acceleration; for instance, if a rod lying on a smooth table is struck at some point near the end by a ball rolling on the table the rod will move in the direction of the blow, and it will also rotate. But we can prove, as will be shown presently, that, if we assume *that a material body may be considered as built up of particles*, there is a geometrical point connected with the body that has the same linear acceleration when the body is acted on by a force, as if this acted on a *particle* whose mass equaled that of the body. This point is called the "centre of mass" of the body.

A few illustrations of this fact may be interesting. If a particle were to fall freely from a projecting height, its path would be vertically down and it would have a constant acceleration. Similarly, when a plank or a chair falls without striking an obstacle, there is a point connected with its figure that moves vertically down with a constant acceleration, however much the body turns. If a particle were struck a blow, it would move in the direction of the force. Similarly, if a rod or a chair is free to move and is struck a blow, there is always a point connected with it that moves in a straight line in the direction of the blow, although rotation may be produced also.

Consequently, whenever we give illustrations of forces, etc., and refer to the linear acceleration "of the body," the linear acceleration of its centre of mass is meant.

Measurement of Mass

The Theory and Practice of the Measurement. — The mass of a particle has been defined in terms of the action of one particle on another; or, using the definition of a force, we may say that, if two particles receive acceleration owing to the action of the same force, their masses vary inversely as their accelerations. For, if m_1 and m_2 are the masses and a_1 and a_2 the accelerations, the value of the force F must satisfy the two equations:

$$F = m_1 a_1,$$

$$F = m_2 a_2.$$

Hence
$$m_1 a_1 = m_2 a_2,$$

or
$$m_1 : m_2 = \frac{1}{a_1} : \frac{1}{a_2}.$$

In order to give a number, then, to the mass of any body, some piece of matter must be arbitrarily selected to which the number 1 is given; and the mass of the body must be compared with this "unit mass" by the experiment just described, *assuming*, what will be proved presently, the property of the centre of mass as just stated.

One manner in which this comparison can be *imagined* done — although it is not practicable — is to place the two bodies on a *smooth*, *horizontal* table, so that gravity has no action and that the motion is not affected by friction; and, attaching to each in turn a spiral spring, to drag it in a horizontal direction in such a manner and at such a rate that the spring is elongated by the same amount. The spring will therefore exert the same force on the two bodies in turn; and, if a_1 and a_2 are the measured accelerations, $m_1 a_1 = m_2 a_2$. (We assume in this that when a spring is stretched a definite amount it exerts a definite force, regardless of the time or place.)

Another method is to make use of a fact which will be discussed later, viz., the acceleration with which a body falls toward the earth — due allowance being made for the effect of the atmosphere — is the same at any one point on the earth's surface for all bodies. A body, then, is acted on by a force due to the earth, which we call "weight," whose value, in accordance with the definition of force, equals the product of the values of the mass of the body and the acceleration which it would have if falling freely. Call the mass m and the acceleration g; then the weight of the body is mg. It follows, then, that if the body is supported from the end of a spiral spring, — an ordinary "spring balance" such as is shown in the cut, — the spring will be elongated and therefore exerts an upward force on the body. Since the body is at rest, — relatively to the earth and the spring, — this upward force must equal mg. If two bodies when suspended in turn from the same spiral spring produce the same elongation, their weights must be the same, proper allowance being made for the buoyancy of the air (see page 165); *i.e.* calling the masses m_1 and m_2, $m_1 g = m_2 g$, and hence $m_1 = m_2$, or their masses are the same. So, if a standard body is chosen whose mass is called 1, another body can be taken which elongates a spring slightly more than does the former when suspended by it; and then, by chipping or filing off minute quantities, it may be so altered as to produce the same elongation as does the standard body. Thus we obtain a second body of mass 1. Similarly, we can obtain a body of mass 2 by first suspending together the two bodies of unit mass, noting the elongation, and then determining a third body which produces this same elongation; etc. To obtain a body

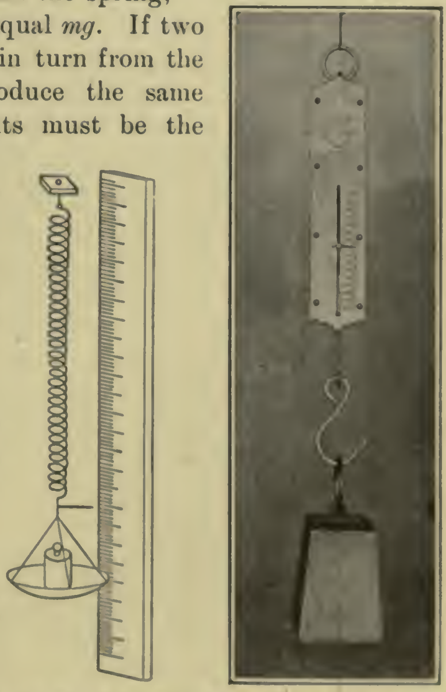


FIG. 29. — Spring Balance used for measuring weights.

whose mass is $\frac{1}{2}$, we must make two bodies of equal mass — as shown by producing the same elongation — which when suspended together will produce the same elongation as does the body of unit mass. Proceeding in this manner, we may obtain a "set" of bodies all of whose masses are known in terms of the standard. Then to obtain the mass of any body, it may be suspended from the spring and that combination of bodies from this set determined which will produce the same elongation. (Other and more accurate methods are used in practice, as will be shown later in speaking of the "chemical balance.")

In using a spiral spring for ordinary purposes a different method is followed from that just described. Experiments show that the elongation of a spiral spring is proportional to the stretching force, most approximately; and a divided scale may be attached to the frame carrying the spring, the readings on which are proportional to the elongation of the spring.

Then if a body whose mass is m_1 produces an elongation h_1, and one whose mass is unknown but which may be called m produces an elongation h, $h_1 : h = m_1g : mg$, since these last are the forces. Therefore, $h_1 : h = m_1 : m$; or $m = \frac{m_1h}{h_1}$. As a rule the instrument maker divides and marks the scale so that it gives the values of the masses directly on some known system; that is, when a body of mass 1 is suspended, the pointer which marks the elongation stands at division 1 on the scale, etc.

Mass and Weight. — It should be carefully noted that this method of comparing the *masses* of two bodies is in reality one which compares their *weights;* but, since g at any one locality is the same for all kinds and quantities of matter, the weight of a body is proportional to its mass. In other words, two bodies that have the same weight at any one point on the earth's surface also have equal masses. This fact is not self-evident; for the weight of a body is dependent upon the presence and proximity of the earth; its mass, upon the acceleration it would receive from a definite force *anywhere in the universe.* The former is a variable quantity, as we shall see, because it is different for the same body at different latitudes on the earth and at different heights above sea level; while the latter is, to the best of our knowledge, an unvarying constant quantity for a given body. There is no more *a priori* reason for believing that two bodies which have equal masses have equal weights, than for believing that they have equal volumes, which is obviously not true in general. It is a question to be settled by experiment, and depends upon proving the constancy of g for all kinds and amounts of matter at any one point on the earth's surface. This was first shown by Galileo in 1590, who allowed two bodies of different weights to fall from the top of the Leaning Tower of Pisa, and observed that they reached the ground at the same instant, thus proving that they had the same acceleration. (Galileo, however, had no conception of the property of mass, and performed his experiment with a different object.) Newton was the first to realize the fact that a material body had a property which we have called "mass"; and he devised a most ingenious experiment for the purpose of learning whether g was the same for all bodies. His method depends upon the use of pendulums and will be described later. It has been used also by Bessel; and all experiments prove that we may regard g as a constant at any one point.

Conservation of Matter. — Few quantities can be measured with the exactness of the weight, and hence mass, of a body, by the use of a "chemical balance" so called. One of the questions first investigated by means of it was as to the constancy of mass of a body, or of two bodies when combined in any way. From what has been said in the introduction, it is evident that a piece of matter can undergo various changes; it may have its shape, its size, its temperature, etc., altered; it may explode into fragments; it may be "electrified" or "magnetized"; it may be melted if it is a solid, or evaporated if it is a liquid, or *vice versa*. Similarly, if two pieces of matter are brought together, they may stick to each other like putty and glass; or they may unite to produce new substances, like a piece of coal burning, a process in which the oxygen in the air unites with the carbon in coal to form a new gas, called "carbon dioxide." In all these changes, however, there is, so far as we know, *absolutely no alteration in the total mass of the body or bodies concerned*. This fact is sometimes called the "Principle of the Conservation of Matter"; or, more properly, the "conservation of mass."

The Unit of Mass. — The standard body whose mass forms the basis of the accepted system of physical units is a piece of platinum which is kept in Paris, and which is called the *Kilogramme des Archives*. It was officially adopted in 1799, at the same time as the metre bar. Its mass is called a kilogram (Kg.); and a body whose mass is one thousandth of this is said to have a mass of 1 "gram" (g.). A mass of one tenth of a gram is a "decigram"; one of one thousandth of a gram is a "milligram," etc.

When the kilogram was originally made it was designed to have a mass equal to that of 1000 cu. cm. of pure water at a temperature of 4° C., because under those conditions the water is more dense than at any other temperature. (The temperature must be specified, because, as it changes, the volume of a given quantity of matter varies.) More exact experiments have, however, shown that this relation is not quite exact. In fact, the mass of 1000 cu. cm. of pure water at 4° C. is about 999.96 g.

In England and the United States the commercial Unit Mass is that of a piece of platinum kept at the Standards Office at Westminster, marked "P. S. 1844 1 lb.," and called the "Imperial Avoirdupois Pound." It has been determined by experiment that the number of grams in one pound is 453.5924277. This unit is subdivided in such a way that 16 ounces equal one pound.

C.G.S. System. — In all scientific work the units in terms of which lengths, intervals of time, and masses are expressed are the centimetre, the gram, and the mean solar second. This is called the "C.G.S. system."

Force $F = ma$

The Unit of Force. — The unit of force on this C.G.S. system is that which corresponds to the product ma being unity; that is, a force which produces in a body whose mass is 1 g. an acceleration of 1 cm. per second each second (or an acceleration 2 in a body where mass is $\frac{1}{2}$, etc.). This force is called a "dyne." A force of one million dynes, *i.e.* 10^6 dynes, is called a "megadyne."

As estimated by our muscles, a dyne is extremely small; for, as we can find by experiment, the value of the acceleration of a falling body is not far from 980 on the C.G.S. system, so the weight of a body whose mass is a *milligram* is the product of 0.001 and 980 or 0.980. Consequently, we feel approximately the force of 1 dyne when we hold a milligram "weight" in our hands. On the pound-foot-second system the unit force is one which gives in 1 sec. an acceleration of 1 ft. per second to a particle whose mass is 1 lb. (This unit therefore equals 13,825 dynes approximately.) Other unit forces often used are the "weight of a body whose mass is 1 g.," or of one whose mass is 1 lb.: these units have the great disadvantage of being variable, owing to variation in "g."

Force Effects. — In defining the numerical value of a force, we made use of the idea of a particle and of the "force-effect" acceleration. But the production of acceleration is, as we have seen, not the only effect of a force. This is evident from the formula of definition itself, which has the form $f = ma$, if f represents the force, m the mass of the particle, and a the acceleration. For we have assumed that forces act independently; that is, they are vector quantities. So, in the formula, f may be the geometrical sum of several forces. In particular, if $a = 0$, $f = 0$; but this does not mean necessarily that there is no force; it may mean that there are two equal and opposite forces acting on the particle.

Thus, if a particle is suspended "at rest" at the end of a vertical wire, it has no acceleration, owing to the fact that the two forces acting on it, its weight and the tension in the wire, are equal and opposite. It may also happen that when a particle is under the action of two opposing forces, its acceleration is not zero; in this case one force is greater than the other. So, in general, we may say that a second effect of a force is to overcome or neutralize, more or less completely, another force. (If we speak of matter when being accelerated as offering an opposing "force of inertia" equal to ma, we may say that the effect of a force is always to overcome another force.)

This second effect of a force is illustrated in many ways: it requires a force to stretch a spring, to bend a stick, to twist a wire, to push a body over a rough table, to drive in a nail, etc. As the properties of matter are gradually better understood, we hope to explain all these effects in terms of the acceleration of particles of matter. We can do this in certain cases already, as we shall see later. (See Kinetic Theory of Matter, Chapter IV.)

Similarly, as we have shown before, forces may be produced in various ways. If a stretched spring or wire or cord is fastened to a body, it will be accelerated unless there is an opposing action; if a moving body strikes another, each exerts a force on the other, etc.

Measurement of a Force. — We make use of some one of these force effects in order to measure a force in practice. We know that a body which is supported at rest free from the earth must be acted upon by an upward force whose numerical value is mg, where m is its mass and g is the acceleration of a body falling freely in a vacuum under the influence of gravity. We have shown how to measure m, and methods will be described shortly for obtaining the value of g. Thus, we can observe how much a spring elongates under the stretching action of different bodies; and, *assuming* that the spring does not change during the operations, we may thus obtain readings on a scale attached to the spring, which correspond to known forces. This process is called "calibration" of the spring. Then, to measure any force, we can observe how much it elongates the spring. (If we wish to exert a known force in a definite direction upon a body, we can attach one end of a cord to the body and the other to a spiral spring, and then pull the farther end of the spring in the desired direction until it elongates the proper amount.) Other elastic bodies may be calibrated in a similar manner and used for the measurement of forces.

Linear Momentum. — The general formula for the value of a force, $F = ma$, may be expressed in a different manner. Since the acceleration is the rate of change of the velocity, the value of the force is the product of the mass by this rate of change; and, since the mass is a constant for a given particle, the force equals the rate of change of the product mv, where v is the velocity. This product is called the "linear momentum" of the particle; so the force equals the rate of change of the linear momentum.

If the acceleration is constant, *i.e.* if the speed is changing at a uniform rate but the direction of motion is unchanged, writing v_2 and v_1 as the velocities at the end and the beginning of an interval of time t, the rate of change of the velocity is $\frac{v_2 - v_1}{t}$. Hence, we may write $F = m\frac{v_2 - v_1}{t}$, or $Ft = m(v_2 - v_1)$. The product Ft is called the "impulse" of the force. If a sudden blow is struck the particle, its momentum in the direction of the force will be changed; and the amount of this change measures the impulse of the blow.

If a particle is moving with a constant velocity, there is no resultant force acting; but to produce a change in the velocity a force is required. A useful illustration of the formula $F = ma$ is furnished by suspending a heavy body by a long cord and attaching a thread to it also, so that it may be drawn sidewise. If it is pulled slowly, a small force is required; if it is jerked suddenly, the force required may be so great that the thread breaks.

Illustrations of Forces. — There are two types of acceleration, one when the speed changes and the direction does not, the other when the direction changes and the speed does not; and corresponding to each of these is a definite type of force. In the one the force produces a change in speed; in the other, a change in direction of motion.

a. Rectilinear force. — Thus, if we observe that a particle is moving in a straight line with varying speed, we know that there is at any instant a force acting in the direction of the line, whose numerical value is ma, where m and a are the values of the mass of the particle and its acceleration at that instant. Illustrations of this type of force are afforded by falling bodies, $F = mg$; by an elevator when rising or falling at a varying rate; a railway train "getting up" speed, etc. When an elevator is rising at a uni-

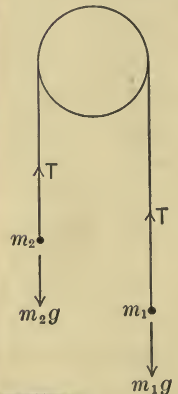


FIG. 30. — Atwood's machine: m_1g and m_2g are the forces due to gravity. T is the tension in the string.

form rate, the upward force exactly balances the downward force of weight (and friction); but, if it is *accelerated*, an additional force is required. A similar statement is true of the train.

If a particle is given a velocity obliquely upward, it will have the path of a parabola as described on page 46, for it will retain a constant horizontal velocity, if we neglect the action of the air; and it will be under the action of a constant downward force. This is an illustration of the independence of two motions, one of uniform velocity, the other of uniform acceleration.

Again, let two particles whose masses are m_1 and m_2 hang from the two ends of a perfectly flexible inextensible cord which passes over a pulley; let us suppose that there is no friction, and let us neglect for the time being the mass of the cord and the inertia of rotation of the pulley. There are two forces acting on each particle: on the one whose mass is m_1 there is a force downward owing to gravity and equal to m_1g, and an upward one due to the tension in the string whose value may be written T, hence the total downward force is $m_1g - T$; on the other particle, the total downward force is, similarly, $m_2g - T$, because the force due to the tension in the string must be the same at both ends if there is no friction and it is inextensible. The *downward* acceleration on one particle must, however, equal the *upward* acceleration of the other, since the string is inextensible; the downward acceleration of the first particle is, however, $\frac{m_1g - T}{m_1}$; and the *upward* one of the other is $\frac{T - m_2g}{m_2}$. Calling this acceleration a, we have, then,

$$a = \frac{T - m_2g}{m_2} = \frac{m_1g - T}{m_1},$$

or, eliminating T,

$$a = \frac{(m_1 - m_2)g}{m_1 + m_2}.$$

Thus, if $m_1 > m_2$, a is positive, and the first particle will descend with a constant acceleration; if $m_1 < m_2$, a is negative, and the first particle will ascend.

The tension in the string is given at once from the formula, viz.,

$$T = m_2g + m_2a,$$

or

$$T = m_1g - m_1a.$$

This shows that in the case of an ascending particle, viz., the one whose mass is m_2, if a is positive, the tension is greater than the weight by an amount m_2a, while in the case of the descending particle it is less than its weight.

The same formula for the acceleration may be obtained in a different way; the total force acting on the system in such a direction as to give the first particle an acceleration downward is $m_1g - m_2g$, the total mass having the acceleration is $m_1 + m_2$, hence the acceleration equals $\frac{m_1g - m_2g}{m_1 + m_2}$ or $\frac{m_1 - m_2}{m_1 + m_2} \cdot g$.

(This apparatus is called Atwood's machine.)

An illustration of the independent action of forces is afforded by the slipping of a body down a rough inclined plane board. Let, at any instant, the body whose mass is m be at P. There are three forces: gravity, acting vertically down, whose value is mg; friction, acting parallel to the plane in such a direction as to oppose the motion, whose value may be called F; and the force of reaction due to the board, opposing the force with which the body presses on it, which is perpendicular to the plane of the board and upward, and whose value may be written R. If N is the angle the inclined plane makes with the level surface of the earth, the component of mg parallel to the plane is $mg \sin N$ and its direction is *down* the plane; F is parallel to the plane, but directed upward; R has no component parallel to the plane, because it is perpendicular to it. The total force down the plane is, then,

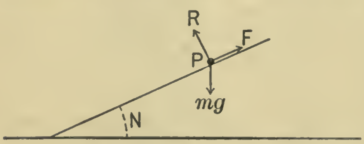


FIG. 31. — Motion down a rough inclined plane.

$$mg \sin N - F,$$

and so the acceleration is this divided by m; *i.e.*

$$\frac{mg \sin N - F}{m}.$$

b. Centrifugal force. — In order to change the *direction* of motion of a particle a force must be applied at right angles to this; and conversely, if a particle is moving in a curved path under the action of a certain force, and the force is removed, the particle will continue to move with a constant velocity in the direction of the tangent to the curve at the point where it was at the instant the force ceased. If a particle is moving with constant speed in a curved path which lies in a plane, the acceleration at any instant is toward the centre of curvature of the path and has for its numerical value $\frac{s^2}{r}$, where s is the speed and r is the radius of curvature. (Whatever the path is, a circle can always be drawn which will coincide with the curved path at any point, and its centre and radius are called the "centre and radius of curvature" at that point.) There must then be a force whose direction is perpendicular to the line of motion and whose numerical value is $\frac{ms^2}{r}$, in order to make the particle whose mass is m and speed s change its direction of motion and move in a circle of radius r. This force, $F = \frac{ms^2}{r}$, is called "centrifugal force." The less the force, the less the change in direction, *i.e.* the greater the value of r; or, if the speed is increased, the radius must be also, unless the force is increased; and so, when a particle is revolving in a circle, if the force toward the centre is diminished, or, if the speed is increased, the particle moves farther away from the centre, if such motion is possible, so as to have a larger radius of motion. This fact is sometimes described by saying "a particle revolving around a centre tends to move as far away from it as possible." Thus, a wet mop may be freed from the water by revolving it rapidly; clothes may be dried by inclosing them in a perforated cylinder which is made to rotate rapidly; etc. The force that is required to hold a particle in a circle varies directly as its mass; and, if the force applied is greater than $\frac{ms^2}{r}$, the particle will move toward the centre, while, if it is less than this, the particle will move away from the centre. Therefore, if an emulsion of two liquids, one more dense than the other, *e.g.* cream in milk, is put into a rapidly rotating cylinder, the heavier of the two — milk — will go to the outside wall, while the lighter — cream — will come closer to the axle.

c. Pendulum. — If both the speed and the direction of motion are changing, the force must be oblique to the latter, *i.e.* it must have a component in the direction of motion and one at right angles to it. One particular case of this kind of force is furnished by the action of the earth on a "simple pendulum," *i.e.* on a particle of matter suspended by a *massless* cord from a fixed support so that it is free to move in a vertical plane. When at rest, it hangs in a vertical direction; but, if disturbed slightly, it will make vibrations. (Of course a simple pendulum cannot be made, but we can approximate to one by using a small but heavy bob and a very fine wire to support it.) Let $\overline{OM}$ be a vertical line through the point of support O; let $\overline{OP}$ be the position of the pendulum at any instant (not necessarily at the end of the swing); let a circle be drawn with radius $\overline{OP}$, thus indicating the path of the particle; draw through P two lines: one vertically down, to indicate the direction of the force of gravity; and one tangent to the circle. Call the mass of the particle m, and the length of the pendulum l. There are two forces acting on the particle; its weight mg vertically down, and the *tension* in the supporting cord acting along the line $\overline{PO}$. The latter has, however, no effect on the speed, serving merely to change the direction of motion; similarly, only that component of mg which is along the tangent has any influence on the speed. This component has the value $mg \sin N$, where N is the angle (MOP). Let us suppose the arc of vibration is made so small that its chord coincides with it; in this

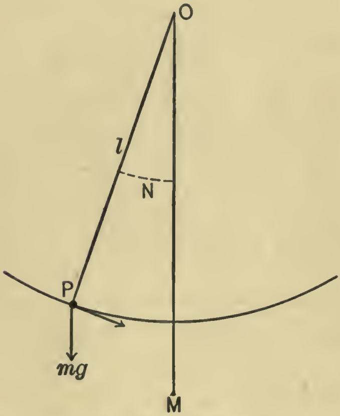


Fig. 32. — Simple pendulum.

limit the sine of an angle equals the angle itself (see page 25); and so the force at any instant along the tangent to the curve — which in the limit is nearly horizontal — has the value mgN. If the displacement of the particle along the arc measured from its lowest point is x, $N = \frac{x}{l}$; and therefore the force has the value $mg\frac{x}{l}$. Its direction is opposite to the displacement; and therefore its value must be written $-mg\frac{x}{l}$. The acceleration is, then, $-\frac{g}{l}x$. It follows that the motion of a particle along its infinitesimal arc is harmonic; and its period is $2\pi\sqrt{\frac{l}{g}}$. (See page 50.)

A method is thus offered for the measurement of g at any point, and therefore for the investigation of the question as to whether it is the same for all kinds of matter and for all amounts.

That portion of the above formula which states that the period of a pendulum varies as the square root of its length was deduced by Galileo as early as 1638.

Parallelogram and Triangle of Forces. — Another method of stating that forces act independently is to say that they can be added geometrically like vectors; and accordingly this theorem is sometimes referred to as the "parallelogram of forces." For if the two forces acting on a particle are represented by $\overline{PQ}$ and $\overline{PR}$ in the cut, their geometrical sum is represented by $\overline{PS}$, the diagonal of the parallelogram which has $\overline{PR}$ and $\overline{PQ}$ for two adjacent sides.

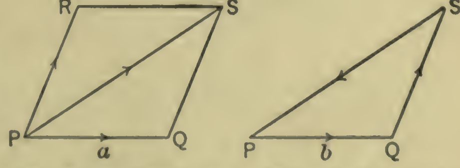


FIG. 33. — *a*. "Parallelogram of Forces": the geometrical sum of $\overline{PQ}$ and $\overline{PR}$ is $\overline{PS}$.
b. "Triangle of Forces": the geometrical sum of $\overline{PQ}$, $\overline{QS}$, and $\overline{SP}$ is zero.

It is evident that if the particle is acted on by three forces,

$\overline{PR}$, $\overline{PQ}$, and one equal but opposite to $\overline{PS}$, their geometrical sum is zero; for, adding these lines geometrically, a closed triangle is formed. Conversely, if a particle under the action of three forces has no acceleration, they will, if added geometrically, form a closed triangle. This theorem is called the "triangle of forces."

This principle was first stated, for a special case, however, by Stevin, as early as 1605.

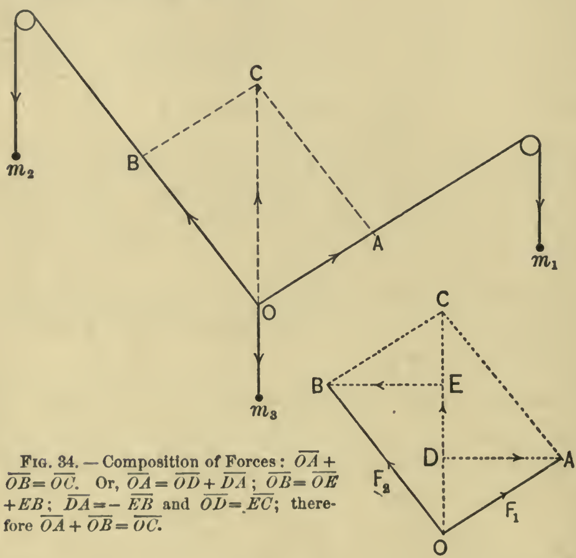


FIG. 34. — Composition of Forces: $\overline{OA}+\overline{OB}=\overline{OC}$. Or, $\overline{OA}=\overline{OD}+\overline{DA}$; $\overline{OB}=\overline{OE}+\overline{EB}$; $\overline{DA}=-\overline{EB}$ and $\overline{OD}=\overline{EC}$; therefore $\overline{OA}+\overline{OB}=\overline{OC}$.

The composition of forces is illustrated by the motion of a boat that is being rowed across a river, there being two forces, one due to the oars, the other to the current; and by many other similar motions.

If a cord carrying at its two ends particles whose masses are m_1 and m_2 is supported by two pulleys, and if a third particle of mass m_3 is attached to it at some point between the pulleys, as shown in the cut, the system will come to rest under the action of gravity in some definite position. There are now three forces acting at the point O: m_3g, vertically down; m_1g_1 in the direction $\overline{OA}$, because all that the pulley does is to change the direction of the force of the earth on the first particle; m_2g, in the direction $\overline{OB}$. If $\overline{OA}$ is a length proportional to the product m_1g, and $\overline{OB}$ is a length proportional to m_2g, their geometrical sum $\overline{OC}$ will be a vertical line proportional to m_3g. (Another mode of considering the geometrical sum of two vectors $\overline{OA}$ and $\overline{OB}$ is to see that, if $\overline{OA}$ is resolved into two vectors $\overline{OD}$ and $\overline{DA}$, where $\overline{OD}$ is along the diagonal $\overline{OC}$ and $\overline{DA}$ is perpendicular to it, and if $\overline{OB}$ is resolved in a similar manner into $\overline{OE}$ and $\overline{EB}$, $\overline{DA} = -\overline{EB}$ and $\overline{OD} + \overline{OE} = \overline{OC}$.)

Action and Reaction. — In the case of the interaction of two particles, the force exerted by one on the other is equal and opposite to that exerted by the latter on the former, and is in the straight line joining them. That is, if m_1 and m_2 are the masses of the particles, and a_1 and a_2 their accelerations owing to their mutual action, these accelerations are in the line joining the particles, and their numerical values are such that $m_1a_1 = -m_2a_2$. This law may be expressed in terms of the linear momenta. It may be written $m_1a_1 + m_2a_2 = 0$; and this is equivalent to saying that the rate of change of the sum $(m_1v_1 + m_2v_2)$ is zero, where v_1 and v_2 are the components at any instant of the velocities of the two particles *in the same direction along the line joining them.* If the rate of change of a quantity is zero, the value of the quantity itself must remain unchanged. So we may write the formula in the form: $m_1v_1 + m_2v_2 =$ a constant, so long as one particle is influenced only by the other.

This product of the mass of a particle by its velocity in a definite direction is, as we have said, called its "lineaı momentum" in that direction; and so the above formula states that, when two particles are left to their mutual interactions, the *sum* of their linear momenta along the line joining them does not change, however much the momentum of either one changes. Thus, the two particles may approach, may collide, may recede, etc.

This law will be given a simple geometrical interpretation when we speak of the properties of a system of particles. In the simplest case, viz., when the particles are moving along the same line, it is illustrated by many familiar facts. When a rifle is fired, it recoils; and, if it is suspended by cords, it will move in a direction opposite to that of the bullet. If m_1 is the mass of the rifle and m_2 that of the bullet, and v_1 and v_2 are the corresponding velocities, the sum $(m_1v_1 + m_2v_2)$ remains unchanged after the explosion of the powder. But before the explosion both v_1 and v_2 are zero; and therefore the sum $(m_1v_1 + m_2v_2)$ must be zero after the explosion also. So $m_1v_1 = - m_2v_2$ or $v_1 = -\frac{m_2v_2}{m_1}$. Similar illus-

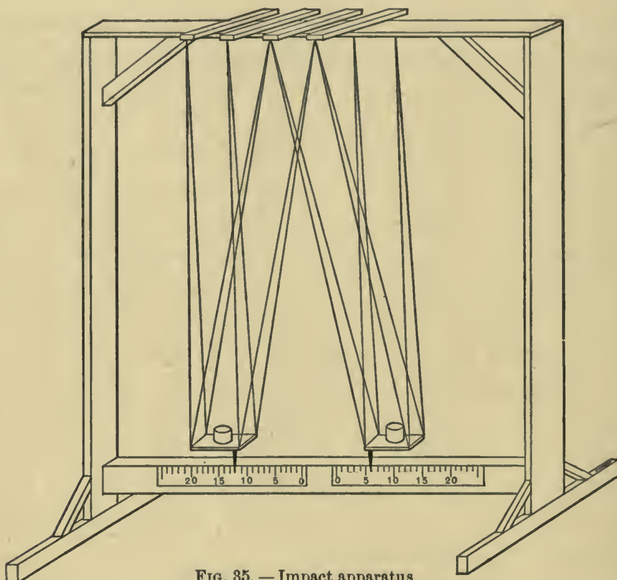


Fig. 35. — Impact apparatus.

trations of recoil or reaction are given when a man jumps sidewise off a chair, when water is allowed to escape from an opening in the side of a can which is suspended free to move, etc. A simple form of laboratory apparatus for studying the impact or collision of two bodies is shown in the cut. Directions for its use are given in all laboratory manuals.

Conservation of Linear Momentum. — The linear momentum of a particle is a vector quantity; and so the quantity $m_1v_1 + m_2v_2$ is a vector along the line joining the two particles at any instant. As there is no force acting on the particles except that due to their mutual action, their momenta in any other direction do not change. The particles are not in general moving in the line joining them; so, if u_1 is the actual velocity of the first particle at any instant, v_1 is its component along the line joining the particles, and m_1u_1 is its total linear momentum. Similarly, if u_2 is the actual velocity of the second particle, its total linear momentum is m_2u_2. The above statements are, then, equivalent to saying that the geometrical sum of m_1u_1 and m_2u_2 remains unchanged provided there are no external forces. The sum is, of course, a vector. This more general law may be illustrated by the impact of two moving billiard balls. Let the total momentum of the first particle be represented by the vector $\overline{BO}$, and that of the second by $\overline{AO}$; their geometrical sum is $\overline{CO}$. After impact, if $\overline{OE}$ and $\overline{OD}$ are the two momenta, their geometrical sum $\overline{OF}$ must be in the same direction as $\overline{CO}$ and equal to it.

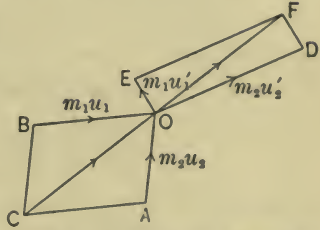


FIG. 36. — Conservation of linear momentum: the geometrical sum of the linear momenta does not change on impact: $\overline{CO} = \overline{OF}$.

These statements in regard to the mutual actions of two particles may be extended at once to a system of several particles. In this case there are forces acting on any one particle owing to the action of all the others, but these are equal and opposite to the reaction of this particle on the others. So, if we add together the components of all these internal forces resolved in any fixed direction, the sum is zero. This fact may be expressed in terms of momenta. Since the algebraic sum of the components in any one direction of the linear momenta of any two particles due to their interaction remains constant, so must the sum for all the particles; and, again, the *geometrical* sum of the total momenta of all the particles must remain constant so long as there is no external action. This is called the "Principle of the Conservation of Linear Momentum."

Centre of Mass. — This principle has a simple geometrical interpretation. Let the perpendicular distances of the particles from a fixed plane at any instant be called x_1, x_2, x_3, etc. If we assign to each of these distances as a measure of its "importance" the value of the mass of the particle at that distance, *i.e.* m_1 to x_1, m_2 to x_2, etc., the *mean* distance of the system of particles from the plane is

$$\frac{m_1x_1 + m_2x_2 + \cdots}{m_1 + m_2 + \cdots}.$$ (See page 30.)

This distance may be called $\bar{x}$; and, writing for the total mass of the system, $m_1 + m_2 + \cdots$, M, we have

$$\bar{x} = \frac{m_1x_1 + m_2x_2 + \cdots}{M},$$

or $M\bar{x} = m_1x_1 + m_2x_2 + \cdots$.

But the particles are moving, and each one has a velocity whose component perpendicular to this plane may be called u with a proper suffix. Thus u_1, this component of the velocity of the first particle, is the rate of change of x_1; etc. Owing to these motions $\bar{x}$ does not, in general, remain constant; and calling $\bar{u}$ its rate of change, we have from the above definition of $\bar{x}$, taking the rate of change of both terms of the equation,

$$M\bar{u} = m_1u_1 + m_2u_2 + \cdots.$$

But the principle of the conservation of momentum states that this sum, $m_1u_1 + m_2u_2 + \cdots$, remains constant so long as there are no external actions; therefore, $M\bar{u}$ is a constant or $\bar{u}$ is constant. Further, the *total momentum of the system* away from the fixed plane is $M\bar{u}$.

In order to describe definitely the position of a particle, its distances from three fixed planes at right angles to each other must be given; for instance, the position of the corner of a table in a room can be described by stating its height above the floor and its distances from two of the walls that make a corner. Thus, to describe the positions of the particles that are being considered, two other planes at right angles to each other and to the first must be chosen; and we have three distances for each particle, which we may call x, y, z, each with its proper suffix. We can then define a point $\bar{x}$, $\bar{y}$, $\bar{z}$, for which $\bar{y}$ and $\bar{z}$ have values similar to that of $\bar{x}$; that is,

$$M\bar{y} = m_1y_1 + m_2y_2 + \cdots,$$

$$M\bar{z} = m_1z_1 + m_2z_2 + \cdots.$$

If $\bar{v}$ and $\bar{w}$ are the rates of change of $\bar{y}$ and $\bar{z}$ at any instant, they are also constant, provided there are no *external* forces. Therefore the geometrical sum of $\bar{u}$, $\bar{v}$, $\bar{w}$, which is the actual linear velocity of the point $\bar{x}$, $\bar{y}$, $\bar{z}$, remains unchanged. (The product of this velocity by M, the total mass, is evidently equal to the total momentum of the system, the momenta of the separate particles being added geometrically.)

The point defined by the distances $\bar{x}$, $\bar{y}$, $\bar{z}$ is called the "centre of mass" of the system of particles; and the statements just proved in regard to its velocity are seen to be equivalent to stating that "the centre of mass of a system of particles free from all external actions moves with a constant velocity." This is another expression of the principle of the conservation of linear momentum.

The acceleration of the centre of mass is, in a perfectly similar manner, seen to be given by three equations,

$$M\bar{a} = m_1a_1 + m_2a_2 + \cdots,$$

and two others, in which a_1 is the rate of change of u_1, etc. But m_1a_1 is the total force on the particle whose mass is m_1, and if there are no external forces, this is due to the action of the other particles, resolved parallel to the direction of u_1; etc. And, since all the internal forces neutralize each other, this sum, $m_1a_1 + m_2a_2 + \cdots$, must equal zero. Therefore $\bar{a}$ is zero, as was shown in the preceding paragraph.

External Forces. — Let us now suppose that the system is acted upon by some external force, such as gravity. Each particle will then be under the action of two sets of forces, internal and external; so that if a is the acceleration in some particular direction of a particle whose mass is m, it is due to the sum of the components in this direction of the forces acting on this particle due to the two sets. If, then, we form the sum $m_1a_1 + m_2a_2 + \cdots$, it is the sum of the components in this direction of all the forces, internal and external; but the sum of the components of the internal forces is zero, and so only the sum of the components of the external forces need be considered. Call this sum X. Then $m_1a_1 + m_2a_2 + \cdots$ equals X. We have just shown, however, that this sum equals $M\bar{a}$, where $\bar{a}$ is the component of the acceleration of the centre of mass in the chosen direction. That is, $M\bar{a} = X$ or $\bar{a} = \frac{X}{M}$. In words, the acceleration in any direction of the centre of mass of a set of particles equals the sum of the components of the *external* forces in that direction divided by the total mass of the system.

We thus see the exact agreement between the properties of a single particle with those of a set of particles, the centre of mass of the set playing the part of the point occupied by the single particle.

The General Problem of Dynamics. — So far we have considered only the applications of forces to particles; but in nature, of course, material bodies are never in this form. The actual cases of forces are always those concerned with extended bodies; and it is evident that the effect of a force on an ordinary material body depends upon three things: its numerical value, its direction, and its point of application. Thus, as has been explained, if a blow is struck a body, the effect depends upon the point where the blow is applied as well as upon its numerical value and direction; and as a rule both translation and rotation are produced. We shall investigate these two questions separately; that is, we shall deduce first the effect of a force in producing linear acceleration and then its effect in producing angular acceleration.

Translation

Translation of an Extended Body. — We can at once apply all our deductions for a set of particles to a material body that has a finite volume or to a system of such bodies, if we *assume* that we can regard one as made up of particles. The centre of mass of such a body is defined by the same equations as for a set of particles; and its position can be calculated in many simple cases, *e.g.* a sphere, a uniform rod, etc. The linear momentum of such a body is the sum of the momenta of its parts, and therefore it equals the product of its total mass by the linear velocity of its centre of mass. The acceleration of its centre of mass in any direction is equal to the sum of the components in this direction of all the external forces divided by the total mass; and, therefore, the actual acceleration of the centre of mass is in the direction of the geometrical sum of the external forces and is numerically equal to the value of this resultant force divided by that of the total mass. This is the great physical property of the centre of mass. In general language we may say that no matter what the shape or size of the body or where the external forces are applied, no matter how it turns or spins in its motion, the centre of mass follows the same path with the same velocity and acceleration that would be observed if these forces were transferred parallel to themselves and acted on a particle whose mass was that of the body.

One or two illustrations may be given. If a particle is thrown obliquely in the air, it will describe the path of a parabola; similarly, when a chair is thrown in the air, or when an acrobat jumps off a springboard, turning a somersault, their centres of mass describe parabolas. So also, when a bombshell explodes in mid air, the centre of mass of its fragments describes a parabola as they gradually fall toward the earth. If a random blow is struck a body, its centre of mass moves off in the direction of the blow.

Translation of a System of Bodies. — If we have a system of bodies, it may be considered, *so far as translation is con-*

cerned, as a system of particles coinciding with the centres of mass of the bodies and having their masses. This follows at once from what has just been said in regard to the properties of a single body. It is this fact that justifies us in using the illustrations we have given in discussing the laws concerning the motion of particles. It also justifies the language used in speaking of mass and its measurement.

The properties of matter in rotation will be considered in the next Section.

Illustrations of the calculation of the position of the centre of mass.

It is easy to calculate the position of the centre of mass of any regular solid provided the matter is distributed uniformly throughout it, *e.g.* a cylindrical wire, a cube, a sphere, etc., and also of a system of bodies whose masses and the positions of whose centres of mass are known.

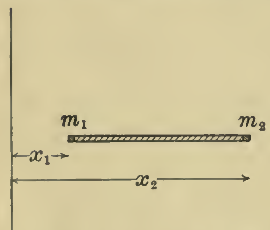


FIG. 37.—Centre of mass of a uniform rod: m_1 and m_2 are minute portions at its ends.

1. *Uniform rod.*—The centre of mass of a uniform rod is its middle point. For, consider the rod as made up of equal separate particles; and let m_1 and m_2 be two which are at the ends. Take as the plane of reference one perpendicular to the rod, and let x_1 and x_2 be the distances of m_1 and m_2 from the plane. Then, by definition, their centre of mass is given by the equation:

$$(m_1 + m_2)\bar{x} = m_1x_1 + m_2x_2.$$

But $m_1 = m_2$; hence $\bar{x} = \frac{x_1 + x_2}{2}$, *i.e.* the centre of mass of these two particles is halfway between them. A similar statement is true for the other masses which make up the rod, always combining those which are equidistant from the two ends; and therefore the centre of mass of the rod is this same point. Q.E.D.

The centre of mass of a uniform sphere (or spherical shell) is also its centre of figure.

2. *Uniform triangular board.*—Draw the three medial lines $\overline{Aa}$, $\overline{Bb}$, $\overline{Cc}$, connecting the vertices with the middle points of the opposite sides. They meet in a point O. Since the straight line $\overline{Bb}$ divides the triangle into two equal halves, the centre of mass must lie on it; for the triangle may be considered built up of a great number of strips parallel to the side $\overline{AC}$, and as the centre of mass of each of these lies on the medial line $\overline{Bb}$, the centre of mass of the entire triangle must lie on it also. Similarly, it must lie on $\overline{Aa}$ and $\overline{Cc}$; that is, it must be the point O, their common point of intersection.

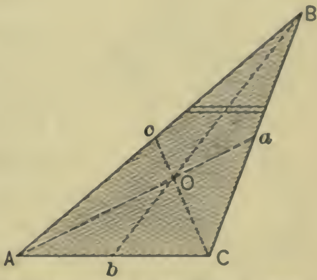


FIG. 88.—Centre of mass of a uniform triangular board.

3. A uniform rod, mass $m_3 = 25$, carrying two symmetrical bobs whose masses are $m_1 = 15$, $m_2 = 20$; the dimensions and distances being as indicated in the cut.

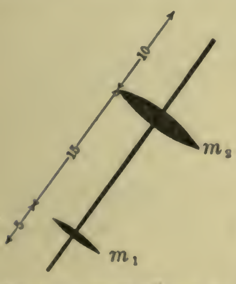


FIG. 89.—Centre of mass of a weighted bar.

The centre of mass of the rod itself is its middle point; which is at a distance 15 cm. from the ends. Take as the plane from which to measure distances one perpendicular to the rod at its left end. Then

$$m_1 = 15,\ x_1 = 5;\ m_2 = 20,\ x_2 = 20;\ m_3 = 25,\ x_3 = 15;$$

and therefore

$$\bar{x} = \frac{m_1x_1 + m_2x_2 + m_3x_3}{m_1 + m_2 + m_3} = \frac{75 + 400 + 375}{60} = 14.17.$$

The centre of mass must, then, lie at a distance of 14.17 cm. from the plane at the end of the rod; and since the bobs are symmetrical, it must lie in the axis of the rod at that distance from the end.

4. *A rigid framework lying in a plane;* two bodies, whose masses $m_1 = 20$, $m_2 = 10$, are connected by massless wires to a uniform rod whose mass $m_3 = 10$; the dimensions being as shown in the cut.

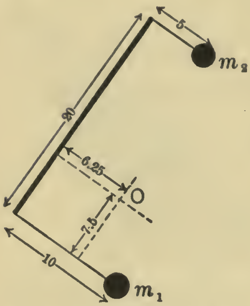


FIG. 40. — Centre of mass of a rigid framework.

Take as the two planes of reference one perpendicular to the rod at its lower end, the other through the rod perpendicular to the two wires.

Hence

$m_1 = 20,$	$x_1 = 0,$	$y_1 = 10,$
$m_2 = 10,$	$x_2 = 20,$	$y_2 = 5,$
$m_3 = 10;$	$x_3 = 10;$	$y_3 = 0.$

So $\bar{x} = \frac{300}{40} = 7.5$; $\bar{y} = \frac{250}{40} = 6.25$.

That is, the centre of mass is a point at a distance 7.5 cm. from the plane perpendicular to the rod at its lower end; and a distance 6.25 cm. from the rod itself in a direction parallel to the wires; therefore it is at the point O as shown.

Rotation

Introduction. — Let us now consider the rotation of a material body when it is acted on by a force. A simple case is that of a body pivoted on an axle, *e.g.* a door. If a push is given it, in general an angular acceleration will be produced; but if the push is so directed that its line of action passes through the axis, it has no effect on the rotation. It is a fact easily observed that the effect increases as the direction of the line of action of the push is removed farther and farther from the axis.

Moment of a Force. — To be definite, let the cut represent a cross section of the body so taken as to be perpendicular to the axis and to include the *point* of application of the force. Resolve the force into two, one parallel to the axis, the other in this plane section. Only the latter has any effect on the rotation. Let P be the intersection of the plane by the axis, and F the position and direction of this component of the force. Prolong the line of action of the force, if necessary, and let fall upon it a perpendicular from P; this is called the "lever arm" of the force F with reference to the axis through P. If its length is l, experience shows that the change in rotation produced by the force F varies directly as F and as l; if either is increased, the angular acceleration is increased, and *vice versa*. To determine the exact effect that the two quantities F and l have in producing angular acceleration, the simplest method is to apply a second force to the pivoted body in such a direction as to neutralize the rotating effect of the first. Experiments lead us to believe that if the relations between the two forces and their two lever arms is such that $F_1 l_1 = F_2 l_2$, one force balances the other so far as rotation is concerned. The product of a force by its lever arm, with reference to any axis, is called the "moment of the force with reference to this axis"; and this is evidently the proper measure of the rotating effect of a force. The moment bears the same relation to rotation that force does to translation.

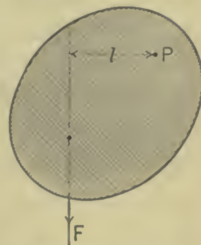


Fig. 41. — Moment of force: plane section of body perpendicular to axis at P; F is component of force in this plane.

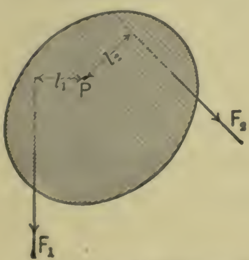


Fig. 42. — Rigid body pivoted by an axis through P, and in equilibrium under the action of two forces. $F_1 l_1 = F_2 l_2$.

A moment is defined by the *position* of its axis, its amount, and the sense of its rotating action. (A moment in the sense of the hands of a watch is given an algebraic sign different from one in the opposite sense.) A moment is, then, a rotor.

Moment of Inertia. — We must next determine the relation between the moment of a force and the angular acceleration produced owing to it. The simplest case of a rotating body that we can imagine is that of a particle attached to a pivot by a massless rod. Let the particle of mass m be at A; let the pivot be at P; and let there be a force F in the direction $\overline{AB}$. Draw the perpendicular $\overline{PQ}$ upon the line of action of the force; let its length be l; and let the angle it makes with the massless rod $\overline{AP}$ be called N. Let the length of this rod be called r. The lever arm of the force F is l, and it equals $r \cos N$; hence the moment of the force F around the axis at P is $Fr \cos N$. There is also a force on the particle owing to the rod; but this force has a lever arm zero, since its line of action passes through P. As a result of the force F the particle will receive an angular acceleration, whose value may be found as follows: resolve F into two components, one along $\overline{AP}$, the other perpendicular to it along $\overline{AC}$; the value of the latter is $F \cos N$, because the angles (BAC) and (APQ) are equal; and this is the only component that produces any angular motion of the particle. Therefore, the *linear* acceleration of the particle in the direction of this component is $\frac{F \cos N}{m}$. But it was proved on page 55 that the angular acceleration of the moving point equals its linear acceleration divided by r. So, calling the angular acceleration A, $A = \frac{F \cos N}{mr}$; and its connection with the moment of the force $Fr \cos N$ is evident; viz., $Fl = mr^2A$. Writing a single symbol, L, for the moment, this equation becomes $L = mr^2A$ or $A = \frac{L}{mr^2}$.

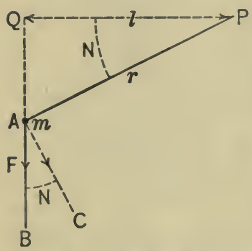


Fig. 43. — A particle of mass m is constrained to move in a circle of radius r about P, and is under the action of a force, F, at any instant.

Since "moment of force," L, plays the same part in rotation that force does in translation, and since angular acceleration in one corresponds to linear in the other, it is seen that, in the rotation of a particle round an axis in a circular path of radius r, mr^2 takes the place of mass in translation; that is, it measures the inertia of the matter for rotation; if it is large, A is small, and conversely. Similarly, if a material body, considered made up of particles, is rotating around a fixed axis, the total inertia of the body for rotation is $m_1r_1^2 + m_2r_2^2 + \ldots$, where r_1 is the radius of the path in which the particle whose mass is m_1 is moving, etc. This sum is, of course, an arithmetical one. It is called the "Moment of Inertia" of the body with reference to the axis of rotation, and is ordinarily written I.

Moments of inertia may be calculated for the regular solids provided the matter is uniformly distributed in them, and in many other special cases. Processes of the infinitesimal calculus are, however, required. A few illustrations may be given. For a sphere of radius R and mass M, the axis being a diameter, $I = \frac{2}{5} MR^2$; for a cylindrical cylinder of radius R and mass M, the axis being the axis of symmetry, $I = \frac{1}{2} MR^2$. (If, however, the axis is a generating line of the surface, $I = \frac{3}{2} MR^2$.) If the length of the cylinder is H, and if the axis passes through its middle point perpendicular to its length, $I = M\left(\frac{H^2}{12} + \frac{R^2}{4}\right)$.

Equation of Motion for Rotation of a Rigid Body around a Fixed Axis. — We must next discuss the rotation of a body of definite size and shape round a fixed axis under the action of an external moment. Such a body is called in dynamics a "rigid" one. As before, describe a cross section through the body perpendicular to the axis and including the point of application of the force; and let the external force be resolved into two components, one parallel to the axis, the other in the plane. Let the latter be represented in the cut by F; let P be the trace of the axis; and l the lever arm. The moment of the force is Fl.

It should be expressly noted that this quantity has the same value wherever the *point of application* of the force is, provided only that the force keeps *its line of action*, *i.e.* provided l does not change. (Thus the force may be applied at A_1, or A_2, or A_3, etc., in the line of action.) In discussing translation it was shown that the effect of a force on the motion of the centre of mass did not depend upon the *position* of the line of action nor of the point of application of the force, but, simply, on its *amount* and *direction;* consequently, the *total* effect of a force upon a *rigid* body depends upon its amount, its direction, and the position of its line of action with reference to the body, *not* upon the position in the line of action of its *point of application*. (If the body is made up of particles so connected as to form a figure of variable size and shape, *e.g.* an elastic body, l and I would change.)

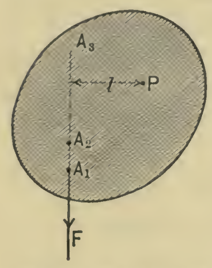


FIG. 44. — A rigid body under the action of a force F. A_1, A_2, A_3 are points in the line of action of F.

There are also internal forces between the particles; in any actual case there is friction between the material pivot on which the body turns and the body itself; and, further, the pivot in general exerts a force on the body. The moment of this last force is zero, because its lever arm is zero; and we shall assume for our present purposes that there is no friction. The moments of the internal forces neutralize each other, moreover, because the forces between any two particles have been assumed to be equal and opposite and *in the line joining them*, so the lever arms are equal, and therefore the two moments are equal, but in opposite senses of rotation.

The total moment around the axis is, then, that of the external force F; and calling, as before, $Fl = L$, we have the fundamental equation for the rotation of a rigid body around a fixed axis,

$$A = \frac{L}{I},$$

where A is the angular acceleration, L is the moment of the external force, and I is the moment of inertia of the body, the last two quantities being referred to the fixed axis.

Illustration. — A simple but important illustration of this formula is afforded by a pendulum. Let $\overline{OM}$ be a vertical line through the point of support, and $\overline{OP}$ the position of the pendulum at any instant, so that the angular displacement is the angle (MOP), or N. If m is the mass of the bob of the pendulum, and r the length $\overline{OP}$, the moment of inertia of the pendulum around an axis at O perpendicular to the plane of motion is mr^2. There are two forces acting on the pendulum bob: its weight, mg, whose moment around the axis through O is $mgr \sin N$, and the tension of the string, whose moment is zero. But this moment due to gravity is in an opposite direction to the displacement; hence, in the formula $A = \frac{L}{I}$, $L = -mgr \sin N$ and $I = mr^2$. Therefore, $A = -\frac{mgr \sin N}{mr^2} = -\frac{g}{r} \sin N$. If the amplitude of vibration is made infinitesimal, we may replace $\sin N$ by N; and $A = -\frac{g}{r} N$. Therefore, the motion is harmonic, and the period is $2\pi\sqrt{\frac{r}{g}}$, as was found before.

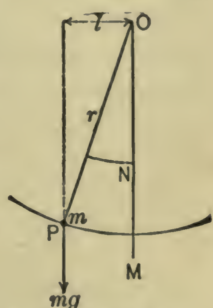


Fig. 45. — Simple Pendulum.

Composition of Moments. — It is evident, since forces act independently, that if there are several external forces acting, L is the sum of the moments of the separate forces, proper attention being given the algebraic signs.

It is easily proved by geometry that if a rigid body is acted on by two forces that lie in a plane, the sum of the moments about any axis perpendicular to this plane equals the moment of their geometrical sum, provided its line of action passes through the point of intersection of the two forces. For, let the two forces have their lines of action in the lines $\overline{OA}$ and $\overline{OB}$, and let them be represented in direction and amount by the vectors $\overline{OA}$ and $\overline{OB}$. Their geometrical sum is then $\overline{OC}$, and its line of action is here represented as passing through O, the intersection of $\overline{OA}$ and $\overline{OB}$. Let P be the trace in this plane of the perpendicular axis; draw the lines $\overline{PO}$, $\overline{PA}$, $\overline{PB}$, and $\overline{PC}$. The numerical value of the moment of the force $\overline{OA}$ about the axis through P is the product of the length of $\overline{OA}$ by the perpendicular distance from P to this line; that is, it equals twice the numerical value of the area of the triangle (POA). Similarly, the moment of $\overline{OB}$ is given by twice the area of the triangle (POB), and that of $\overline{OC}$ by twice the area of the triangle (POC). But the area of the latter triangle equals the sum of the areas of the three triangles (POB), (PBC), and (OBC); and the combined areas of these last two equals the area of the triangle (POA), because all three have bases of the same length, viz., $\overline{OA}$ or $\overline{BC}$, and the combined altitude of the first two equals that of the third. Therefore the area (POC) equals the sum of the areas (POB) and (POA); and it follows at once that the moment of $\overline{OC}$ equals the sum of the moments of $\overline{OA}$ and $\overline{OB}$. (In the cut, these last two moments are in the same direction, viz., that corresponding to rotation counter-clockwise. If P is placed elsewhere, the two moments might be in opposite directions; in which case they would

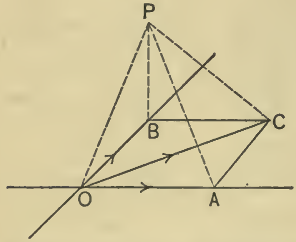


FIG. 46.—Composition of the moments of $\overline{OB}$ and $\overline{OA}$ about axis at P.

be given opposite signs, and their algebraic sum must be taken.)

Conservation of Angular Momentum. — Returning to the general formula, it is seen that $A = 0$, if $L = 0$; that is, the angular velocity of a body turning on a fixed axis remains constant if either the moment of the external force is zero or if the sum of the moments of the external forces is zero. This is perfectly analogous to the case of translation when $F = 0$, and is illustrated by the rotation of a wheel whose friction with its axle can be neglected.

If the value of the angular velocity at any instant is written h, the product Ih is called the "angular momentum" of the rigid body around the given axis; and the general law may be stated by saying that the moment of the force around the axis equals the rate of change of the angular momentum about the same axis. So, if the total moment is zero, the angular momentum remains constant.

If the rotating body is not rigid, the angular momentum is the sum

$$m_1 r_1^2 h_1 + m_2 r_2^2 h_2 + \cdots,$$

where m_1, r_1, h_1 apply to one particle of the body; etc. The statement that this sum is constant when the external moment is zero is still true, however. Several illustrations are worth noting. If the angular velocity of all the particles is the same, the angular momentum may be expressed $(m_1 r_1^2 + m_2 r_2^2 + \cdots) h$; and now, if owing to any *internal* cause the values of r_1, r_2, etc., become smaller, the value of h must increase. This was the case with the planets in their early history and is so with the sun at present. There are forces acting on these bodies, but their moments about the axes of rotation are zero; and the formula may be applied. As time goes on, the planets and sun have contracted owing to internal gravitational forces; and therefore, as proved above, their angular velocities have increased. Again, as an acrobat turns a somersault in the air, while at the same time he jumps over an obstacle, his centre of mass describes the path of a parabola; but he can increase his angular velocity by drawing in his arms and legs, thus diminishing his moment of inertia, because there is no *moment* due to the force of gravity.

Illustrations of Rotation. — If a rigid body is turning on a fixed axis, a moment round the same axis will change the angular speed, either increasing or decreasing it, as is illustrated by setting in motion a grindstone by means of a crank handle or in stopping one by means of a brake. If, however, the moment is around an axis at right angles to that of the existing angular velocity, the direction of this axis will be changed; this is illustrated by the motion of a rolling hoop whose upper edge is pushed sidewise, as explained on page 54, or by the motion of a spinning top whose axis is inclined to the vertical.

Principal Axes. — When a material body is rotating on a fixed axle there are in general certain forces and moments which the body exerts on the axle and which are borne by the bearings that hold the axle. If the axis does not pass through the centre of mass, there is a pull on the axle toward this point as it moves in a circle around the axis, due to the reaction of the centrifugal force. Its amount is Mrh^2, if r is the radius of this circle, M the mass of the body, and h its angular speed. At any point of the moving body there are three directions, called "the principal axes at that point," such that if the axis of rotation does not coincide with one of them there is a twist on the axle tending to make it turn. This push and twist must of course be withstood by the axle or its bearings. So, if the body is to turn freely, producing no forces or moments on the axle, the axis of rotation must pass through the centre of mass and must be one of the principal axes at this point. In other words, to make a body maintain its axis of rotation in a definite position and direction other than one which is a principal axis at its centre of mass, a force or moment is required; and, if no such force or moment is applied, the position or the direction of the axis of rotation will change. But, if a body is set spinning about a principal axis at its centre of mass, it will maintain its rotation unchanged in every respect, if no moment acts upon it. This last statement is illustrated by the throwing of a quoit, whose axis remains parallel to itself if it is set spinning in the proper way; by the motion of the earth on its axis, which moves in space parallel to itself (omitting small perturbations and the effect due to the protuberances at the equator); by the motion of projectiles shot out by "rifled" guns; etc.

Translation and Rotation

It is interesting to arrange in parallel columns corresponding properties of translation and of rotation around a fixed axis.

Translation of a Particle	*Rotation of a Rigid Body*
a. mass	*a.* moment of inertia
b. force	*b.* moment of force
c. linear momentum	*c.* angular momentum
$F = ma$	$L = IA$
Forces act independently.	Moments act independently.
If $F = 0$, the linear momentum remains constant.	If $L = 0$, the angular momentum remains constant.
If the direction of the force is perpendicular to that of the motion, the direction of the latter is changed.	If the axis of the moment is perpendicular to that of the motion, the direction of the latter is changed.

Motion in General of a Material Body

General Description. — A material body will, in general, receive both linear and angular acceleration when acted upon by external forces; but these are independent of each other. The centre of mass of the body will receive a linear acceleration; and, as this point moves in space, the rotation will take place about it exactly as if it were a fixed point in the figure. Several illustrations have been given already; viz., the motion of an acrobat, that of a chair thrown in the air, etc. If a rigid body is struck a blow at random, its centre of mass will move in the direction of the blow, and the body will rotate, in general; but, if the line of action of the blow passes through the centre of mass, there is no rotation. Consequently, if two lines of action are found such that blows along them do not produce rotation, they must intersect at the centre of mass.

Therefore, to discuss completely the most general problem in dynamics, all that is necessary is to know the laws of motion of translation and those of rotation about an axis passing through a fixed point.

Resultant. — There are certain cases in which the resulting accelerations of a body under the action of several forces might have been produced under the action of a single force; if such is the case, this force is called the "resultant" of the others. The action of a single force on a body is to accelerate the centre of mass in its direction and to cause angular acceleration around an axis through the centre of mass at right angles to the plane including it and the line of action of the force.

If the various forces acting in a body all lie in a plane, or if they are all parallel, it may be shown that they have a resultant, with the exception of one case, which will receive due attention.

Non-parallel Forces. — Let the body be acted upon by two coplanar non-parallel forces F_1 and F_2. Their geometrical sum R may be found as usual; and its effect in accelerating the centre of mass equals the combined effects of F_1 and F_2. But if R is to be the resultant, a position for its line of action must be found such that its moment shall equal the combined moments of F_1 and F_2. It has been shown on page 92 that this is the case if the line of action of R passes through the intersection of the lines of action of F_1 and F_2. Therefore the obvious geometrical method of determining the resultant of two coplanar non-parallel forces is to take a plane section through the body so as to include the lines of action of the two forces, prolong these lines until they meet at a point, O; lay off from O along the lines of the forces, distances $\overline{OA}$ and $\overline{OB}$ proportional to their numerical values; and complete the parallelogram. The diagonal $\overline{OC}$ represents in direction, amount, and position the resultant R; its point of application may be anywhere along the line $\overline{OC}$, *e.g.* at Q.

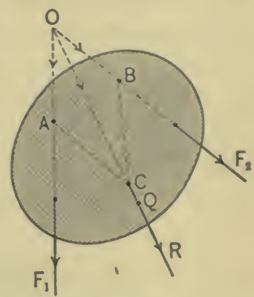


FIG. 47. — Rigid body under the action of two non-parallel coplanar forces F_1 and F_2. Their resultant is R.

In a perfectly similar manner, by continuing the process, the resultant of any number of coplanar non-parallel forces may be determined.

If M is the mass of the body; I, its moment of inertia about an axis through the centre of inertia perpendicular to the plane of the forces; and l, the perpendicular distance from the centre of inertia to the line of action of the resultant R; the linear acceleration of the centre of mass is $\frac{R}{M}$; and the angular acceleration of the body is $\frac{Rl}{I}$.

Equilibrium. — It is evident that if there are three forces acting on the body, F_1, F_2, and one equal and opposite to R along the same line of action, they will neutralize each other completely: there will be no acceleration of any kind. Such a state is called one of "equilibrium." The general law may then be stated that, if a body is in equilibrium under the action of three non-parallel forces, they must lie in a plane, their lines of action must meet in a point, and any one must be equal and opposite to the geometrical sum of the other two,—or, what is the same thing, if the three are added geometrically, they will form a closed triangle. (It was by considering the special case of equilibrium of a body on an inclined plane that Stevin was led to the statement of the parallelogram of forces. See page 76.)

Parallel Forces. — If the two forces acting on the body are parallel, their resultant, so far as translation is concerned, must be a force parallel to them, whose numerical value is their algebraic sum. That is, if F_1 and F_2 are their *numerical* values, that of the resultant is $F_1 + F_2$, if the forces are in the same direction; but, if they are in opposite directions, and if F_2 is the greater, the resultant is in the direction of this force, and has the value $F_2 - F_1$. Further, in order to satisfy the requirements in regard to rotation, this resultant must have such a position that its moment around any axis equals the algebraic sum of the moments of F_1 and F_2 around the same axis. Describe a plane section through the body, including the parallel lines of action of F_1 and F_2. The line of action of the resultant must also lie in this plane; otherwise the resultant would have a moment about any axis lying in it, which F_1 and F_2 do not.

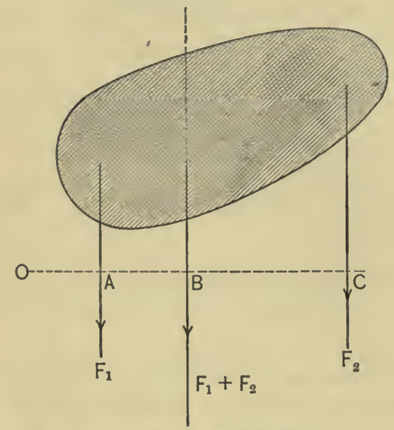


Fig. 48. — Rigid body under the action of two parallel forces F_1 and F_2 in the same direction. Their resultant is $F_1 + F_2$.

We shall consider first the case when the two forces are in the same direction. Imagine an axis perpendicular to their plane, and let its trace on the plane be O. From O draw a line $\overline{OAC}$ perpendicular to the lines of the two forces; if the parallel force $(F_1 + F_2)$ is to be the resultant, it must be so placed that its moment around the axis through O equals the sum of the moments of F_1 and F_2. Let its position be indi cated as shown in the cut, its intersection with the line $\overline{OAC}$ being at B. The condition that B must satisfy is that

$$(F_1 + F_2)\,\overline{OB} = F_1\overline{OA} + F_2\overline{OC},$$

or

$$F_1\overline{AB} = F_2\overline{BC}, \text{ i.e. } \frac{AB}{BC} = \frac{F_2}{F_1}.$$

This may also be expressed as follows:

$$F_1\overline{AC} = (F_1 + F_2)\overline{BC}, \text{ i.e. } \frac{BC}{AC} = \frac{F_1}{F_1 + F_2},$$

or

$$F_2\overline{AC} = (F_1 + F_2)\overline{AB}, \text{ i.e. } \frac{AB}{AC} = \frac{F_2}{F_1 + F_2}.$$

These relations are independent of the position of O, and therefore hold true for any axis. They determine uniquely the line of action of the resultant.

If the forces are parallel but in opposite senses of direction, and if F_2 is the greater, the resultant is $F_2 - F_1$ and is in the direction of F_2, and it is so placed that its moment around any axis is equal to the *difference* in the moments of F_1 and F_2. The same formulæ as above apply, giving F_1 in these a negative sign.

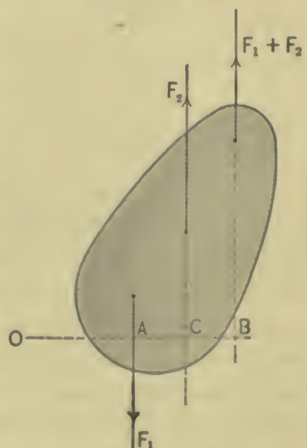


FIG. 49. — Rigid body under the action of two parallel forces F_1 and F_2 in opposite directions.

As in the previous case, the acceleration of the centre of mass is in the direction of the resultant and has the value $\frac{F_1 + F_2}{M}$; and the angular acceleration has the value $\frac{(F_1 + F_2)l}{I}$, where l is the perpendicular distance from the centre of mass to the line of action of the resultant, and I is the moment of inertia about an axis through the centre of mass and perpendicular to the plane of the forces. This process may be continued so as to determine the resultant of any number of parallel forces.

"Couples." — An ambiguity arises when the two parallel forces are equal but in opposite directions, *i.e.* when $F_1 = -F_2$. In this case there is no resultant. (On substituting these values in the previous solution, it is seen that, if there were a resultant, its value would be zero, and its line of application would be at an infinite distance.) Such a combination of two equal but opposite parallel forces is called a "couple."

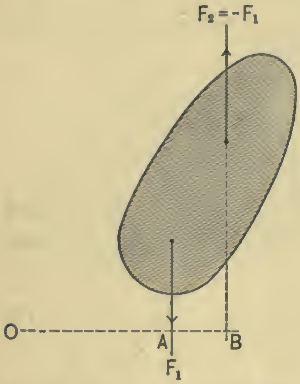


FIG. 50. — Rigid body under the action of a "couple."

Their sum is zero, therefore the linear acceleration of the centre of mass is zero, *i.e.* it has a constant linear velocity; but there is an angular acceleration around an axis through the centre of mass perpendicular to the plane of the couple. Describe a plane section through the body so as to include the two parallel forces, and consider any axis perpendicular to this plane. Let its trace be O; and from it drop a perpendicular $\overline{OAB}$ upon the lines of action of the forces. The sum of their moments, taken contrary to the direction of rotation of the hands of a watch, is $F_1\overline{OB} - F_1\overline{OA} = F_1\overline{AB}$. This product is called the "strength of the couple," and is evidently independent of the situation of the axis. Therefore, while the centre of mass of the body retains a constant velocity, the angular acceleration around an axis through the centre of mass and perpendicular to the plane of the couple equals its "strength" divided by the moment of inertia of the body about this axis.

Equilibrium. — If a body is under the action of three parallel forces that lie in a plane whose algebraic sum is zero, and the algebraic sum of whose moments around any axis is zero, there is neither linear nor angular acceleration: the body is in equilibrium. Conversely, if a body is in equilibrium under the action of three, *or of any number of*, parallel forces, their algebraic sum is zero, and that of their moments around any axis is zero.

A couple can evidently be balanced by another couple that is in the same plane, but that is not necessarily parallel to the first, of equal strength but of opposite sign.

Centre of Gravity. — The most important illustration of parallel forces is furnished by gravity. A body, considered made up of particles, is acted upon by as many forces as there are particles; and these forces are parallel because their lines of action join the separate particles to the centre of the earth. Their resultant is their arithmetical sum, and is called the "weight of the body"; its numerical value is the product of the total mass of the body by g, the acceleration of a falling body; and the position of its line of application may be found as follows: Let m_1 and m_2 be the masses of any two particles of the body at an invariable distance apart, $\overline{AB}$. The forces of gravity are vertical and their numerical values are m_1g and m_2g; their resultant is $(m_1+m_2)g$, so placed that $(m_1+m_2)g\overline{QR} = (m_1g)\overline{PR}$, where $\overline{PQR}$ is a horizontal line, intersecting the lines of the three forces in the points P, R, Q. That is, $(m_1+m_2)\overline{QR} = m_1\overline{PR}$. This line of action of the resultant meets the line joining the particles in a point C, such that, by similar triangles,

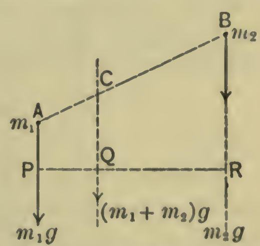


FIG. 51. — m_1 and m_2 are two particles of a rigid body under the action of gravity. Their centre of gravity is C.

$$\overline{AB} : \overline{CB} = \overline{PR} : \overline{QR}.$$

But $\overline{PR} : \overline{QR}$, as we have just seen, equals $\frac{m_1 + m_2}{m_1}$. Therefore, $\frac{AB}{CB} = \frac{m_1 + m_2}{m_1}$, or $\overline{CB} = \frac{m_1}{m_1 + m_2} \cdot \overline{AB}$. It is seen, then, that the point on the line joining the two particles through which their resultant weight passes, depends upon the masses of the particles and upon their distance apart, but is independent of the *position or direction* with reference to the earth of the line joining them; it is therefore a fixed point on this line. Similarly, in the general case, there is a point fixed, with reference to the particles, through which the line of action of the resultant weight passes, however the body is situated with reference to the earth. This is called the "centre of gravity" of the body; and it is seen from the above equation that it coincides with the "centre of mass."

N.B. — The above proof of the existence of a centre of gravity and of its coincidence with the centre of mass depends upon the fact that "g" is a constant for all amounts and all kinds of matter.

Equilibrium

The state of equilibrium of a body has already been defined as that in which there is no acceleration, either linear or angular; and the obvious conditions are that both the sum of the components of the forces in any direction and the sum of the moments around any axis should be zero. (We may speak in the same way of the equilibrium of a system of bodies.) If the body is at rest with reference to any standard figure, — *e.g.* a book lying on a table is at rest with reference to the table, — the equilibrium is called "statical"; while, if the body is in motion, — unaccelerated, of course, — the equilibrium is called "kinetic," *e.g.* a sphere rolling on a smooth horizontal table. There are several kinds of equilibrium, however, depending upon what changes in the motion of the body (or system of bodies) take place when a slight impulse or blow is given it.

Stable. — If the equilibrium of the body is such that, as a result of the impulse, it does not continue to move away from its former position, but makes oscillations about it, it is said to be "stable." This is illustrated by practically all bodies in nature that are in equilibrium. An ordinary pendulum when at rest, a block when it rests on a table, a body hanging at rest from a spiral spring, etc., are illustrations of statical stable equilibrium. If an ellipsoidal body is set spinning around its longest or shortest axis, the motion is stable.

Unstable. — But if the equilibrium is such that as a result of the impulse the body departs farther and farther from its former position, it is said to be "unstable." This is illustrated by a board balanced in a vertical position on one corner, by a conical body balanced on its point, etc., or by an ellipsoidal body spinning around its intermediate axis. It is obvious that when a body is in unstable equilibrium

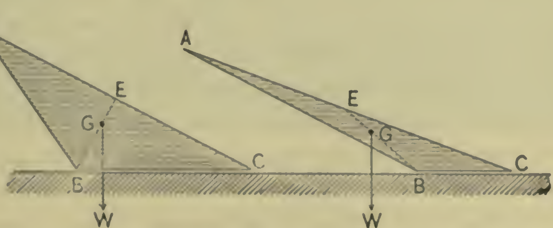


FIG. 52. — *G* is the centre of gravity; if the vertical line falls inside the base *BC*, there is equilibrium.

for impulses in some directions, it may be stable for others; and again a body may be stable for an extremely small impulse and unstable for a larger one, so that there are "degrees of stability." Thus a block shaped as shown in the cut and resting on a horizontal support is in stable equilibrium, because the force of gravity acting vertically down through the centre of gravity is balanced by an upward force due to the table. But this last force can pass through the centre of gravity and therefore neutralize gravity completely only so long as the line of action of the force of gravity falls inside the area of contact between the block and the table. If the shape of the block is so changed that this line of action approaches the edge of this area, the stability becomes less and less — for an impulse in the proper direction. When the line of action of the weight falls outside the edge, this downward force forms a couple with the upward force due to the table, and the block will turn over.

Neutral. — Another state of equilibrium is also recognized; namely, that in which, when an impulse is given the body, the change in motion produced remains permanent. This kind of equilibrium is called "neutral," and is illustrated by a sphere or a cylinder lying on a smooth horizontal table, by a body pivoted on an axis passing through its centre of gravity, etc.

It is evident that when the condition of a body in unstable equilibrium is disturbed, it passes over into either a stable or a neutral condition; and, as disturbances are always occurring in nature, the condition of unstable equilibrium can exist for only infinitesimal intervals of time.

Principle of Stable Equilibrium. — Any disturbance of stability must produce a reaction which tends to restore the body or system to its previous condition; and this principle can be applied to any stable condition, whether it is a purely mechanical one or not. Consider some illustrations of stability. (1) A body hanging suspended by a spiral spring is in stable equilibrium. If a blow downward is given it, the initial velocity will be decreased owing to the increased tension of the spring. Hence, if the tension of a stretched spiral spring is increased by any means, it will raise the suspended body. (2) An iron bar surrounded by some medium, *e.g.* water, at a constant temperature is in stable equilibrium; for if its temperature is suddenly increased in any way, the tendency will be for it to return to the temperature of the surrounding medium. Now, when the temperature of an iron bar is increased, its length is increased; but this act of increasing in length produces a tendency for the bar to return to its former temperature. That is, if an iron bar is stretched by mechanical means, its temperature will fall. (3) Just the opposite effect happens with a piece of rubber cord from which a weight is hanging. When its temperature is lowered, it elongates; consequently stretching a rubber cord raises its temperature.

FIG. 53. — Illustration of three kinds of equilibrium.

Work and Energy

Measurement of the Effect of a Force. — In the previous sections of Dynamics we have considered, generally speaking, only one property of a force, viz., the fact that it produces a change in momentum. It was shown, however, on page 45 that when a particle is moving in a straight line with a constant acceleration, this quantity could be expressed in two ways:

$$a = \frac{s_2 - s_1}{t_2 - t_1} \text{ and } a = \frac{1}{2}\frac{s_2^2 - s_1^2}{x_2 - x_1},$$

where s_1 is the speed of the particle at the instant T_1 when it has reached a point at a distance x_1 from a fixed point of reference in the line of motion; and s_2 is the speed at the instant T_2 and position x_2. Multiplying each of these values of a by m, the mass of the moving particle, we have the force. Thus,

$$f = \frac{m(s_2 - s_1)}{T_2 - T_1}, \quad f = \frac{\frac{1}{2} m (s_2^2 - s_1^2)}{x_2 - x_1}.$$

The former is the ordinary expression for the value of a constant force, stating that it equals the change in the linear momentum in a unit of time. The product $f\,(T_2 - T_1)$ is called the impulse of the force; so this formula expresses the fact that the impulse of the force equals the change in linear momentum of the particle. If the force varies, we must understand by $f\,(T_2 - T_1)$ the sum of a series of terms each of which is a force multiplied by its time of action.

Definition of Work and Kinetic Energy. — The latter formula, however, is a new expression. Suitable names have been given its terms: $\frac{1}{2}\,ms^2$ is called the "kinetic energy" of the particle whose mass is m when its speed is s; $f\,(x_2 - x_1)$ is called the "work done *by* the force" f in the distance $x_2 - x_1$, provided the speed is increasing, or the "work done *against* the force" f if the speed is decreasing — thus this equation reads: either, "the work done by the force in the distance $x_2 - x_1$ equals the increase of the kinetic energy of the moving particle in that space"; or, "the work done against the force f in the distance $x_2 - x_1$ equals the decrease of the kinetic energy of the moving particle in that space." Several things should be noticed:

1. The distance $x_2 - x_1$ is measured in the line of action of the force; if the line of motion makes an angle N with the line of the force, the work is the product of $F \cos N$ and $(x_2 - x_1)$.

2. The idea of work involves *both force and motion in the direction of the force;* no work is done unless there is motion;

and this motion must be in the direction of the force. Thus, a pillar supporting a building does no work, neither does a horizontal table on which a ball rolls.

3. In the expression for the kinetic energy s is the *speed*, not the *velocity;* in other words, kinetic energy does not depend upon the direction of motion. This is evident, because to produce a change of direction (and no change of speed) the force must be at right angles to the direction of the motion, and therefore, by what has just been said, no work is done.

4. The same relation between work and change in kinetic energy holds true even if the force is not constant; because we can in that case consider the force as constant for a short distance, during which the formula holds, then assume another constant value for a short distance, etc. The total work done — that is, the sum of the amounts done by the separate forces — will then equal the total change in the kinetic energy.

Illustrations. — Let us consider several cases of motion from the standpoint of both momentum and kinetic energy. When a ball is thrown, the momentum gained depends upon the impulse of the force, *i.e.* upon the time during which it acts; while the kinetic energy gained depends upon the work done by the force, *i.e.* upon the distance through which it acts. If a bullet is set in motion by a compressed spring, as in a toy gun, the spring in relaxing exerts an impulse and also does work, producing momentum and kinetic energy. As a bullet fired from a rifle enters a wooden target, the *distance* it penetrates depends upon its kinetic energy; the *interval of time* required to bring it to rest depends upon its *momentum*. The destructive power of a moving body is due to its kinetic energy; its power to strike a blow, to its momentum.

Conservative Forces. — For the time being we shall limit ourselves to the consideration of what we may call purely mechanical forces, *e.g.* those produced by gravity, by elastic bodies which are deformed, such as a bent bow, compressed spring, etc. In all these cases we make the assumption, which is based upon the results of experiments, that the work done by a force during a displacement from one point to another depends upon the initial and final positions, *not upon the path followed;* such forces are called "conservative" for reasons that will soon appear.

This may be illustrated by the force of gravity. Describe two horizontal planes $\overline{PQ}$ and $\overline{P_1Q_1}$ at a distance h apart. Draw an oblique line $\overline{AB}$, making an angle N with the vertical line $\overline{AC}$; its length is $\frac{h}{\cos N}$. Let a particle of mass m move under the action of gravity down $\overline{BA}$; at any point the force in a vertical direction is mg, and therefore the component along the line $\overline{BA}$ is $mg \cos N$. The work done, then, between B and A is

$$mg \cos N \cdot \frac{h}{\cos N} \text{ or } mgh.$$

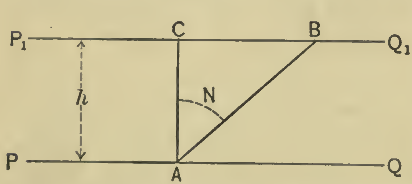


FIG. 54. — $\overline{PQ}$ and $\overline{P_1Q_1}$ are two horizontal planes at a distance apart, h.

Similarly, if the particle in passing from B to A had followed the path $\overline{BC}$ and $\overline{CA}$, the work done would have been the same; for, in the path $\overline{BC}$ the motion is perpendicular to the force of gravity, and therefore no work is done; while in the vertical motion $\overline{CA}$ the work done is mgh. Since any line curved or broken may be considered made up of straight lines, it is seen that when a particle of mass m falls from any point in one horizontal plane to any point in a lower parallel plane at a vertical distance h, the work done by gravity is mgh, and is independent of the path. This is a consequence of the fact that the force of gravity is vertical and is constant in amount at all points near the surface of the earth at any one locality.

Potential Energy. — Since, then, in all the cases that we are to consider at present the work done by forces depends upon only the initial and final points, we may write

$$F(x_2 - x_1) = V_1 - V_2,$$

where V is a quantity whose value depends simply upon the point considered. Thus, our fundamental formula becomes

$$V_1 - V_2 = \frac{1}{2} m (s_2^2 - s_1^2),$$

or

$$V_1 + \frac{1}{2} m s_1^2 = V_2 + \frac{1}{2} m s_2^2.$$

This means that during the motion the quantity $V + \frac{1}{2} m s^2$ remains constant. Consider an illustration: a ball being set in motion by a compressed spring. The above formula states that as the speed of the ball increases, the value of V decreases; or, *vice versa*, if the ball is made to strike the spring and compress it, as the speed of the ball decreases, the value of V increases. Again, in the case of a falling body, as the speed increases, the value of V decreases; and, if the ball is thrown upward, as its speed decreases, the value of V increases.

The quantity whose value is V is called "potential energy"; and it is seen by the above illustrations that when a spring is compressed, the potential energy increases; when a particle is raised vertically upward, the potential energy increases; and, conversely, when the spring relaxes or the particle falls, the potential energy decreases. We say that the "compressed spring has potential energy," and that the "system of the particle and the earth has potential energy"; or, in the latter case, more simply, "the particle itself has potential energy"; but these words are only a *description* of the experiments just mentioned. In the case of the particle and the earth, the former has not changed its size, its shape, its mass, or any of its physical properties; *it* has therefore not been changed nor has anything been added to or taken from it; but its relation to the earth has been altered. The same is true of the particles of a compressed spring; their relative positions are changed. In a similar way, a twisted wire, a bent bow, a clock spring that has been wound up, etc., all have potential energy; and, in general, a body or a system of bodies has potential energy if the particles composing it are in such a condition that a force is required to maintain it.

The formula $F(x_2 - x_1) = V_1 - V_2$ gives a means of calculating only the *change* in the potential energy; and so what is meant by "the potential energy for a given position or condition" is the work required to bring the system into that condition from some other one which is taken as the standard one. Thus, in dealing with gravity, it is customary to reckon from the surface of the earth; and the potential energy of a particle of mass m at a vertical height h above the earth's surface is therefore mgh. In compressing a spring, the standard condition is that when the spring is entirely relaxed; and since experiments show that the force which the spring exerts at any instant when compressed varies directly as the amount of compression, this force may be written cx, where x is the compression and c is a constant to indicate the proportionality; but, as the spring is compressed more and more, the force varies, and therefore during the compression from 0 to x, the mean value of the force is $\frac{1}{2}cx$ (the average of 0 and cx — see page 32); and the potential energy of the compressed spring is the product of this mean value of the force by the distance, *i.e.* it is $\frac{1}{2}cx^2$.

Conservation of Energy. — If a particle has kinetic energy, or if a system has potential energy, it is in a condition such that it can do work. A falling body can compress a spring or bend a board, thus overcoming a force; or it may strike another body and change its speed, thus doing work also. Similarly, a bent bow may change the speed of an arrow or it may raise a body up from the earth. Two things should be noted: (1) There are two ways of doing work corresponding to the two types of forces referred to on page 68, namely, producing acceleration in a particle, in which case it gains kinetic energy; or overcoming some opposing force, *e.g.* gravity, in which case the system on which the work is done gains potential energy. (2) If a particle or a system does work on another particle or a system, the latter gains energy and the former loses energy. The exact relation between this gain and loss is stated in the general formula $V + \frac{1}{2} ms^2 =$ constant, which is true only for so-called "conservative" forces. (See page 108.) This says that, if a system has both kinetic and potential energies, the sum of the two remains constant; if one decreases, the other increases by an equal amount. This is a special case of the principle of the "Conservation of Energy." (See page 115.) Thus, if one part of a system does work on the other, *e.g.* a compressed spring and a ball, a bent bow and arrow, one loses a certain amount of energy, the other gains it. Similarly, in the system made up of the earth and a falling body, the potential energy decreases by an amount equal to that by which the kinetic energy increases. If there are no changes in the potential energy of a system, the total kinetic energy does not change, the loss in one part equals the gain in another; an illustration is given by the impact of two billiard balls. (All cases of impact between inelastic bodies are excluded from consideration here, because the forces acting during the impact of such bodies do not satisfy our assumption made above in regard to mechanical forces. As will be shown shortly, part of the energy in the case of impact of inelastic bodies disappears from view and is manifest to our senses in the production of heat effects, such as rise of temperature, etc.)

Unit of Work and Energy. — The act of transfer of energy from one particle or one system to another involves what we have called "work." Its numerical value is, from its definition, the product of the values of the force and the displacement in the direction of the force. Work, kinetic energy, and potential energy are, then, all similar quantities and are all measured in terms of the same unit. On the C. G. S. system, this is the work done by a force of one dyne in a displacement of one centimetre, or the energy of a particle whose mass is two grams moving with a speed of one centimetre per second; it is called an "erg"; but it is too small for practical purposes, and so 10,000,000, or 10^7, ergs is the unit in common use; it is called a "joule," in honor of the great English physicist who did so much to teach correct ideas in regard to energy. Another unit often used is the "foot pound," or the work required to raise in a vertical direction a distance of one foot a body whose mass is one pound. This, then, equals approximately 1.356 joules, assuming g to equal 980.

Motion in a Vertical Circle. — One case of transfer of energy deserves special notice; it is that of a particle suspended by a massless cord and making vibrations in a vertical circle under the action of gravity. As it swings through its lowest point, it has its greatest kinetic energy and its least potential; and, as it gradually rises, the former decreases and the latter increases, until at the end of its swing it has its energy entirely in the potential form. If the arc of vibration is extremely small, this particle is a simple pendulum; but even when the arc is large, we can deduce certain general laws. If the particle moves freely from P_1 downward along the circle of radius r, starting from rest, its speed at the bottom A is such that

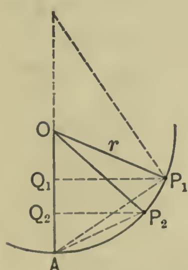


FIG. 55. — Motion of a particle in a vertical circle under the action of gravity.

$$s^2 = 2\,g\,\overline{Q_1A},$$

where Q_1 is the projection of P_1 on the vertical diameter.

But, by geometry, we know that

$$\overline{Q_1A} = \frac{\overline{P_1A}^2}{2\,r},$$

where $\overline{P_1A}$ is the chord, and r is the radius.

Hence,
$$s = \overline{P_1A}\sqrt{\frac{g}{r}}.$$

If the particle had moved from P_2 down the circle, its speed at the bottom would have been

$$s = \overline{P_2A}\sqrt{\frac{g}{r}}.$$

So the speed at the bottom of the path is directly proportional to the length of the chord of the arc through which it falls.

Work and Energy in Motion of Rotation. — When work is done in producing angular acceleration of a rigid body about a fixed axis, somewhat different expressions for the work and the kinetic energy may be deduced, which are more useful. Let a plane section be taken through the body, perpendicular to the axis, and passing through the point of application of the force, and let the force be resolved into one parallel to the axis and one parallel to this plane. In the cut let P be the trace of the pivot; F, the component of the force, whose point of application is O and whose line of action is $\overline{OA}$; and $\overline{PA}$, the lever arm of length l; and call the length of the line $\overline{PO}$, r. In the motion neither the component of the force parallel to the axis nor the force exerted by the pivot on the body does any work. (We may suppose that gravity does not act; or, the axis may be considered as being vertical.) The force F may be resolved into two components,

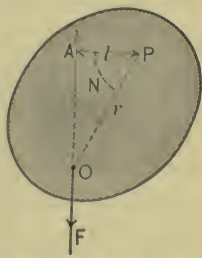


FIG. 56. — Plane section of a rigid body perpendicular to axis at P; F is the component of the force in this plane.

one along $\overline{OP}$ which therefore does no work, the other perpendicular to $\overline{OP}$. Calling the angle (OPA) N, this last component has the value $F \cos N$; and, as the body turns, this component remaining always perpendicular to the line $\overline{OP}$, the work done equals the product of $F \cos N$ and the length of the arc described by the point O. This arc equals the product of the angle turned through and the length of the radius $\overline{PO}$. Calling this angle M, the work is then $rMF \cos N$; but $r \cos N = l$, the lever arm; and $Fl = L$, the moment of the force F around the axis through P. Consequently the work equals the product of the *moment of the force by the angular displacement.*

The kinetic energy of the rotating rigid body, which we may consider made up of particles, is the sum of the energy of these taken separately. A particle of mass m_1 at the distance r_1 from the axis has the linear speed r_1h, if h is the angular speed of the body, and therefore the kinetic energy $\frac{1}{2} m_1(r_1h)^2$ or $\frac{1}{2} m_1r_1{}^2h^2$. The sum of the energy of all the particles is then $\frac{1}{2}(m_1r_1{}^2 + m_2r_2{}^2 + \cdots)h^2$ or $\frac{1}{2} Ih^2$, where I is the moment of inertia about the given axis. (Notice the exact correspondence of the values for work and energy in translation and rotation.)

Properties of Potential Energy. — The connection between potential energy and force deserves consideration. The formula of definition is $F(x_2 - x_1) = V_1 - V_2$. This is equivalent to saying that, if by a displacement from a point 1 to a point 2, the potential energy *decreases*, *i.e.* if $V_1 > V_2$, F is positive, and is, therefore, in the direction from point 1 to point 2; so there is a force acting in the *direction of the displacement*, whose value is $\frac{V_1 - V_2}{x_2 - x_1}$. This is illustrated by the force of gravity, that of a compressed spring, etc.

Another interpretation is this: in a system left to itself under the action of its own forces, motions take place, if at all, in such a manner as to produce a decrease in the potential energy. If a system is in equilibrium, the total force in any direction is zero, and therefore any slight displacement may be produced without there being any work done; hence the potential energy remains constant during these displacements. This shows, then, that at a position of equilibrium, the potential energy has either a maximum or a minimum value. If the equilibrium is stable, it is not difficult to prove that the value of the potential energy is a minimum; as is illustrated by a pendulum being in stable equilibrium when it is hanging at rest at the bottom of its path.

Power. — The rate at which work is done, that is — if this rate is uniform — the work done in a unit time, is called the "activity" or "power." On the C. G. S. system the unit is "one erg per second"; but for practical purposes "one joule per second" is taken: this is called a "watt," in honor of the great Scottish engineer who made so many improvements in the steam engine. Since work equals the product of force by displacement or of moment of force by angular displacement, power equals the product of force by linear speed or of moment by angular speed.

Power is also often measured in terms of a unit called "one horse power," which is defined to be 33,000 foot pounds per minute. This equals 746 watts approximately.

Other Forms of Energy

Conservation of Energy. — There are many other forms of forces than those which have been considered. Some of these correspond to forms of potential energy, such as the surface tension of liquids, forces due to electric charges or to magnets, etc.; others, however, do not. Among the latter the force of friction is the most important. This force is manifest whenever two pieces of matter in contact with each other move relatively; and in all the cases of motion so far discussed the conditions have been so expressed as to assume the complete absence of friction. It is a force that always opposes the motion: and its numerical properties will be discussed later. Let us consider several cases of friction and the immediate results. If two blocks of ice are rubbed together, some of the ice is melted; if the bearing between a wheel and its axle is not well lubricated so as to avoid friction, there is a "hot box," the bearings become hot and the parts expand; if a paddle is stirred rapidly in water, thus producing friction between different currents of the water, the temperature of the water rises and it will finally boil; if an inelastic body, like a piece of lead or putty, is deformed, different layers move over one another, there is friction, and the temperature rises. These various changes — melting, boiling, rise in temperature, and expansion — are called "heat effects," and will be discussed more fully in the next section of this book; but what is of fundamental importance here is to recognize that these effects are all produced when work is done against friction. It will be shown later how we can measure these effects numerically; and experiments show that their amount is proportional to the work thus done against friction. In doing this work, energy is lost by the body or system doing the work; and so it is natural to assume that the heat effects are manifestations of the addition of energy to those parts of the bodies which are directly affected by the friction, namely, the most minute portions — in certain cases, the molecules. This assumption is completely supported by all experiments and observations.

We have seen that in the transfer of energy in purely mechanical cases, there is no loss — what one body loses, another gains; so we extend this idea to all processes in nature, and state our belief that in no case is there any loss in energy. It may be present as energy of bodies of sensible size, of molecules or their parts, or of the ether. This statement is called the "Principle of the Conservation of Energy."

Nature of Potential Energy. — A few words more should be said in regard to what is meant by "potential energy." As we have used the expression, it describes a condition of a body with reference to other bodies or of the parts of a body with reference to each other, which is primarily concerned with the idea of force and its production. We cannot explain it in terms of such simple quantities as intervals of space or time and mass. We understand, however, its transformation into kinetic energy; and it is possible that it is the manifestation to our senses of the existence of the kinetic energy of portions of a medium which has inertia and which is intimately connected with ordinary matter, but which does not appeal directly to us.

Friction

External and Internal Friction. — As has been said in discussing different methods of doing work, there is a force that opposes the relative motions of any two pieces of matter that are in contact; this is called the force of "friction." Its discussion does not properly belong to mechanics; but it is convenient to give it here.

Distinction must be made between two kinds of friction, internal and external. The latter is illustrated when one solid body is made to move in contact with another, or when one layer of a fluid flows past another; *e.g.* a block of wood moving over a table, currents of water produced in a vessel by stirring a paddle in it, currents in the air produced by blowing. In all these cases the relative motion is soon stopped unless some force maintains it. Friction between moving layers of fluids is said to be due to "viscosity." Internal friction is illustrated when a solid body is deformed in any way, for in every case, to a greater or less extent, portions of the body move over each other. The only case of friction which will be considered now is that of one solid moving over another; the discussion of viscosity and of internal friction is deferred.

Sliding Friction. — The most important cases of friction between solid bodies are those when the two surfaces in contact are plane and when one body rolls on the plane surface of another. It will be seen that the explanation of the friction in these two cases is quite different.

Consider the motion of a rectangular block over a plane. Let AB be the section of the plane by the paper and $CDEF$ be that of the block. Let a force whose value is F produce acceleration of the block parallel to the plane surface; there is a force of friction opposing this, call its value F_1; then the total force producing the acceleration is $(F - F_1)$, and if m is the mass of the moving block, its acceleration is $\frac{F - F_1}{m}$. If there is no acceleration, and therefore the speed remains constant, the applied force exactly balances the friction, $F = F_1$; and we have thus an experimental method for determining the force of friction between two given materials, over a definite area of contact, when a definite force presses the two solids together and when the speed has a definite value.

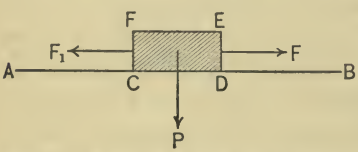


FIG. 57. — Motion of a block along a rough plane. F is the force producing the motion.

Experiments show that the friction is independent of the relative speed of the bodies, if this is small, and of the area of contact, provided the force pressing the two bodies together remains constant; that it varies directly as this force, and that it is different for different materials and varies with their condition. If P is the force pressing the bodies together and F the force of friction, the above statements may be expressed, $F = cP$, where c is a constant for two definite bodies in a definite condition. It is known as the "coefficient of friction."

A simple method for the determination of c is as follows: Let a block be placed upon an inclined plane which is gradually made more and more steep until when the block is given a slight velocity by a push, it maintains its motion unchanged. Let $\overline{AB}$ represent the inclined plane and M the block moving over it. If m is the mass of the block, and N the angle of inclination of the plane, the force with which the block is pressed against the plane P is $mg \cos N$, and the force parallel to the plane due to gravity and friction is $mg \sin N - F$. But $F = cP = c\, mg \cos N$; and therefore, if the plane is so tipped that there is no acceleration, as above described, the force parallel to the plane must be zero; that is, N must be such that

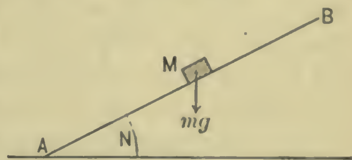


FIG. 58. — Motion of a block down a rough inclined plane.

$$mg \sin N - c\, mg \cos N = 0.$$

Hence $$c = \tan N;$$

and so can be measured. (If the body is at *rest* on the plane, it must be tipped farther than in the experiment just described before motion will begin. The coefficient, then, for what may be called "statical" friction is greater than for that which may be called "kinetic.") This coefficient of friction varies greatly with the condition of the surfaces of the body, with the lubrication, etc.

Rolling Friction. — When one solid *rolls* on another, as for instance a cylinder on a plane, there is no friction, properly speaking, because in a rolling motion there is no sliding of one surface over another; but yet experiments show that a force is necessary in order to maintain the cylinder rolling at a constant speed over a horizontal plane. This is due to the fact that the latter surface does not in general remain horizontal, but is deformed slightly by the moving cylinder in such a manner that there is a hump in front of the latter which it starts to ascend before it moves on. This so-called "rolling friction" is then due to the fact that a minute force is required to keep the rolling body from sliding back along the hump. In some cases the plane remains horizontal, but

Fig. 59. — As a body rolls on a plane surface, both are deformed slightly.

the surface of the cylinder is flattened where it is in contact with the plane. The explanation of the "friction" is evident. Rolling friction is always much less than sliding.

Dynamometers. — Friction occurs in all actual mechanisms where there are moving parts, but is always diminished in practice as much as possible by all known means. There is one form of instrument, however, that depends upon its presence, the "friction dynamometer," which is devised to measure the power furnished by a revolving shaft. One simple form of this instrument is represented in the cut. The revolving shaft whose cross section is shown at A is clamped between two halves of a block by means of bolts; and this block is kept from turning by means of a weight attached to a lever as shown. If the adjustment is exact, so that the block does not move in either direction, the moment on it due to the weight must exactly balance that due to the friction between it and the revolving shaft. Call this latter moment L, the hanging weight mg, and its lever arm l; then, when there is a balance, $L = mgl$. But as the

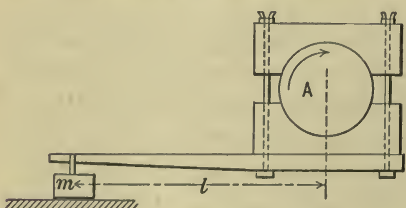


Fig. 60. — A Prony Brake. The turning moment on the block, due to the friction as the shaft A revolves, is balanced by the weight of the hanging body m.

shaft revolves it does work against this frictional moment; and, if h is the angular speed of the shaft, the *power* it furnishes is Lh or $mglh$. Therefore by applying a brake of this kind and balancing it, the power may be measured. There are many modifications of this instrument, the most important one being the substitution of liquid friction for solid. A set of blades is fastened to the shaft, which is then inclosed in a cylinder containing water and so supported that it may be kept from turning by suitable levers. As the shaft revolves, currents are produced in the water and a moment is required to keep the outer cylinder from turning; if L is the amount of this moment and h the angular speed of the shaft, the power it furnishes is Lh. (The energy goes into heat effects.)

Machines

General Principles. — A mechanism consisting of rigid or inextensible parts by means of which energy is transferred from one point to another is called a machine. A force applied at one point overcomes another force applied at a different point. Thus, the lever forming a pump handle is a machine, because when work is done by the man pumping at one end, the lever does work in raising the water at the other end. A pulley over which a cord passes is a machine, because if one end of the cord is attached to a body and the other is held by a man, the latter by doing work pulling on the cord may raise the body off the ground.

In all machines there are parts which move over each other and therefore produce friction. Consequently, a machine never delivers as much energy as it receives; part of the latter is spent in overcoming friction and is therefore "lost" in heat effects. Again, if the force acting produces an acceleration of the working parts of the machine, part of the work done is thus spent in producing kinetic energy and not in overcoming a force; while, if the velocity of the moving parts of the machine is decreasing, they themselves do work in helping to overcome the opposing force. Therefore, *in the discussion that follows, the effect of friction will be neglected; and we shall assume that there is no acceleration.* Consequently, the energy furnished the machine equals the work it does. Although this is the case, the *force* which the machine exerts need not be the same in amount or in direction as that which is exerted on it; for the distances through which these forces are displaced need not be the same. In fact, the distinct object of a machine is to obtain in return for a given force a larger one in a suitable direction.

The problem, then, is to determine the connection between the force that is doing work on the machine and the force that is being overcome by the machine. The ratio of the latter force to the former is called the mechanical advantage of the machine. There are two general methods for the discussion of this problem. One is to consider the energy relations; that is, to express by equations the fact that the work done by one force equals that done against the other. The other method is to express the fact that the two forces are keeping the machine in a condition of equilibrium, because we have assumed that there is no acceleration. We shall discuss a few simple machines and deduce the ratio of the two forces.

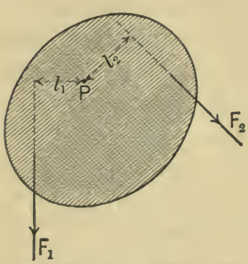


Fig. 61. — Equilibrium of a rigid body which is pivoted at P, when under the action of two forces F_1 and F_2. Principle of the lever.

The Lever. — In its general form the lever consists of any rigid body capable of rotation around a fixed axis; but most levers in actual use consist of straight rods. If a force F_1 is applied to this body at any point, it will have a moment around the axis equal to F_1l_1, where l_1 is the lever arm; and, if there is motion, the work done by the force is the product of this moment by the angular displacement. If this moment does work by overcoming another force F_2 acting on the body, whose lever arm is l_2, the work done is F_2l_2 multiplied by the angular displacement. But since the body is rigid, the angular displacements of all points are the same; and, therefore, since the force due to the reaction of the pivot does no work, the work done by F_1 equals that done against F_2, *i.e.* F_1l_1 equals F_2l_2 or $\frac{F_2}{F_1} = \frac{l_1}{l_2}$.

This relation also expresses the fact that the body is in equilibrium under the action of two forces whose values are F_1 and F_2, and which are properly directed. Illustrations of the lever are given by a pump handle, a crowbar, a pair of scissors, a pair of tongs, nutcrackers, etc. The formula for a lever was given first by Archimedes in the simple case of a straight bar acted on by two weights. The more general case was first solved by Leonardo da Vinci.

The Pulley. — A pulley consists of a circular wheel which has a groove in its edge to hold a cord, and which turns on an axle supported by a framework called the "block." A pulley is used in two ways: in one its block is fastened to a firm support, and in the other it is kept from falling by being carried in the bight of a cord passing round the wheel. The former arrangement is called a "fixed pulley"; the latter, a "free" one. In the former case, if a force F_1 is applied to one end of the cord and does the work F_1x_1 by producing the displacement x_1, it can overcome a force F_2 at the other end whose displacement is x_2, provided $F_1x_1 = F_2x_2$. But if the cord is inextensible, $x_1 = x_2$; and hence $F_1 = F_2$. This is the same relation that follows if the pulley is in

FIG. 62. — A simple form of pulley.

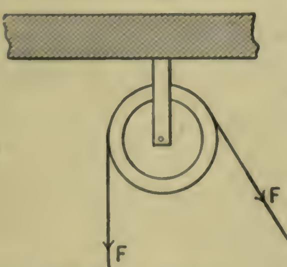


FIG. 63. — A fixed pulley.

equilibrium under the two forces F_1 and F_2. Thus a pulley simply changes the direction of the force.

As an illustration of a free pulley, let the arrangement be as shown in the cut: the cord, one end of which is fastened to a fixed support, passes under the pulley, and its two branches make an angle N with the vertical; to the free end of the cord a force F_1 is applied which balances a force F_2 applied to the block vertically downward. Since a pulley simply changes the direction of a force, there are two upward forces acting on the block owing to the two branches of the cord, which are equal in amount but inclined at equal angles N to the vertical on opposite sides. If the end of the cord is displaced *in the direction of the force* a distance x_1, the work done is x_1F_1; but the vertical displacement of the block owing to this motion is x_2, where $x_1 = 2x_2 \cos N$. The work done against the force F_2 is F_2x_2; and therefore $F_1x_1 = F_2x_2 = \frac{F_2x_1}{2\cos N}$, or, $2F_1 \cos N = F_2$.

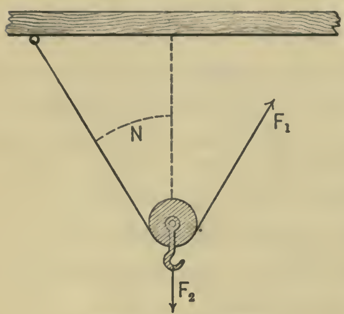


Fig. 64. — A free pulley.

The proof of this relation between x_1 and x_2 is as follows: Let AOB be the original position of the cord; and ACD its final position after the displacement; so that $\overline{AD} = \overline{AO} + \overline{OB}$. Lay off $\overline{AC}$ equal to $\overline{OB}$, thus making $\overline{CD} = \overline{AO}$; and drop a perpendicular $\overline{BE}$ upon $\overline{AD}$. Then the vertical displacement of O, *i.e.* x_2, is $\overline{AO}\cos N$; and the displacement of B parallel to the line $\overline{OB}$, *i.e.* x_1, is $\overline{ED}$. But it is evident from the figure that $\overline{CD} = \overline{AO} = \overline{CB}$; and hence

$$\overline{ED} = \overline{CD} + \overline{CB}\cos(BCE) = \overline{AO}(1 + \cos 2N)$$
$$= 2\,\overline{AO}\cos^2 N.$$

Fig. 64*a*. — Geometrical treatment of the motion of a free pulley. *B* is the end of the cord before displacement; *D*, its position after.

Hence,

$$x_1 = 2\,\overline{AO}\cos^2 N,$$

$$x_2 = \overline{AO}\cos N.$$

$$\therefore \frac{x_1}{x_2} = 2\cos N.$$

This is, again, the condition that the block should be in equilibrium under the three forces.

If, as is usually the case, the two branches of the cord are themselves vertical, $N = 0$, $\cos N = 1$; and $2F_1 = F_2$, *i.e.* $\frac{F_2}{F_1} = 2.$

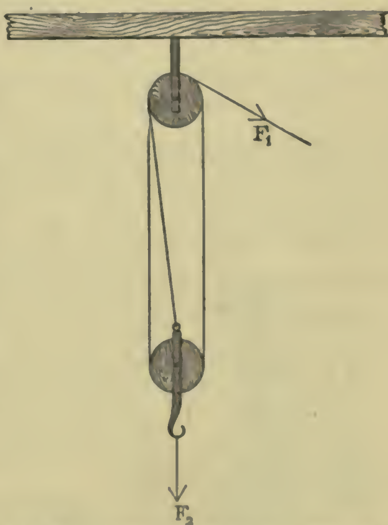


Fig. 65. — Combination of pulleys: upper fixed pulley has two free wheels; lower pulley has one wheel.

Pulleys are combined in many ways; but all forms can be easily explained on the above principles. Thus let there be a free pulley and a fixed pulley with two wheels turning independently on the same axle; let a cord, one end of which is fastened to the block of the free pulley, be run over them as shown; and let a vertical force act downward on the free pulley. Let the pulleys be so small compared with their distances apart that the various branches of the cord between the pulleys are parallel. If a force F_1 is applied to the free end of the cord and produces a displacement x_1, the free pulley will rise a vertical distance $x_2 = \frac{x_1}{3}$; and therefore a vertical force F_2 can be overcome, whose value is given by

$$F_1 x_1 = F_2 x_2, \text{ or } F_2 = 3F_1, \text{ i.e. } \frac{F_2}{F_1} = 3.$$

This is evidently the condition that the free pulley should be in equilibrium under the action of F_1 and F_2.

The formulæ for pulleys were first deduced by Stevin about 1600.

The Inclined Plane. — If a body of mass m_1 is moved upward along an inclined plane whose angle of inclination is N, the force $m_1 g \sin N$ must be overcome. So, if a body of mass m_1 lying on an inclined plane is joined by an inextensible cord, parallel to the plane and passing over a pulley, to a body of mass m_2 which hangs free, the weight of the latter is just sufficient to balance that of the former provided that $m_2 g = m_1 g \sin N$, or $m_2 = m_1 \sin N$.

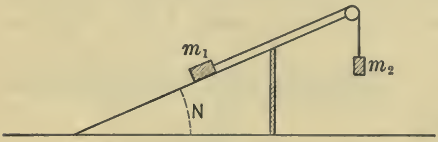


Fig. 66. — Principle of the inclined plane.

If the first body is displaced up the plane a distance x_1, the second will sink an equal distance x_1. Hence the two amounts of work, $m_1 g \times \sin N$ and $m_2 g x_1$, are equal. That is, $m_1 \sin N = m_2$, as before.

The theory of the inclined plane was also given by Stevin.

The Screw. — If a piece of paper the shape of a right-angle triangle (ABD) is cut out and wrapped around a circular cylinder, the edge $\overline{AD}$ being kept perpendicular to the axis of the cylinder, the hypotenuse $\overline{AB}$ will trace a spiral line on the surface of the cylinder. If a groove is cut following this line, we have what is called a cylindrical "screw." The portions of the cylinder between the grooves are called the "threads" of the screw; and the distance from the edge of one thread

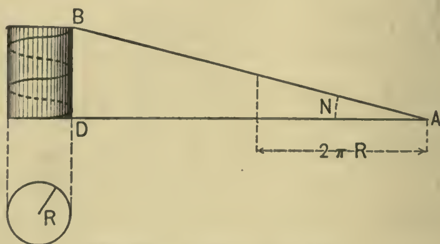


Fig. 67. — Principle of the screw.

to the corresponding edge of the next measured parallel to the axis of the cylinder is called the "pitch" of the screw. If a distance equal to the circumference of the cylinder is marked off on the line $\overline{AD}$, beginning at A, and a perpendicular line erected, the length included between the base and the hypotenuse is evidently equal to the pitch of the screw. If R is the radius of the cylinder, the length of the circumference is $2\pi R$; and the vertical distance on the triangular piece of paper which corresponds to this horizontal distance is given by the proportion,

$$\text{vertical distance} : 2\pi R = \overline{BD} : \overline{AD};$$

or, the vertical distance equals $2\pi R \tan N$. This, then, is the pitch.

If a spiral groove is made on the inner surface of a hollow cylinder, we have what is called a "nut"; and by cutting the grooves of a screw and a nut at the same pitch and of suitable depths, the screw will fit inside the nut. Then, if the nut is held fixed, the screw can advance through it by a rotation on its axis; or, if the screw is fixed, the nut can only advance along it by a rotation round its axis. In either case, for one complete rotation, the advance equals the pitch of the screw.

FIG. 68. – Screw and nut lifting jack.

If a moment L is applied to the screw turning in a fixed nut, and if by means of this a force F_2 is overcome, which is so applied as to oppose the advance of the screw, the work done by the moment in one rotation is the product of the moment by the angle 2π, *i.e.* $2\pi L$, and that done against the force F_2 is $F_2 2\pi R \tan N$. If the moment L is due to a force F_1 applied to the circumference of the screw in such a direction as to have the lever arm R, $L = F_1 R$. Consequently

$$2\pi F_1 R = F_2 2\pi R \tan N,$$

or $$F_1 = F_2 \tan N;$$

so that $$\frac{F_2}{F_1} = \frac{1}{\tan N}.$$

Screws are used in letter presses, cotton presses, lifting jacks, etc.

It is thus seen that by these various machines a force may be magnified as much as desired, with a corresponding decrease in the displacement; that the direction of a force may be changed; and that a moment may produce a force.

Chemical Balance. — This is not a machine in the ordinary use of this word, but is an instrument involving the principle of the lever, which is used to determine when the weights of two bodies are equal. It consists essentially of a horizontal beam, carrying at its ends, by means of knife edges, two "pans," and supported at its middle point by a knife edge which rests on a fixed support. The bodies whose weights are to be compared are placed one in each pan; and by adding "weights" from a set, the balance may be brought to a state of equilibrium. The arrangement of the parts of the apparatus is such that the centre of gravity comes below the knife edge. For details of the instrument, reference may be made to Ames and Bliss, *Manual of Experiments in Physics*, page 151.

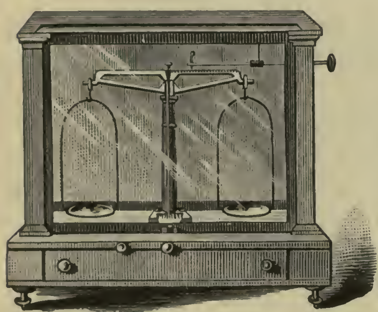


Fig. 69. — Chemical balance.

CHAPTER III

GRAVITATION

Law of Universal Gravitation. — The property of matter that is called inertia and forms the basis of dynamics was not recognized until Sir Isaac Newton stated the laws of motion in his great treatise, *Philosophiæ Naturalis Principia Mathematica*, which was published in 1687; but the property of weight was familiar to every philosopher. The laws of falling bodies were first stated in 1638 by Galileo, who also showed that all bodies fall with the same acceleration — neglecting the effect of the air. Previous to this, in 1609 and 1618, Kepler had announced the laws that bear his name concerning the motion of the planets around the sun; and in searching for an explanation of these laws and of the motion of the moon around the earth, Newton was led to a great generalization concerning matter. He thought that inasmuch as a body falls to the earth, and as there was no reason why such a phenomenon should be limited to the earth, there might be forces acting between all portions of matter in the universe, and that the moon is held in its orbit and the planets in theirs by forces of the same nature as that which draws an apple as it falls toward the earth. Other philosophers had had this idea before, notably Hooke; but no one had expressed it clearly until Newton did so in the *Principia*. The conception occurred to him as early as 1666; but for various reasons he did not make it known until later. He found that the observed facts would be explained if he assumed that this force obeyed the following law: Between two particles of matter whose masses are m_1 and m_2 and which are at a distance r apart, there is a force of attraction proportional to $\frac{m_1 m_2}{r^2}$. This can be expressed by writing $f = G\frac{m_1 m_2}{r^2}$, where G is a constant of proportionality and is independent of the material of the particles or their distance apart. This law is in accord with all known observations and experiments.

Illustrations. — We shall consider a few special cases; and, in discussing them shall make use of the fact first proved by Newton that, if the above law is true, a solid spherical body acts as if all the matter were concentrated into a particle at its centre, provided the body is homogeneous or can be regarded as made up of spherical shells each of which is homogeneous. The sun and planets and the various satellites may for present purposes be considered as satisfying these conditions.

1. *Falling bodies.* — In the case of a body falling toward the earth, we may let m_1 be the mass of the falling particle, m_2 be the mass of the earth, and r its radius. Then the force between them, acting on each, is as above, $f = \frac{G m_1 m_2}{r^2}$. Considering the motion of the falling particle, we may write $f = m_1 g$, where g takes the place of $\frac{G m_2}{r^2}$, and is therefore a constant. This is the ordinary formula for the weight of a body whose mass is m_1. If G is independent of the kind of matter in the particle and its amount, g should be also, and to test this Newton performed some experiments with pendulums; for, as has been shown in the discussion of the simple pendulum, the period of vibration is $2\pi\sqrt{\frac{l}{g}}$, where l is the length of the pendulum. So, if g is different for different kinds of matter, it would be apparent if the periods of pendulums of different materials were determined. This question was investigated by Newton and later by Bessel; and all experiments agreed in showing that g is independent of the

kind of matter. (This fact is also shown by the experiments of Galileo on the two cannon balls falling from the Leaning Tower at Pisa; for, as they fell in the same time, they had the same acceleration.)

Variations in "g." — Since the earth is not a sphere, but is slightly flattened at the poles, the distance from its centre to the surface decreases as one proceeds from the equator to either pole, and for this reason g increases. Again, owing to the fact that the earth is spinning rapidly on its axis, a certain force is required to make any particle on its surface move in its circular path; consequently a portion of the force of gravitation is spent in accomplishing this, and the difference between the two produces the acceleration of the falling body. Since the radius of the circle of motion of a particle is greatest at the equator, and therefore the centrifugal force greatest, the weight of a body is least there, or g is least. Therefore, owing to both these causes, g varies at different latitudes on the earth, increasing as the latitude is increased. Several formulæ have been advanced to connect these two variable quantities; the most satisfactory of which gives as the value of g at any latitude l, $g = 978(1 + 0.005310 \sin^2 l)$.

The fact that a pendulum of a constant length had different periods at different points on the earth's surface was recognized as early as 1671, and it was explained by Huygens as due to differences in the centrifugal force at these points. (This is, as we now know, only part of the cause.)

Of course g varies also, owing to local causes, such as the nearness of a mountain, great inequalities in the constitution of the surface of the earth, etc.; but these variations are as a rule most minute. Methods of measuring g exactly will be discussed presently.

2. *The motion of the moon.* — The moon moves around the earth in an orbit that is nearly circular with a period of approximately 27 days 8 hr.; and so there must be an acceleration toward the centre of its orbit, *i.e.* the centre of

the earth, equal to $\frac{s^2}{r}$, where r is the radius of this orbit and s is the linear speed of the moon; or, substituting for s its value in terms of the period T, $s = \frac{2\pi r}{T}$, this acceleration is $\frac{4\pi^2 r}{T^2}$. So, since r and T are both known, it can be calculated. If this acceleration toward the earth is due to gravitation and if Newton's law is true, it can also be calculated in terms of the acceleration of a falling body at the surface of the earth, *i.e.* g. For, using again the general formula, in which m_1 is the mass of a particle, m_2 that of the earth, and r the distance from the centre of the earth to the particle, $F = \frac{Gm_1m_2}{r^2}$; and therefore the acceleration of the particle toward the earth is $\frac{Gm_2}{r^2}$. Consequently, calling a the acceleration of the moon, r_1 the radius of the moon's orbit, and r_2 the radius of the earth, $a : g = \frac{1}{r_1^2} : \frac{1}{r_2^2}$; and thus, since g, r_1, and r_2 are known, a can be calculated. Newton showed that the two values, one based upon direct observation, the other upon his law of gravitation, agreed.

3. *The motions of the planets.* — As a result of a laborious study of numerous observations on the motions of the planets around the sun, and after many futile trials, Kepler succeeded in discovering three laws in regard to these motions, with which all observations are approximately in accord. These are:

(1) The areas swept over by the straight line joining a planet to the sun are directly proportional to the time; *i.e.* equal areas are described in equal intervals of time.

(2) The orbit of a planet is an ellipse, having the sun at one of its foci.

(3) The squares of the periods of different planets are proportional to the cubes of the major axes of their orbits.

Newton showed that these laws, and many slight variations from them, were direct consequences of his law of gravitation. The first law follows because the force of gravitation, acting on a planet, is always directly toward a fixed point, viz., the centre of the sun, which in the statement of Kepler's laws is supposed not to move. (Forces like this which are directed toward a fixed point are called "central forces.") The second and third laws follow because gravitation is a central force which varies inversely as the square of the distances between the bodies.

There are of course many irregularities in the motions of the moon and of the planets because of the action of other portions of matter than the earth and the sun, of variations in their distances apart, of the departure of the earth from a spherical form, etc.; but all these irregularities can be fully explained as consequences of this law of gravitation. This is the science of "Gravitational Astronomy."

4. *The "Cavendish experiment."* — Various experiments have been performed since the days of Newton to see whether the force of gravitation between bodies of ordinary size could be measured. The first of these was carried out by Cavendish in 1797–8. His method was to place two bodies on the ends of a light rod which was suspended horizontally by a fine vertical wire attached to its middle point, then to bring up near these suspended bodies two others so placed as by their force of gravitation to turn the rod and thus twist the supporting wire. He observed an effect, and measured the force exerted. This experiment has been repeated often and in various forms. (In one it was shown that the force varied with the masses of the bodies and inversely as the square of the distance.)

Having thus measured the force between two bodies of known mass at a known distance apart, and assuming Newton's law to be true, one can at once calculate the value of G in the formula. It is 0.000000066576, or 6.6576×10^{-8}.

on the C. G. S. system. This leads to a value for the mass and the average density of the earth. (By the value of the "density" of a homogeneous body is meant the ratio of the value of the mass of a certain portion of it to the value of the volume of this portion. Thus if D is the value of the density, and m and v are those of the mass and volume of any portion, $D = \frac{m}{v}$ or $m = Dv$.) We have seen that in accordance with Newton's law $g = \frac{Gm}{r^2}$, where m is the mass and r is the radius of the earth. Thus $m = \frac{gr^2}{G}$; and all the quantities in the second term are known. Further, if the earth can be considered as a sphere, its volume is $\frac{4}{3}\pi r^3$; and therefore in terms of the average density $m = \frac{4}{3}\pi r^3 D$, and accordingly $D = \frac{3}{4}\frac{g}{\pi G r}$ and can be calculated. Its value is 5.5270 on the C. G. S. system. As will be shown later, the value of the density of water at ordinary temperature does not differ far from 1 on this same system; and so the density of the earth is about $5\frac{1}{2}$ times as great as that of water.

The student should consult Mackenzie, *The Laws of Gravitation*, Scientific Memoir Series, New York, 1900.

Centre of Gravity. — A few more things should be said in regard to gravitation as we observe it here on the surface of the earth. It is a force directed toward the centre of the earth approximately; and therefore the forces acting upon the particles of a body are parallel to each other. Their resultant is called the weight of the body, and we have proved that there is a fixed point connected with the body through which this resultant always passes, however the body is turned. This point is called the "centre of gravity." If a rigid body is pivoted so as to be free to turn around a horizontal axis, but is at rest with reference to the earth, a vertical line through the centre of gravity must intersect the axis; otherwise the weight would have a moment about it, and the body would turn. The equilibrium is evidently stable if the axis is above the centre of gravity; unstable, if it is below; and neutral, if it passes through this point. The fact that the centre of gravity lies vertically below the axis of suspension when the equilibrium is stable furnishes a method for its experimental determination. If the body is suspended at a *point*, the centre of gravity must lie in the vertical line through it; and so, if the body is suspended in turn from two points, the centre of gravity must be the intersection of the two corresponding vertical lines.

Compound Pendulum. — If a body is suspended free to turn about a horizontal axis, it is called a compound pendulum; and if it is set vibrating through an infinitesimal amplitude, it will have harmonic motion. Let the cut be a section through the centre of gravity of the body G, and perpendicular to the axis at P. Call the length of the line $\overline{PG}$, h, and the angle it makes at any instant with a vertical line through P, N; this is then the angular displacement. The force mg acting vertically down through G has a moment about the axis equal to $mgh \sin N$, which is in the opposite direction to the displacement; therefore the angular acceleration is $-\frac{mgh}{I}\sin N$, where I is the moment of inertia of the body about the axis. If the amplitude is small, $\sin N$ can be replaced by N, the angle itself; and the acceleration has the value $-\frac{mgh}{I}N$. Consequently the motion is harmonic; and the period is $2\pi\sqrt{\frac{I}{mgh}}$. The period may be observed, and in pendulums of certain shapes, m, h, and I may all be measured; therefore g may be determined. Calling the period

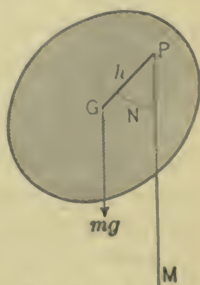


FIG. 70. — Compound pendulum. Plane section of a rigid body pivoted by a fixed horizontal axis, the plane being perpendicular to the axis P, and including the centre of gravity, G.

T, $g = 4\pi^2 \frac{I}{mhT^2}$. A simple pendulum is a special case of a compound one, in which there is only a particle vibrating; so $I = mh^2$, and the period becomes $2\pi\sqrt{\frac{h}{g}}$. A simple pendulum cannot actually be constructed; but one can be imagined of such a length that its period equals that of any compound pendulum.

By suspending the compound pendulum described above so as to vibrate in turn about different axes, all parallel to the original axis, one may be found on the opposite side of the centre of gravity, so placed that this point lies on a line perpendicular to the two axes and such that the period of vibration about it is the same as that about the original axis. This fact was deduced by Huygens as early as 1673. The distance between the two axes may be shown to be equal to the length of a simple pendulum having the same period as that of the compound one. Therefore, although it is impossible to construct a simple pendulum, a compound one may be made provided with two axes of vibration and so adjusted that the periods of vibration about them are the same. If T is this period, and l is the distance apart of the axes, $T = 2\pi\sqrt{\frac{l}{g}}$, or $g = \frac{4\pi^2 l}{T^2}$.

Galileo, in 1583, first called attention to the apparent isochronism of a pendulum; and he made use of this fact in certain observations. He also, in 1641, described a plan of using a pendulum to regulate a clock, and had a drawing of his invention made. This fact was not generally known, however; and in 1656 Huygens independently invented a pendulum clock, which came into immediate use.

Historical Sketch of Mechanics

Although the main facts in regard to the historical development of mechanical principles have been stated in connection with them, it may be well to give a brief review. Up to the time of Newton the fundamental property of matter was thought to be weight; and the only forces considered were those produced by weight. Archimedes and da Vinci had stated the laws of the lever and Stevin had explained the equilibrium of a body on an inclined plane, making use of what we call the parallelogram of forces; the "proofs" in each case were made to rest upon certain assumptions which appealed to the philosopher as being fundamental and which could not be proved themselves. Galileo made a great step in advance, because he undertook the experimental study of *dynamics*, and formulated certain statements in regard to the properties of matter in motion. He assumed that if a body were free from any force, it would continue to move in a straight line with a constant speed; he showed that the acceleration of a falling body is constant; and he deduced many well-known theorems; he further assumed what is equivalent to saying that a force acting on a body produces its action independently of the existing motion of the body. Galileo's experiments on Mechanics were published in 1638.

Newton's attention was attracted to a property of matter different from weight, and to other forces than that of weight, by his conception of the explanation of the motion of the planets. In his *Principia*, published in 1687, he proposed three laws of motion which are equivalent to the following: 1. A body left to itself will maintain its velocity constant. 2. If a body is acted on by an external force, it will receive an acceleration such that $f = ma$; forces act independently of each other. 3. Action and reaction are equal and opposite. These laws are in accord with the principles enunciated on page 59, and they have served for over two hundred years as the basis of all work in Mechanics. Newton thus introduced the ideas of mass and of the proper measure of any force.

Before Newton's ideas were accepted, there was a dispute as to the value to be assigned "quantity of motion of matter."

Descartes maintained that the proper value was wv, where w was the weight of the body and v its velocity, while Leibnitz adopted wv^2. It was shown by d'Alembert, in 1743, that both were correct; mv measures the momentum or the impulse of the force, $\frac{1}{2}mv^2$, the kinetic energy or the work done by the force.

The greatest of Newton's contemporaries was Huygens, whose treatise, *De Horologio Oscillatorio*, published in 1673, is equaled in importance only by the *Principia*. In this he discusses the motions of pendulums, simple and compound; the laws of centrifugal force, etc. He had no conception of mass and used less elementary assumptions than did Newton. He adopted in his work, as the fundamental property of matter in motion, wv^2, and showed the great importance in questions of rotation of the quantity which we call the moment of inertia.

Since the publication of the *Principia* progress in Mechanics has been (1) in a philosophical study of the nature of the postulates and definitions of Dynamics, (2) in deducing from Newton's laws other principles which are more useful for particular classes of motion.

BOOKS OF REFERENCE

MACH. Science of Mechanics. Chicago. 1893.

This gives an interesting history of the progress of Mechanics, together with a critical study of the principles on which the science is based.

POYNTING. The Mean Density of the Earth. London. 1894.

This contains a full account of the experimental investigation of Newton's Law of Gravitation.

ZIWET. Mechanics. New York. 1893.

This is a more advanced text-book than the present one, but will be found most useful for purposes of reference.

PERRY. Spinning Tops. London. 1890.

This is a most interesting series of lectures on the mechanics of spinning bodies.

WORTHINGTON. Dynamics of Rotation. London. 1892.

This is a most useful book of reference for elementary students.

CHAPTER IV

PROPERTIES OF SIZE AND SHAPE OF MATTER

Solids and Fluids; Liquids and Gases. — The most obvious property of a material body is that it has a certain shape and size, both of which can be changed by suitable forces. As has been explained before, the name "solid" is given to a body that keeps its size and shape under all ordinary conditions; and the name "fluid," to a body that yields to any force, however small, that acts in such a manner as to make one layer move over another. Fluids are divided into two classes, according as they can form drops or not; the former are called "liquids"; the latter, "gases." See Introduction, page 16.

Elasticity, Viscosity, etc. — Some bodies when deformed slightly by a force will return to their previous condition after the force is removed; they are called "elastic." Thus, glass, steel, ivory, etc., and all fluids are elastic. Certain solids, however, when deformed in a similar manner, remain so after the force ceases; they are called "inelastic" or "plastic." Such bodies continue to yield to a force as long as it is applied. In the deformation of all inelastic bodies there is a sliding of portions of matter over each other, as when a piece of lead is bent or hammered; and consequently there is what has been called "internal" friction between these parts. There is a sliding of this kind whenever any actual solid, however elastic, is deformed, although its amount may be very small. This is shown by the fact that, if a body as elastic as a glass rod or a steel tuning fork is set in vibration, the motion soon ceases, and the temperature of the body is raised slightly. Similarly, when currents are produced in a liquid or a gas by stirring them in any way, the motion soon ceases, and the temperature is found to be increased. A fluid which offers great frictional opposition to the relative motion of its parts is said to be "viscous," while those which flow easily are said to be "limpid." As will be explained later, when a fluid is made to flow through a long tube, the quantity that escapes from the open end is independent in most cases of the material of which the tube is made. This proves that the fluid layers on the inner surface of the tube stick to it, and so the fluid actually flows through a tube of the same material as itself. Consequently, in this flow the velocity is zero at the surface of the tube and increases toward the axis; and so layers of the fluid flow over each other. It requires work to accomplish this; and the quantity of fluid escaping under a given force or "head" measures the viscosity of the fluid, being inversely proportional to it. Similarly, when a solid moves in a fluid, there is a layer of the fluid attached to it, which moves, then, over other layers of the fluid. Consequently, if a pendulum vibrates, or if a disk is supported by a vertical wire which twists to and fro around its axis, making torsional vibrations, the rate of decrease in the amplitude of the oscillations measures the viscosity of the surrounding fluid. In this way the viscosity of various fluids has been measured; and it is found that it varies greatly for different fluids and with the temperature of any one fluid. Rise in temperature decreases the viscosity of a liquid, but increases that of a gas.

Diffusion. — Whenever any two gases are brought together, they mix; and after a short time the mixture is homogeneous. This process is called "diffusion." If two liquids like water and alcohol are brought in contact, one will diffuse into the other; and, even in other cases like mercury and water, where there is no apparent mingling, it may be proved that at the surface of contact there is a slight diffusion.

The most important investigation of the phenomena of diffusion was carried out by Graham (1850). He was led to divide substances into two classes — "crystalloids" and "colloids." The former diffuse much more rapidly than the latter, and can as a rule be obtained in a crystalline form, while the latter are amorphous. The mineral acids and salts are crystalloids; the gums, starch, and albumen are colloids. If the former are dissolved in water, the solutions have properties most markedly different from the water; while if the latter are dissolved in small amounts in water, they have little, if any, effect, in some cases the colloids being merely suspended in the water in a very finely divided state. If colloids are mixed with not too much water, they form jellies or membranes; and crystalloids are able to diffuse through many of these with almost as much ease as through pure water. This process is called "osmosis," and one case of it will be discussed later. (This evidently offers a method for the separation of crystalloids from colloids if there is a mixture of them. Osmosis was first observed by the Abbé Nollet in 1748, who used a piece of bladder as the membrane. Parchment paper is often used.)

Similarly, gases can pass through a thin sheet of india rubber; the latter absorbs the gas on one side and gives it off on the other. Many gases can pass through metals with ease if the latter are red hot; thus, hydrogen can pass through red-hot platinum, oxygen through red-hot silver, etc.

If two soft solids like lead and gold are brought into contact, experiments show that after the lapse of sufficient time there has been diffusion of one into the other; and the same is believed to be true to a certain extent at the surface separating any two solids, or in fact any two portions of matter.

Rates of diffusion are measured by placing two bodies in contact over a known area and determining the quantities of either which pass this surface in a given time and the distances they permeate. There are great differences in the rates of diffusion of different bodies, and these rates vary with the temperature.

Solution. — One of the most important phenomena dealing with the properties of matter is illustrated when some common salt is put into a basin of water: the salt as a solid disappears, it is said to be "dissolved." The salt is called the "dissolved substance" or "solute," and the water the "solvent." Mixtures which are homogeneous and from which the constituent parts cannot be separated by mechanical processes, are called "solutions." The formation of a solution is evidently closely connected with the process of diffusion. Similarly, we can have solutions of other solids in liquids, of solids in solids, of liquids in liquids, etc.

In the case of salt dissolving in water, it is found that, if the temperature is kept constant and more and more salt is added, a condition is reached such that, if more is introduced, it does not dissolve, but remains as a solid precipitate: the solution is said to be "saturated." If the temperature is lowered, salt will be precipitated if some solid salt is already present; otherwise, this does not in general take place. If the liquid thus contains in solution more salt than would saturate it at a given temperature, it is said to be "supersaturated"; and its condition is unstable, for by adding a minute piece of salt, all the salt in solution in excess of that required to produce saturation is precipitated. Similar phenomena are observed in many other solutions, but not in all.

It is found also that, as one substance dissolves in another, there are temperature changes; thus, as common salt is dissolved in water, the temperature of the water falls, while if sulphuric acid is dissolved in water, the temperature rises. These changes will be discussed later.

Kinetic Theory of Matter. — It is impossible to explain these facts of diffusion without assuming that the minute portions of material bodies — their molecules — are endowed with motion of translation; while, if they are free to move and are moving, the general explanation of the phenomena is at once evident and needs no statement. To account for the differences between solids, liquids, and gases, it is only necessary to assume different degrees of freedom of motion of the molecules. Since solid bodies offer great opposition, in general, to changes in size and shape, we assume that in them the molecules are held together as if by a "framework," so that they can vibrate, but cannot move from one part of the body to another unless the "framework" breaks down; this may happen with difficulty or with ease, thus causing the difference between elastic and inelastic bodies. The word "framework" is used to describe a condition, not an actual thing; we mean simply that there are forces between the molecules which hold them together exactly as the framework of a building or bridge holds it together. In a liquid we assume that the molecules are so free that they can and do move about from one point to another, but yet that the forces are sufficiently strong to prevent them on the whole from getting far apart. Of course, if a molecule strikes the surface with sufficient velocity it may escape, and thus evaporation is explained. In a gas we assume that the forces between the molecules are so minute that the freedom of motion is practically perfect; they can move freely from any point to another in the space open to them; and we think of the molecules as having rapid motion to and fro through this space. We shall show later how simple it is to explain in general terms all the properties of a gas as due to the motion of its molecules. It is not known whether these forces between the molecules which are so evident in the case of solids and liquids are due to gravitation or not, but it is at least possible.

The phenomenon of viscosity is at once explained by this assumption of moving molecules; for, if one layer of a fluid is moving over another, molecules will pass between the layers, and each one that passes from the layer flowing more slowly into the other retards the latter; while, if one moves from this layer into the former, it accelerates it. Thus, owing to the continual interchange, the two layers finally have no relative velocity; and, if one of them is at rest being in contact with a solid, all the fluid comes to rest. (This is a case where a force is explained in terms of the motion of particles of matter. See page 68.)

Coefficients of Elasticity; Hooke's Law. — When an elastic solid is subjected to a force, it will, in general, yield slightly and then come to rest, *e.g.* bending a bow, stretching a wire. This means that the changes in the molecular forces which are called into action at any point by the deformation are sufficient to neutralize the action of the external force. There is thus at any point of a deformed elastic solid a change in the position of the molecules immediately around it and a corresponding "force of restitution." When there is equilibrium between this internal force and the external one, the former may be determined from a knowledge of the latter. The elasticity of the body is measured by the ratio of this change in the internal force at any point, which is produced by the deformation of the matter near it, to the amount of this deformation. These internal forces between portions of the body are called "stresses"; and their numerical value is defined as follows: let the internal force between two portions of the body whose area of contact is A have the value F, then the limiting value of the ratio $\frac{F}{A}$ as A is made smaller and smaller is that of the stress at that point. (If the stress is uniform, it equals the force per unit area.) Owing to the deformation the internal forces are changed; and calling the change in the force ΔF, the ratio $\frac{\Delta F}{A}$ is the stress corresponding to the deformation. The deformation which produces this stress is called the "strain"; and its numerical value is defined differently, depending upon the kind of deformation. Thus, if the volume of each minute portion of the body is changed, without there being at the same time a change in shape, let v be the value of the original volume of a minute portion of the body at any point, and Δv be the change in this; then the value of the ratio $\frac{\Delta v}{v}$ is defined to be that of the strain at that point. Similarly, if the length of a wire or rod is changed by stretching or compression, the strain is defined to be the ratio of the change in length to the original length. If the shape of a solid body is changed, the measure of the strain may be defined also, as will be shown later. It is found that, if the strain is small, the corresponding stress is proportional to it: this is called Hooke's law, and was first stated by Robert Hooke in 1676, in the form "*Ut tensio sic vis.*" The ratio of the stress to the corresponding strain in any elastic body is called the "coefficient of elasticity" of that body with reference to the type of strain. Hooke's law, then, states that all coefficients of elasticity are constants for a given body; or, in more common language, the amount of the deformation of an elastic body is proportional to the force applied.

Since this proportionality between internal force of restitution and displacement is true, and since one is in the opposite direction to the other, the elastic vibrations of any body must be harmonic; because, when in the course of its vibrations the body has a certain strain, *i.e.* displacement, the elastic force of restitution is, in accordance with Hooke's law, proportional to it, and therefore the acceleration is also. Again, if Hooke's law is true, the elastic force corresponding to any displacement must be directly proportional to it, as has just been said; so, if f is the value of this force and x that of the displacement $f = cx$, where c is a factor of proportionality depending upon the nature and dimensions of the strained body and the character of the strain. As the displacement, then, increases from 0 to some value x_1, the average force of restitution is $\frac{1}{2} cx_1$, and therefore the work done in producing the displacement x_1 is the product of x_1 by $\frac{1}{2} cx_1$, or $\frac{1}{2} cx_1^2$.

A solid body can undergo two independent deformations: a change in shape and a change in volume; and corresponding to these an elastic solid has two coefficients of elasticity. If the coefficient with reference to change in volume is large, the body is said to be nearly "incompressible"; while if the one with reference to a change in shape is large, the body is said to be "rigid." A fluid, on the other hand, has only one coefficient of elasticity, that corresponding to a change in volume. Gases are very compressible; liquids are not. These kinds of matter will be discussed separately.

Density and Specific Gravity. — Before, however, proceeding to do this, a physical quantity should be defined which must be used often in the following pages. This is the "density" of a body. It is that property of a body which expresses its denseness, using this word in its ordinary meaning. If the body is homogeneous, and if m and v are the values of its mass and volume, the ratio $\frac{m}{v}$ is defined to be the value of the density. But, if the body is heterogeneous, the density *at any point* is defined to be a quantity whose value equals that of the ratio $\frac{\Delta m}{\Delta v}$ in the limit, where Δv is the volume of a small portion around the point and Δm is the value of its mass. On the C. G. S. system, the density of pure water at 4° C. is almost exactly one; for the kilogram was so constructed that its mass almost perfectly equaled that of 1000 cu. cm. of pure water at 4° C.; *i.e.* the mass of this volume of water is most approximately 1000 g., and therefore the density of the water is as stated above.

The ratio of the density of a body to that of a standard body is called its "specific gravity" with reference to the latter. Thus, if water at 4° C. is chosen as the standard substance, its density is $\frac{m}{v}$, where m is the mass of a volume v; and, if M is the mass of an equal volume of another substance, its density is $\frac{M}{v}$, and its specific gravity is therefore $\frac{M}{m}$. Thus "specific gravity" is independent of the choice of units in terms of which the mass and the volume are measured, but some standard body must be selected.

In order to measure the density of any body — solid, liquid, or gaseous — one of two methods must be followed: either its volume and mass must be measured directly, or its specific gravity must be determined with reference to some body whose density is known. The details of these different processes may be found in various Laboratory Manuals.

In the following table are given the values of the densities of some substances in ordinary use.

DENSITIES

SOLIDS

Aluminium (cast)	2.58	Iron (wrought)	7.86
Brass (about)	8.5	Iron (gray cast)	7.1
Copper	8.92	Lead	11.30
Diamond	3.52	Platinum	21.50
Glass (common)	2.6	Silver	10.5
Glass (heavy flint)	3.7	Tin	7.29
Ice at 0° C.	0.9167	Zinc	7.15

LIQUIDS

Ethyl alcohol at 0° C	0.791	Mercury	13.596
Ethyl ether at 0° C.	0.736	Water at 4° C.	1.000

Water at other temperatures, see p. 148.

Gases at 0° C. and 76 cm. of Mercury

Air (dry)	0.001293	Helium	0.00021
Argon	0.001700	Hydrogen	0.0000895
Carbon dioxide	0.001977	Nitrogen	0.001257
Chlorine	0.003133	Oxygen	0.001429

Water at Different Temperatures

0° C.	0.999878	16° C.	0.999004
1°	0.999933	17°	0.998839
2°	0.999972	18°	0.998663
3°	0.999993	19°	0.998475
4°	1.000000	20°	0.998272
5°	0.999992	21°	0.998065
6°	0.999969	22°	0.997849
7°	0.999933	23°	0.997623
8°	0.999882	24°	0.997386
9°	0.999819	25°	0.99714
10°	0.999739	26°	0.99686
11°	0.999650	27°	0.99659
12°	0.999544	28°	0.99632
13°	0.999430	29°	0.99600
14°	0.999297	30°	0.99577
15°	0.999154	31°	0.99547

CHAPTER V

SOLIDS

General Description of the Strains of a Solid. — A solid body, being characterized by a definite shape and size, can, as has been seen, be deformed in two independent ways; and, in general, under the action of forces both the size and shape are changed. These changes, if small enough, will disappear in the case of an elastic body when the force is removed; but, if the force is too large, certain permanent effects are experienced. These will be described in one particular case, that of a vertical wire whose upper end is fastened to a fixed support and to whose lower end is attached a scale pan into which weights may be loaded. So long as the stretching force is not too great, the elongation varies directly as the load; and, if this is removed, the wire returns to its original length. If the load is increased, however, a point is reached, known as the "elastic limit," such that if the force exceeds this in value, the wire acquires a permanent elongation or "set" which does not disappear when the load is removed. If the load is increased still more, the elongation becomes greater; and at length a condition is reached such that, if a greater force is applied, the extension increases very rapidly and the wire becomes plastic, because the amount of the extension now varies with the time the load acts. This point is called the "yield point." If the load is increased beyond this, the cross section of wire will contract until the "breaking point" is reached. Changes similar to these go on when the shape of a body is altered by twisting it. In what follows we shall discuss only those changes which take place below the elastic limit; so that we can consider the strain as proportional to the stress.

Change of Volume. — In this case the strain is, as has been explained, $\frac{\Delta v}{v}$, where Δv is the change in volume of a portion of the body whose volume originally was v; and, if the corresponding stress or force per unit area is Δp, the coefficient of elasticity is $\Delta p \cdot \frac{v}{\Delta v}$.

In order to produce a change in size of the minute portions of a solid without changing their shape, it is necessary to immerse the solid in a liquid and then to compress the liquid. This is done by having the liquid inclosed in a stout transparent cylinder, one end of which is closed by a piston which can be screwed in or out. Such an instrument is called a "piezometer." When the piston is pushed in, the liquid presses against the immersed solid and compresses it, the volume of each minute portion of the solid being decreased proportionally. To measure the change in volume of the solid, the latter is as a rule made in the shape of a rod; two fine parallel lines are scratched on it, one near each end; and by means of a comparator the distance apart of these lines is measured before and under the compression. If l_0 is the original length and l the length when the stress is increased by an amount Δp, experiments show that $l - l_0 = -cl_0\Delta p$, where c is a factor of proportionality and is extremely small. If Δp is measured, the value of c may be determined. If the body is homogeneous, we may assume that similar changes take place in a plane at right angles to the length of the rod. So, if a *cube* were subjected to the increase in stress Δp, its change in volume would be found as follows: the original volume v_0 is l_0^3; when the stress is increased by Δp, the volume is l^3, or $l_0^3(1 - c\Delta p)^3$, which equals $l_0^3(1 - 3\,c\Delta p)$, since c is so small that terms involving c^2 and c^3 may be neglected; hence, the change in volume is $-3\,l_0^3c\Delta p$ or

$-3 v_0 c \Delta p$. So, in the original experiment the strain corresponding to the stress Δp, *i.e.* the ratio of the decrease in volume to the original volume, is $-3 c\Delta p$; and the coefficient of elasticity corresponding to a change in volume is then $\frac{1}{3c}$. The stress Δp is due to the change in the *internal* forces, but, since there is equilibrium between it and the forces due to the pressure of the liquid, it is determined by measuring the force per unit area with which the liquid presses on the solid, as can be done in a manner to be described presently.

The result of subjecting a hollow solid body like a bottle to such a liquid pressure as this is to decrease its volume exactly as if there were no hollow spaces, provided the liquid has access to them; for each minute portion of the body has its volume decreased exactly as if there were no cavities. It is as if, before the compression, the solid were built up of large "bricks"; while, under compression, it is made up of the same number of smaller ones.

This coefficient of elasticity is sometimes called the "bulk modulus."

Change of Shape. — An illustration of the change in shape of a large solid without any appreciable change in volume is given by the following experiment. Make a rectangular block of wood; to two opposite faces fasten two boards; and push these boards sidewise in their own planes, but in opposite directions. The shape will be changed as shown, the edges of the block being now oblique; and the angle between the two positions of an edge, that is (BAB'), is taken as the measure of the strain. If the sidewise force on either board is F, there is an equal but opposite force of restitution; and, if the area of the cross section at the block is A, the stress corresponding to the strain is $\frac{F}{A}$. *This stress is acting across all the planes of the block that are parallel to the boards.* The coefficient of elasticity for a change in shape is, then, the ratio of this quantity to the angle referred to above. This coefficient is also called the "coefficient of rigidity"; and the particular kind of stress that arises when one layer is moved parallel to another is called a "shearing stress," because it is like the force produced by a pair of shears.

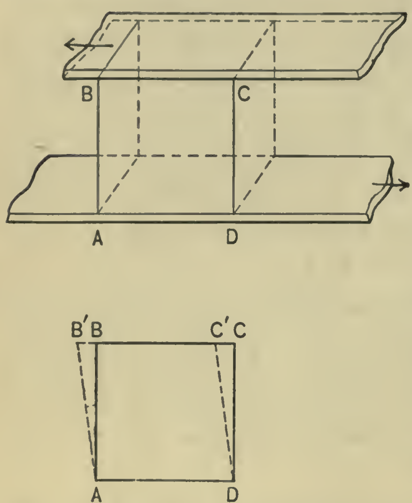


FIG. 71. — Change of shape or shearing strain. The deformation is produced by pushing the two boards in opposite directions.

When a rod or wire is twisted on its axis of figure, each minute portion of the body experiences a shearing stress, because different plane cross sections of the rod or wire are turned through different amounts; and, therefore, this is an illustration of a pure change in shape.

If one end of the rod or wire is held firm, and the other twisted by a moment L around the axis of figure, the angle through which the lower end will be turned is given by the following formula, which may be deduced by the help of higher mathematics:

$$N = \frac{lL}{nB},$$

where l is the length of the rod or wire, n is the coefficient of rigidity, and B is a constant for any one rod or wire depending upon its dimensions. For a circular cylinder of radius r, $B = \frac{\pi r^4}{2}$; and so, for a wire of circular cross section,

$$N = \frac{2\,lL}{\pi n r^4}.$$

If the wire is maintained in this state of torsion by this moment, and if it is in equilibrium, there must be in *each* plane cross section *two internal* moments equal to L, due to the elasticity of the wire; one moment acts on one face at the cross section, the other, on the opposite face; and the two moments are equal and opposite in amount. It is exactly analogous to the case of a wire or cord under the action of a stretching force; at each of its points there are two equal and opposite tensions, each equal to the stretching force. If such a wire is twisted and then set free and allowed to make torsional vibrations, the moment tending to *oppose* the motion at any instant, due to the reaction of the wire, will be

$$L = -\frac{\pi n r^4 N}{2\,l},$$

when the angle of torsion is N. Hence, if a disk whose moment of inertia around the axis is I is fastened to the free end of the wire, its angular acceleration at any instant of the vibration will be $\frac{L}{I}$ or $-\frac{\pi n r^4 N}{2\,Il}$; so the vibrations will be harmonic, and the period of one complete vibration will be

$$T = 2\,\pi\sqrt{\frac{2\,lI}{\pi n r^4}}.$$

The elasticity of a helical spring, that is, a wire coiled up in a spiral, is due to the fact that, when it is pulled out, the wire is *twisted*, not elongated.

Actually there is a sliding of one layer over another when a wire is twisted, or, in general, when there is a shearing stress; but in an *ideal* solid there would be no such sliding, if the shape were changed; and in such bodies as steel, glass, ivory, etc., the slipping is very small.

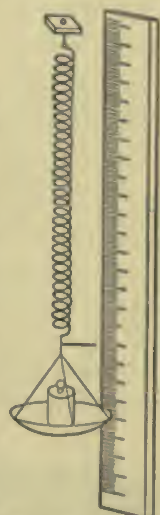


Fig. 72. — Helical or spiral spring.

Young's Modulus. — When a rod or wire is elongated, or when a pillar is compressed, both the shape and size of its minute portions are altered; and so both the coefficients previously discussed are involved. Since, however, the case is such an important one in all practical work, a new coefficient is defined which refers to the elasticity shown when a wire or rod is stretched or compressed. Let l be the original length; Δl, its increase when there is an increase ΔF in the stretching force; A, the area of the cross section. Since the wire or rod comes to equilibrium, the change in the internal force equals ΔF, too; and so the stress is $\frac{\Delta F}{A}$. The corresponding strain is $\frac{\Delta l}{l}$; and the coefficient is $\frac{\Delta F}{A} \div \frac{\Delta l}{l}$. This is called "Young's Modulus"; and, writing its value E,

$$E = \frac{l \cdot \Delta F}{A \cdot \Delta l}.$$

This is found to be a constant for any one kind of matter, *e.g.* a definite kind of brass, iron, etc.; and, if its value is known, the elongation of a particular wire or rod under the action of a given force can be calculated, because

$$\Delta l = \frac{l \cdot \Delta F}{EA}.$$

If a pillar supports a building, its length is less than if it carried no weight; but, if the weight supported becomes too great, the pillar will buckle and give way. It may be shown by a more elaborate discussion that a circular pillar pressed between two supports, and having both its ends fixed, becomes unstable when the force of compression equals $\frac{\pi E r^4}{l^2}$, where r is the radius of the pillar and l is its length.

If a beam is supported horizontally by resting on two horizontal knife edges placed near its ends, and if a weight is hung from its middle point, it will be bent. It is obvious that the lower side will be stretched, while the upper is shortened. There will therefore be a layer of lines in the beam about halfway through, which are neither lengthened nor shortened; this is called the "neutral section." This fact is made use of in constructing standards of length, such as metre bars. One form adopted has a cross section like the letter H, and the upper

FIG. 78. — Standard form of metre bar, with the divided scale along its neutral section.

surface of the "bridge" is the "neutral section"; on this the scale divisions are marked.

It is evident, further, that the coefficient of elasticity involved in flexure is Young's modulus. Formulæ giving the amount of bending for beams of different dimensions may be found in books of reference.

Impact. — When two solid bodies in motion meet, there are changes in their velocities that obey certain laws. The principle of the conservation of momentum, of course, holds. This states that, if the two centres of mass are moving along the straight line joining them, and if m and M are the two masses, v_1 and V_1 the velocities before impact, and v_2 and V_2 those after, proper attention being given the algebraic sign, $mv_1 + MV_1 = mv_2 + MV_2$. The relative velocity before impact is $v_1 - V_1$; after impact, $V_2 - v_2$; and experiments show that for two *spheres* the ratio of these quantities may be regarded as a constant, depending only on the material of the two bodies. This statement may be written $V_2 - v_2 = e(v_1 - V_1)$, where e is a constant of proportionality. It is called the "coefficient of restitution." (More accurate experiments show that e is not constant, but varies slightly with the velocities.) If the impinging bodies are not spherical, there will be rotational acceleration also, in general.

If the bodies are inelastic, $e = 0$ and $v_2 = V_2$; that is, after impact the two bodies proceed on together. If the bodies are "perfectly elastic," $e = 1$ and $V_2 - v_2 = v_1 - V_1$. If this equation is combined with that given by the principle of the conservation of momentum, it will be found equivalent to the expression of the conservation of energy:

$$\tfrac{1}{2} mv_1^2 + \tfrac{1}{2} MV_1^2 = \tfrac{1}{2} mv_2^2 + \tfrac{1}{2} MV_2^2.$$

The method by which two elastic bodies affect each other's motion during impact can be compared with what happens when two railway cars provided with spring buffers collide. The springs are gradually compressed up to a certain point and then expand, pushing the cars apart. So, when two spheres collide, they are deformed, producing elastic reactions which increase in intensity as the centres approach, and then decrease as they cause the bodies to separate.

The laws of impact were investigated and stated at about the same time, 1668, by Wren, Wallis, and Huygens. In their present form they were first given by Newton.

Energy Relation. — The potential energy of a solid in a state of strain may be deduced at once if its elastic coefficients are known. (See page 110.) As an illustration consider a stretched wire. Using the same symbols as before,

$$E = \frac{l \cdot \Delta F}{A \cdot \Delta l}, \text{ or } \Delta F = \frac{EA\Delta l}{l};$$

which means that the force required to produce an elongation Δl is $\frac{EA\Delta l}{l}$; consequently that required to produce an elongation x is $\frac{EAx}{l}$; and the potential energy of the wire when it is stretched this much is $\frac{1}{2}\frac{EAx^2}{l}$.

If a heavy body is attached to one end of a vertical spiral spring, the other end of which is fastened to a fixed support, it can be set in vertical vibrations which are harmonic. The change in the potential energy during the motion is due in part to alterations in the spring and in part to the to-and-fro motion of the body with reference to the earth; and the kinetic energy is due to the motion of the body and to that of the spring. The velocity of different parts of the spring varies; and it may be proved by higher mathematics that the kinetic energy of the system is the same as if the spring had no inertia and the mass of the suspended body were increased by one third the actual mass of the spring.

CHAPTER VI

FLUIDS

General Properties. — Fluids have been defined as those bodies which yield to any force, however small, which acts in such a manner as to cause one layer to move over another; that is, they yield to shearing forces. There are two classes of fluids: liquids and gases. The former have definite volumes, to change which requires great forces; and, if left to themselves, they form drops, but if placed in a solid vessel, assume its shape. The latter are easily compressible and assume both the shape and size of the containing vessel. When a fluid is not flowing, it is said to be at "rest," although this does not imply that the molecules are not moving; it simply means that there is no motion of portions of the fluid over each other. We shall discuss in turn the two conditions: that of rest and that of flowing.

Fluids at Rest

Thrust. — The fluid exerts a force against the walls of the vessel that contains it; and, conversely, the wall reacts against the fluid. For instance, if a toy balloon is inflated, the gas presses outward against the rubber envelope, and this presses inward, tending to compress the gas; a dam holding back a river is pressed against by the water; a tank containing water presses inward with sufficient force to withstand the outward force of the water, otherwise it bursts. Similarly, there are forces against any foreign body immersed in the fluid. If the fluid is at rest, this force between the fluid and the wall or immersed body is *perpendicular* to the separating surface at each point; for, if there were a component *parallel* to the surface, the fluid would flow. The total force acting on the surface is called the "thrust."

Fluid Pressure. — The properties at various points of a fluid are best described in terms of what is called the "pressure." If any small portion of a fluid is considered as inclosed in a solid figure with plane faces, there is a stress across each plane face due to various causes; but, if the fluid is at rest, this force is perpendicular to the face, as just explained. The limiting value of the ratio of this perpendicular force to the area over which it acts is defined to be the value of the "pressure" at the point considered, in the direction of the force. If the pressure is uniform, it equals, then, the force per unit area. We thus speak of the pressure at the bottom of a tank of water, etc.

Pressure at a Point. — At any point of a fluid at rest the pressure has the same value in all directions; for, consider a small portion of the fluid inclosed in a tetrahedron, $ABCD$, and express the condition that it shall be in equilibrium under the action of the forces on its faces. The sum of the components of these forces in any direction must equal zero.

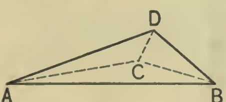


FIG. 74. — A portion of a fluid inclosed in a tetrahedron, or triangular pyramid.

Choose as this direction that of the line $\overline{AB}$; then the forces on the two faces $\overline{BAD}$ and $\overline{ABC}$ have no components parallel to this line, because they are perpendicular to it. Call the area of the face $\overline{ACD}$, A_1, and the force acting on it F_1; the area of the face $\overline{BCD}$, A_2, and the force acting on it F_2. The component of F_1 parallel to $\overline{AB}$ is F_1 multiplied by the cosine of the angle between $\overline{AB}$ and a line perpendicular to the plane $\overline{ACD}$. Imagine a plane perpendicular to $\overline{AB}$; the projection on this plane of the triangle $\overline{ACD}$ is another triangle whose area may be called A, and which is the same as the projection on this same plane of the triangle $\overline{BCD}$. Therefore $\frac{A}{A_1}$ equals the cosine of the angle between the line $\overline{AB}$ and a line perpendicular to the plane $\overline{ACD}$; and the component of the force F_1 along $\overline{AB}$ is, accordingly, $F_1 \frac{A}{A_1}$. Similarly, the component of F_2 along $\overline{AB}$ is $F_2 \frac{A}{A_2}$; and since there is equilibrium, $F_1 \frac{A}{A_1} + F_2 \frac{A}{A_2} = 0$, or $\frac{F_1}{A_1} = -\frac{F_2}{A_2}$. Therefore, in the limit, when the tetrahedron becomes infinitesimal, the *pressure* in the direction of F_1 is equal numerically to that in the direction of F_2. But these directions may be any two; and consequently the pressure at the point around which we have imagined the tetrahedron has the same numerical value in all directions.

Work done when the Volume of a Fluid is Changed. — If a fluid is contained in a cylinder one end of which is closed by a movable piston, the work done on the fluid in order to compress it may be calculated at once. If the pressure of the fluid against the piston is uniform, let its value be p (if the pressure is not uniform, let p be its mean value); then, if A is the area of the piston, the *force* that must be overcome is pA. Let the piston be displaced inward a distance x; the work done on the fluid is xpA; the decrease in volume is xA; and, therefore, the work done in compressing the fluid equals the product of the pressure against the piston by the decrease in volume of the fluid. Similarly, the work done by the fluid as it expands equals the product of the pressure by the increase in volume, provided there is no acceleration.

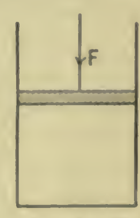


FIG. 75. — The fluid is inclosed in a cylinder by a movable piston.

The work done by a fluid, or on it, as its volume changes can be represented graphically provided the changes are so slow that there is practically a uniform condition throughout the fluid at any instant. Lay off two lines, Ov and Op, making a right angle at O; let distances along one correspond to values of the volume, and along the other, to values of pressure. The former is called the "axis of volume"; the latter, "the axis of pressure." From any point D in the plane of the two lines drop a perpendicular $\overline{DA}$ upon Ov. The point D represents the condition of a fluid whose volume is equal to $\overline{OA}$ and whose pressure is equal to $\overline{AD}$. If the pressure remains constant while the volume expands from $\overline{OA}$ to $\overline{OB}$, the new condition of the fluid is given by the point C where $\overline{BC}$ is parallel to the axis of pressure and $\overline{DC}$ to the axis of volume, and the work done *by* the fluid is equal to the area of the rectangle $ABCD$. If now the pressure suddenly increases to a value $\overline{BC'}$, no work against external forces is done, since the volume does not change. The condition of the fluid is given by the point C'; and, if the fluid again expands until the volume is $\overline{OE}$, its condition is given by the point F, and the work done by it equals the area of the rectangle $BEFC'$. It is at once evident that, if the changes in pressure and volume occur, not in a discontinuous manner as from D to C to C' to F, etc., but continuously, as represented by a *smooth curve* $\overline{PQ}$, the work done during any change in volume will be the area included between this curve, the axis of volume, and two perpendicular lines marking the initial and final volumes. If the fluid is being compressed, the changes may be represented by a curve in the opposite direction, from Q to P; and the area just described gives the work done *on* the fluid by external forces. If the curve describing the changes in the fluid is a closed one, it means

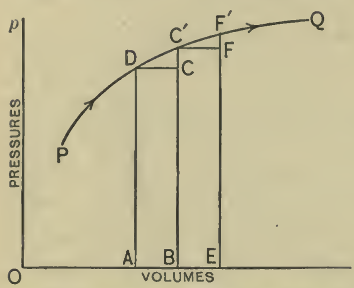


FIG. 76.—Diagram illustrating work done by a fluid as it expands.

that after a series of operations the fluid returns to its initial condition of pressure and volume; it is said to have passed through a "cycle." If the curve is a *right-handed one*, that is, if the series of changes is such that if a man were to walk from point to point along the curve the inclosed area would lie on his right hand, this inclosed area gives the total *net* work done *by* the fluid during the cycle. For, consider two portions of the curve $\overline{AB}$ and $\overline{CD}$ which are intercepted by the same two perpendicular lines through F and E; during the process represented by the curve $\overline{AB}$, the fluid does an amount of work equal to the area $ABEF$; during the process represented by the curve $\overline{CD}$, work is done on the fluid equal in amount to the area $DCEF$; consequently the excess of work done by the fluid equals the area $ABCD$. In a similar manner other pairs of portions of the curve may be considered; and, in the end, the entire work done by the fluid in excess of that done on it equals the area inclosed by the curve.

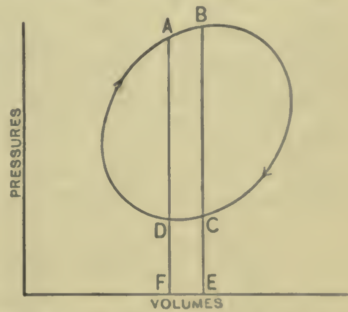


FIG. 77. — Diagram showing work done by a fluid when it passes around a cycle.

Conversely, if the curve is described in the opposite direction, that is, if it is a left-handed one, its area represents the net work done on the fluid by external forces.

Cause of Fluid Pressure. — The pressure at any point in a fluid at rest is due to two causes: (1) the reaction inward of the walls of the vessel that contains the fluid; (2) external forces, such as gravity — this is, in fact, the only such force which we need, in general, consider. Illustrations of the former cause, as shown by balloons, dams, and tanks, have been given before; but to have one where the pressure is due entirely to the containing walls, we must imagine a fluid in its vessel carried to some point where gravity and other "external" forces cease to exist; for instance, to the centre of the earth or far off in space. (There is also, of course, another pressure at any point in a fluid due to the forces between the molecules. This we cannot measure; but we can form an estimate of its value for a liquid by measuring the amount of work required to evaporate it, assuming that we can neglect this pressure for a gas.)

Pascal's Law. — Let us, then, consider the properties at the centre of the earth of a fluid inclosed in some vessel which it *fills;* for instance, a cylinder closed with pistons of different areas. To maintain the fluid in a definite condition, forces F_1 and F_2 must be applied to the pistons from without; and there is, therefore, a corresponding pressure throughout the fluid. But, since the fluid is at rest, this pressure must be the same at all points; for, if it were not, the fluid would flow from a point of high pressure to one of low, there being no force to counterbalance the difference in pressure. Therefore, *the fluid pressure due to the reaction of the walls of the containing vessel is the same at all points throughout the fluid.*

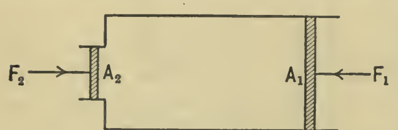


Fig. 78. — A fluid is inclosed in a vessel closed by two pistons of different areas. Neglecting gravity, $\frac{F_1}{A_1}=\frac{F_2}{A_2}$, Pascal's Law.

If this pressure is p, and if the area of one piston is A_1, the force necessary to keep it from moving outward is pA_1; and, if the area of the other piston is A_2, the force acting on it is pA_2. Thus we have a "machine" by means of which a force F_1 which equals pA_1 can balance one F_2 which equals pA_2; and so a small force may produce a great one. If the fluid is a gas, a great pressure, and therefore large forces, cannot be secured unless the volume is made very small; but if it is a liquid, the pressure may be made as great as desired without any marked decrease in volume. Thus, using water or some other liquid as the fluid in a cylinder provided with two pistons, a small force acting on the smaller piston may produce a great force by means of the larger one. This is the principle of the "hydrostatic press," which is shown in the cut. A pump forcing a small piston down produces an upward motion of the large piston; and thus a force is exerted as much greater than the original one at the pump as the area of the large piston is greater than that of the other. (Of course, since this press is used at the surface of the earth, there are additional pressures due

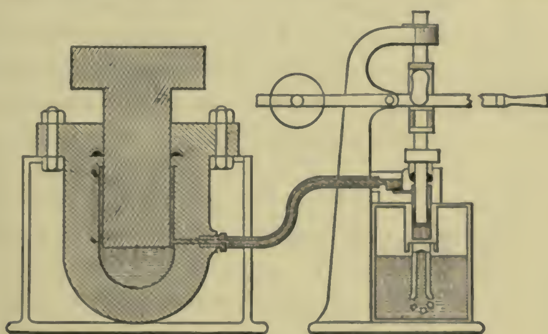


FIG. 79. — Hydrostatic press.

to gravity; but they are nearly the same on the two pistons, and in any case their effect can be neglected when compared with that due to the pistons.)

In a piezometer (see page 150) the force per unit area on the immersed solid body equals the pressure throughout the liquid; and, if the force acting on the piston and the area of the latter are known, the value of the pressure equals their ratio.

This law that the pressure is the same at all points in a fluid when not under the action of external forces was first stated by Pascal, in 1653, and is called "Pascal's Law."

Pressure due to Gravity. — The external force to which all fluids on the surface of the earth are subject is that of gravity; and since, when the fluid is at rest, the lower layers have to maintain the weight of the fluid above, there are differences in pressure at different points in a vertical line in the fluid. Thus imagine two horizontal planes at a distance h apart, and consider in them two portions having equal areas, of value A, one vertically above the other. Let their traces on the paper be $\overline{AB}$ and $\overline{CD}$. The pressure over the plane $\overline{CD}$ due to gravity depends upon the fluid above it; similarly for the pressure over the plane $\overline{AB}$; consequently the excess in the upward *force* across $\overline{AB}$ over that across $\overline{CD}$ equals the weight of the fluid contained in a cylinder whose cross section equals A, the area of either of the planes, and whose height is h, their distance apart. If d is the density of the fluid, this weight is $dhAg$, because hA is the volume and dhA the mass. Therefore the excess of *pressure* due to the vertical height h is dgh.

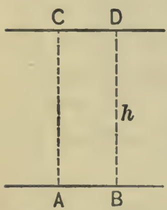


FIG. 80. — $\overline{CD}$ and $\overline{AB}$ are two horizontal planes in a fluid, at a distance apart h.

It follows that the pressure at all points in the same *horizontal* plane of a fluid is the same; because, if it were different, there would be a flowing of the fluid from the point of high pressure to that of low, as there would be no force to oppose the motion.

Archimedes' Principle. — If a solid body (or a body with a solid envelope) is immersed in a fluid at rest, or if a liquid drop or bubble is immersed in a gas, it is acted on by the pressures against its surface; and these produce a resultant force in a vertical direction upward, as can be easily seen. For if the solid body were replaced by a portion of the fluid having the same size and shape inclosed in a massless envelope, there would be identically the same surface pressures on this envelope as there were on the solid. But under these conditions the fluid in the envelope is at rest; and this shows that the forces due to the surface pressures have a resultant acting vertically upward whose amount equals the weight of the fluid in the envelope and whose line of action is through the centre of gravity of this portion of the fluid. Therefore, when a solid is immersed in a fluid, the surface forces due to the weight of the fluid combine to form a buoyant force which is equal in amount to the weight of the fluid displaced by the solid and whose line of action is vertically upward through the centre of gravity of the displaced fluid. (If the solid body is homogeneous, its centre of gravity coincides with that of the fluid displaced; otherwise, it will not in general.) This statement, having been first expressed in words by the great philosopher of Syracuse, is called "Archimedes' Principle." It is illustrated by floating balloons and soap bubbles; by all solid bodies here on the earth, whose apparent weight is therefore less than their true weight; by bodies immersed in water; etc.

If the solid is denser than the fluid, the buoyant force is less than its own weight, and it sinks; while, if its density is less than that of the fluid, it will rise. In both cases the motion takes place in such a manner as to make the potential energy decrease.

This principle furnishes a method for the determination of the specific gravity of a solid body in terms of any liquid which does not dissolve or otherwise affect it. If the weight of the solid is measured when it is hanging free, its value is mg or dvg, when d is its density and v its volume, due allowance being made for the buoyancy of the air; if it is weighed again, hanging immersed in the liquid, the *difference* in weight is d_1vg, where d_1 is the density of the liquid. Thus the ratio of the original weight to this loss in weight is $\frac{d}{d_1}$, or the specific gravity of the solid. If either density is known, the other may be calculated at once.

Unit of Pressure. — The unit of pressure on the C. G. S. system is "one dyne per square centimetre"; but, since pressures are as a rule produced and measured by using columns of liquids, a more convenient practical unit has been chosen. This is "the pressure due to a vertical height of one centimetre of mercury at the temperature of 0° Centigrade, under the force of gravity which is observed at sea level at 45° latitude." This unit is called a "centimetre of mercury"; and its value in terms of dynes per square centimetre may be calculated at once by substituting proper values in the formular $p = dgh$. On the C. G. S. system, the density of mercury at 0° C. = 13.5950; and the value of g at sea level at 45° latitude is 980.692. Consequently the pressure of *one* "centimetre of mercury" is the product of these two quantities; that is, it is 13332.5 dynes per square centimetre. The pressure of 75 cm. of mercury is called a "barie"; and its value in dynes per square centimetre is 75×13332.5, or almost exactly 10^6 dynes, *i.e.* a "mega-dyne." The pressure of 76 cm. of mercury is called "one atmosphere," because, as we shall see later, this is about the pressure of the atmosphere at sea level. Other units are "one pound per square inch," "one ton per square foot," etc., where "one pound" means the *weight* of one pound, etc.

Atmospheric Pressure. — Owing to the smallness of the density of a gas, there are only slight variations in pressure at different points in a gas confined in a reservoir of any moderate dimensions. There are, however, marked differences in the pressure of the surrounding atmosphere in which we live as one rises far above the earth's surface or goes up a mountain. This is owing to the large value of h in the above formula, which may in this manner be secured.

Owing to the presence of the gases forming the atmosphere, there is a pressure exerted by it against every solid or liquid surface with which it is in contact. This is called the "atmospheric pressure." It may be measured at any point on the earth's surface by balancing it against a column of some liquid of known density, as will be shown presently.

The fact that the atmosphere exerts a pressure on solid and liquid surfaces was first clearly understood by Torricelli, a pupil of Galileo's, and by Pascal. Conditions at the surface of the earth were in their minds comparable with those at the bottom of an ocean of water, so far as fluid pressure was concerned. Torricelli devised the famous experiment, which bears his name, of filling with mercury a long glass tube closed at one end, inverting it, and placing the open end under the surface of mercury in an open vessel, care being taken to prevent the entrance of any air. The mercury column stands in the tube to a height of about 76 cm. (at the sea level); being held up in the tube by the downward pressure of the atmosphere on the mercury in the open vessel. The space above the mercury column is called a "Torricellian vacuum"; and it is evident that the only matter present in it is mercury vapor, if the experiment has been carefully performed. This experiment was performed for Torricelli by his friend Viviani in 1643. Pascal varied it by carrying a "barometer," as this apparatus of Torricelli's is called, to different heights and noting the change in the height of the column. Many experiments to show the effects of the atmospheric pressure were devised after the air pump was invented by Von Guericke (about 1657) and improved by Boyle.

Liquids are in general contained in open vessels which they only partially fill. The atmosphere presses against the free surface of the liquid, exactly as if there were a piston pressing down on it. Therefore in the case of liquids in open vessels the pressure due to the containing walls equals the atmospheric pressure, and, as said above, this pressure is the same at all points in the liquid.

Fluids in Motion

If there is a difference in pressure between two points in a fluid, there will be motion from high to low pressure unless some force prevents it. We shall consider several cases.

Uniform Tubes. — Experiments show that, within certain limits, the quantity of fluid that flows through a tube under a difference of pressure at its two ends is independent of the material of the tube. This proves that what actually happens is that a layer of the fluid sticks to the walls of the tube, and the escaping fluid thus moves through a tube made of the same material as itself. There is therefore friction between the moving layers of the fluid, *not between the tube and the fluid;* and the quantity of fluid that escapes under definite conditions varies inversely as its viscosity.

If the tube is horizontal, the pressure is uniform throughout it, so long as there is no flow; this is called the "statical" pressure. But as soon as the motion begins, the pressure falls throughout the tube. If a constant pressure is maintained at one end A — as in water or gas mains — and if the other end B is closed, there is a uniform pressure, as just said; but if B is open so as to allow the fluid to escape, owing to the friction the pressure will decrease uniformly from A to B if the cross section of the tube is the same throughout; and, if the tube is sufficiently long, the fluid will barely flow out, however great is the pressure at A. If the tube is bent, the flow is still further decreased.

If the fluid is flowing uniformly, the quantity that passes through a cross section of the tube is the same at all points along the tube. If v is the average velocity of the fluid over any cross-section whose area is A, and d the density of the fluid, the quantity that passes in a unit of time is vAd.

The manner in which the fall in pressure along a tube and the quantity of fluid flowing through it depend upon its length and cross section, has been found as the result of numerous experiments, and is expressed in various empirical formulæ.

Irregular Tubes. — If the fluid is flowing *uniformly* through a tube of irregular cross section, the quantity passing all cross sections in a unit of time is the same; and that passing any one cross section is, as shown above, vAd, where v is the mean velocity over the section, A is the area of the section, and d is the density of the fluid at that point. If the fluid is flowing slowly, there is no change in density at different points, and therefore there is the following relation between the velocities at different cross sections, $v_1A_1 = v_2A_2$, in which v_1 is the velocity at a point where the cross section is A_1, and v_2 is that at a point where the cross section is A_2. This means that, as the cross section diminishes, the velocity increases, and *vice versa*. (Compare the flowing of a river into a lake and out through a narrow defile.) If the velocity *increases* from one point to another in the direction of motion of the fluid, there must be a *fall* of pressure in this direction so as to produce this increase in the velocity. Hence, if the velocity increases in one direction, the pressure must fall in this direction; and conversely. Wherever the velocity is greatest the pressure is least, and where the velocity is least the pressure is greatest. This is illustrated by the atomizer, the steam injector, the ball nozzle, etc.

Another illustration of this principle is furnished by the "curves" of a baseball. If a ball were in motion of translation only, everything would be symmetrical about its line of motion. As it moved through the air, a layer of gas would stick to it; and, owing to friction between this layer and the surrounding air, currents would be produced which were symmetrical on all sides. But, if the ball is spinning also as it moves, things are different. Let the ball be advancing, as indicated in the cut, in the direction of the arrow and be spinning on an axis perpendicular to the plane of the paper counterclockwise. The *linear* velocity of a point of the ball on the side marked *A* is equal to the *sum* of *v*, the linear velocity of the ball, and *rh*, where *r* is the radius of the ball and *h* its angular velocity; while the *linear* velocity of a point of the ball on the side marked *B* is the *difference* between these. Consequently, the currents produced in the air by friction are much greater on the former side; and the relative velocity between the air and the ball is less; therefore the *pressure* in the air is greater. Owing to this fact the ball is pushed sidewise in the direction from *A* to *B*.

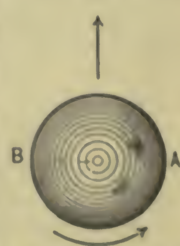


FIG. 81. — Motion of a twisting baseball.

Solid moving through a Fluid. — As a solid moves through the air, there are forces that oppose its motion; and many experiments have been performed to determine the connection between this force and the velocity of the moving body. Newton proposed the law that the force varies as the square of the velocity; but the accepted relation to-day is due to Duchemin: $R = av^2 + bv^3$, for speeds below 1400 feet per second, where R is the force, v the velocity, and a and b are constants. The forces acting on a solid moving through a fluid are the same as those it would experience if it were at rest and the fluid flowed past it in the opposite direction. If a board is placed obliquely across a current of a fluid, the pressure will be greater at the edge which is "up stream" than at the other, because the stream strikes it directly and then flows down along the board. There is thus a moment which turns the board directly across the current. Similarly, as a board or a piece of paper moves through a fluid, for instance the air, it turns so as to move broad face forward. This is illustrated by a sheet of paper or a leaf falling in air, by a flat shell falling through water, etc.

Liquids and Gases

As has been said several times, both liquids and gases are fluids; but liquids are distinguished by the fact that when contained in an open vessel they have a free upper surface in contact with the air, or if left to themselves they form drops inclosed in spherical surfaces, while gases completely occupy any space open to them; liquids are comparatively incompressible, while gases can be easily compressed. In discussing, then, the properties of a liquid as distinct from a gas, its surface of separation from other bodies and its incompressibility form the features to be studied; while the corresponding properties of a gas as distinct from a liquid are its power to expand so as to fill any space and its great compressibility.

CHAPTER VII

LIQUIDS

Compressibility of Liquids. — For many years it was thought that liquids were absolutely incompressible; but later it was shown that all liquids could have their volumes changed by the application of sufficiently great pressure. This is done by the piezometer, which has already been described. The liquid whose compression is to be studied is placed in a glass bulb provided with a fine-bored stem; this stem is open and contains an index which moves up or down as the volume of the liquid in the bulb is altered; the bulb is placed in the piezometer, which is filled with some transparent liquid like water. As the piston is screwed in, the pressure increases, and both the bulb and its contents are compressed. The pressure may be determined if the force acting on the piston and its area are known. Owing to the compression of the bulb itself the index would rise; but, if the contained liquid is more compressible than the bulb, the index is lowered, the amount of the fall depending upon the *difference* in the compressibility of the glass bulb and the liquid. If the volume of the bulb and of each unit length of its stem is known, this total decrease in volume may be determined; and, if the compressibility of the glass in the bulb is known, its decrease in volume may be calculated because the change in pressure is known; and by *adding* it to the decrease in volume observed, the true decrease in volume of the liquid is determined. The strain, then, corresponding to the measured *increase* in pressure Δp is this change in volume Δv divided by the original volume v. The coefficient of elasticity is then $\frac{\Delta p}{\frac{\Delta v}{v}}$ or $\frac{v\Delta p}{\Delta v}$. The "compressibility" of the liquid is the reciprocal of this, or $\frac{\Delta v}{v\Delta p}$. (For water at 8° C., the compressibility is about 0.000047, if the unit of pressure is "one atmosphere"; for ether at 10° C. it is not far from 0.000147; for ordinary alcohol at 14° C., 0.00010. At higher temperatures, the compressibility is much greater.)

Similarly, liquids can be stretched. If a tube which is closed at one end and which is *thoroughly cleaned* inside, is partially filled with a liquid and then carefully inverted, the liquid will not run out, but will stick to the end of the tube. It is therefore under a stretching force owing to its weight; and experiments show that under these conditions its length is increased. (This experiment must be done in such a manner as to avoid the action of the atmospheric pressure, or this must be taken into account in the calculations.)

Liquids at Rest

Form of Free Surface. — All liquids with which we deal are under the influence of gravity; and, therefore, if they are contained in large open vessels which they only partly fill and are at rest, their upper surfaces are horizontal, that is, are perpendicular to the force of gravity. For, if they were not, this force would have a component parallel to them, which would cause the liquids to flow. Similarly, if the horizontal surface is disturbed, for instance by dropping in a stone, the liquid which is thus forced below the horizontal level of the rest of the surface returns, and owing to its inertia continues to rise until it is above the level, thus mak-

FIG. 82. — Forces acting at any point of the surface of a liquid when it is not horizontal.

ing a "hump" or crest; this in turn sinks and makes a depression or "trough"; etc. These up-and-down motions affect the neighboring portions of the surface; and, as a result, what are called "waves" spread out from this centre of disturbance. Their energy comes from that of the original disturbance; and it is wasted away in friction between the moving parts of the liquid. These waves will be discussed in detail later, but it should be remembered that they owe their origin, not to any compression of the liquid, but to the fact that under the force of gravity the surface of a liquid is naturally horizontal.

If a liquid is contained in a vertical circular cylinder which is spinning rapidly around its axis of figure, the surface is no longer horizontal; in fact, it is a paraboloid of revolution; that is, any plane section through the axis is a parabola. The reason for this is clear. Consider any particle of the liquid in the surface. It is under the action of two forces, that of gravity acting vertically down and that due to the reaction of the rest of the liquid against which it presses and which acts perpendicularly outward from the surface. The surface must then have such a form that this force of reaction and the force of gravity have a resultant in a horizontal direction toward the axis; for the actual acceleration of the particle is in this direction since it is moving in a circle with constant speed. If this statement is expressed mathematically, it leads at once to the fact that the section of the surface is a parabola.

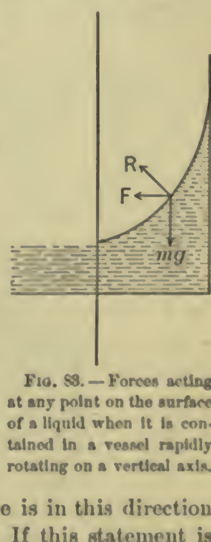


FIG. 88. — Forces acting at any point on the surface of a liquid when it is contained in a vessel rapidly rotating on a vertical axis.

Thrust; Centre of Pressure, etc. — If a liquid contained in a large open vessel is at rest, the pressure at any point depends upon two causes, atmospheric pressure and gravity. If the former is written P, and if the vertical depth of the point in question below the surface is h, the density of the liquid being d, the total pressure at the point is $P + dgh$. This pressure is entirely independent of the shape or size of the vessel, depending simply on the vertical distance between the point and the plane in which lies the upper surface. For, consider any point Q in a liquid at rest which is contained in a cylindrical vessel. The pressure is the same at all points in a horizontal plane through it, otherwise the liquid would flow, and the same is evidently true of a liquid contained in any other shaped vessel, *e.g.* a conical one. But the pressure at Q is $P+dgh$; and therefore all other points in the horizontal plane through it have the same pressure, wherever they are. (The pressure at a point under the slope of the side wall in the vessel shown in the cut owes its value directly to the weight of the column of liquid vertically above it and to the component downward of the reaction of the sloping side wall against the thrust of the liquid.)

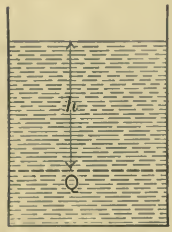


Fig. 84. — Pressure at a point whose distance below the free surface is h.

The pressure at a point in the liquid increases uniformly as the point is taken lower and lower; and therefore the mean pressure in the case when the vessel is a cylinder is the average of the pressure in the surface, P, and that at the bottom, $P + dgH$, where H is the vertical depth; that is, it is $P + \frac{1}{2} dgH$. The *force* on the bottom of the vessel, if it is horizontal and if its area is A, is $(P+dgH) A$; and the thrust on the side wall, if it is rectangular and vertical, is the average pressure multiplied by the area of the wall in contact with the liquid. If this area is A, the thrust is then $(P + \frac{1}{2} dgH) A$. Its point of application is found from the consideration that it is the resultant of a great number of parallel forces whose values increase uniformly from the surface down. In the case of a rectangular wall, this "centre of pressure," as it is called, is at a distance of one third the depth of the liquid

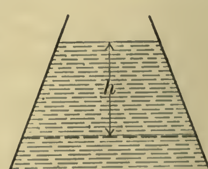


Fig. 85. — A vessel with oblique walls.

from the bottom. In the general case, when the wall is not rectangular, the total thrust is found by adding all the infinitesimal parallel forces which act on the elementary portions of the wall; and the centre of pressure is found by expressing the fact that the moments of these infinitesimal forces around any axis must equal the moment of the thrust when applied at the centre of pressure. These operations require, however, the use of the infinitesimal calculus.

If a liquid at rest is contained in a vessel that has several vertical branches of different shapes and sizes, its upper surface is at the same horizontal level in them all, provided they are not of such small bore as to cause capillary phenomena. This is evident from the fact that the pressure at all points in a horizontal plane *AB* through the body of the vessel is the same — otherwise the liquid would flow; and therefore the free surface must be at the same height above this level in all the branches. Similarly, the pressure is the same at all points of the liquid in the branches that lie in the same horizontal plane; *e.g.* in the plane *CD*.

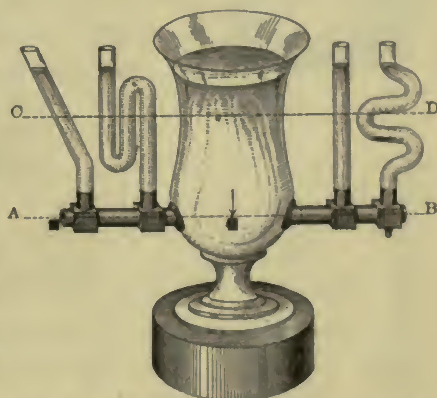


FIG. 86. — Same liquids in connecting tubes. *AB* and *CD* are horizontal planes.

Liquids in Connecting Tubes. — If two liquids that do not mix are placed in the same vessel, the denser will sink to the bottom, because by so doing the potential energy becomes less. A heavier liquid may, however, rest upon a lighter one provided there is no jarring; but the equilibrium is unstable.

If a vertical U tube contains two liquids that do not mix, the levels of the upper surfaces of the two liquids in the two branches are not the same. The heavier of the two liquids will occupy the bottom of the tube and rise to a certain height in one of the arms, while the lighter one will stand to a greater height in the other. The pressure at the surface of contact between the two liquids is the same as at a point in the heavier liquid at the same horizontal level, from what has been said in the previous paragraph; so, if the upper surface of the heavier liquid is at a vertical distance h_1 above this level, and that of the lighter at a distance h_2, $P + d_1gh_1 = P + d_2gh_2$, where d_1 and d_2 are the densities of the heavier and lighter liquids respectively. Therefore, $d_1h_1 = d_2h_2$; and $\frac{d_1}{d_2} = \frac{h_2}{h_1}$. This is, then, a method for the determination of the specific gravity of one liquid with reference to another; and, if the density of either is known, that of the other may be at once calculated.

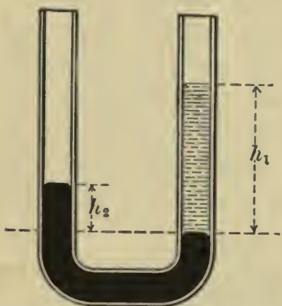


FIG. 87. — Two liquids of different density which do not mix.

Barometer. — The pressure of the atmosphere is, as a rule, measured by balancing it against a column of mercury. The apparatus consists of a long, wide tube, which is closed at one end and which contains a column of mercury, but no air or other gas (except mercury vapor). The tube is placed in a vertical position; and either its open end dips into a basin of mercury or the tube is bent into the shape of the letter J. The space above the mercury may be entirely freed from gases by different means. (One is to hold the tube, closed end down, *fill* it with mercury, cover the open end with the finger, invert it carefully, and place it upright in a basin containing mercury with the open end under the surface.) The pressure on the surface open to the air holds the mercury in the tube and is the same as that at a point at its level in the mercury in the tube; so, if the surface of the mercury in the tube stands at a height h above this outer open surface, the pressure due to the atmosphere equals dgh, where d is the density of the mercury at the temperature of the air, because there is no pressure on the top of the column, and the pressure at any point is that due to gravity alone. (h is observed to be about 76 cm. when the pressure is measured at sea level; and so the tube must have at least this length.) It should be noted that this height is independent of the shape or cross section of the tube, provided only that it is so wide as to avoid capillary action. See page 188.

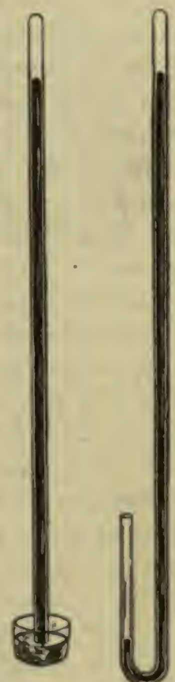


FIG. 88. — Mercury barometers. Cistern and siphon forms.

If a liquid other than mercury were used, it would stand at a height as much greater than this as the density of mercury is greater than its own. (Thus, since the density of mercury is about 13.6 times that of water, this liquid would be forced up in a barometer by ordinary atmospheric pressure to a height 13.6×76 cm.) The pressure of the air as given by a barometer is dgh dynes per square centimetre on the C. G. S. system. It is, however, ordinarily expressed in "centimetres of mercury," that is, in terms of a unit pressure equal to that due to a vertical column of mercury at 0° C., 1 cm. in height, when g has the value it possesses at sea level and 45° latitude. (See page 166.) Calling this value of the density of mercury d_0, and that of g, g_{45}, this unit pressure equals $d_0 g_{45}$ dynes per square centimetre. So, if the observed height of the barometer is

h cm., when the temperature is $t°$ C. and the latitude is l, we may write the value of the pressure in dynes per square centimetre, $P = dgh$; and its value in "centimetres of mercury," therefore, is

$$P = \frac{dgh}{d_0 g_{45}}.$$

This is a height; and so we speak of a barometric "pressure of 76 cm.," meaning "76 cm. of mercury." The relation between g and g_{45} is known (see page 131); and that between d and d_0 is also known. For, as will be shown presently, if t is the temperature on the Centigrade scale, $d_0 = d(1 + 0.0001818\,t)$. Therefore, if the temperature and latitude at the place of observation are known, the pressure in terms of centimetres of mercury may be at once calculated from the observed barometric height. In the formula, h is the height in centimetres. Sometimes, however, the reading is made on a divided scale which is correct at 0° C.; and in this case the readings must be corrected in order to give h. If the scale is made of such a material that each centimetre increases in length an amount a cm. for each degree rise in temperature, two divisions which are 1 cm. apart at 0° are a distance $(1 + at)$ cm. apart at $t°$ C. Consequently, if the observed reading is H scale divisions, the height is $H(1 + at)$ cm.; and therefore $h = H(1 + at)$.

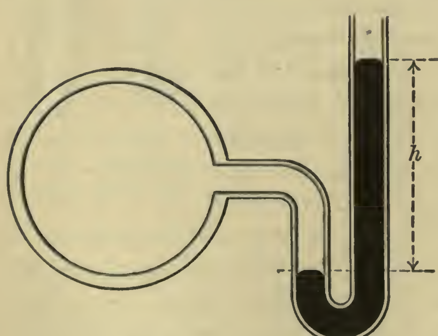


FIG. 89.—Open manometer.

Open Manometer. — In a similar manner, the pressure in a gas inclosed in any vessel can be measured. Let a bent tube containing some liquid be joined to the vessel as shown. If there is a difference in pressure between the gas inside the vessel and the air outside, there will be a difference in level of the columns of the liquid in the two arms. Call this difference h, and the density of the liquid d. Then, if the level is lower in the arm attached to the vessel, the pressure in the gas inside is $P + dgh$. (If the level is higher in this arm, the pressure is $P - dgh$.) This instrument is called an "open manometer." The pressure, as here expressed, is in dynes per square centimetre, if the C. G. S. system is used; its value in centimetres of mercury can be deduced, as has just been explained for a barometer.

Floating Bodies. — There is one application of Archimedes' principle to liquids that is of special interest. It is to the case of a body floating on the surface of a liquid. If a solid of less density than a liquid is immersed in it and allowed to move, it will rise to the surface, but will come to a position of equilibrium when, as it floats, it displaces a volume of the liquid whose weight equals its own; for, under these conditions, the upward buoyant force due to the liquid equals the downward weight of the solid. The line of action of the former is vertically through the centre of gravity of the displaced liquid; that of the latter, vertically through the centre of gravity of the floating body. Therefore, when there is equilibrium, these two centres of gravity must lie in the same vertical line; otherwise there would be a moment which would make the body turn around a horizontal axis.

This equilibrium is stable if, when the body is tipped slightly, the resulting moment is in such a direction as to turn it back again; it is unstable if this moment, under similar conditions, is such as to tip it over. Thus, a board floating on its side is in stable equilibrium; but, if made to float upright, its equilibrium is unstable.

Osmosis and Osmotic Pressure. — As one substance dissolves in another, it breaks up into small particles which diffuse through the solvent. These particles by their presence affect its molecular forces, as is shown by many facts. One of these may be mentioned here. It is found that certain solid bodies allow some liquids to pass through them, but not others (see page 141); and it is possible to make a membrane that will permit the molecules of a liquid to pass through perfectly free, but will not permit the passage of any dissolved molecules. Such a membrane is called "semi-permeable." If now a solution is placed in a wide tube closed at one end with such a membrane, and is supported upright in a large vessel containing the pure liquid, which can pass through the membrane, it is observed that the levels of the liquid in the tube and in the outer vessel are not the same, as they would be if the membrane were absent; the height of the level in the tube is the greater. There is therefore on the two sides of the membrane a difference in hydrostatic pressure which is maintained by some force due to the difference in the conditions on these two sides. The molecules of the solvent can pass freely through the membrane, and they continue to do so until, when equilibrium is reached, the hydrostatic pressure prevents any further passage. (Of course molecules may still continue to pass through; but, if they do, an equal number pass out.) There is therefore a difference between the pure solvent and the solution. If the density of the solvent is d, that of the solution does not differ much from this; and if the difference in level of the two free surfaces is h, the hydrostatic pressure dgh measures this tendency of the pure solvent to pass through the membrane into the solution, that is, it measures the effect the dissolved substance has upon the solvent in affecting its molecular forces. This passage of a liquid through a membrane or

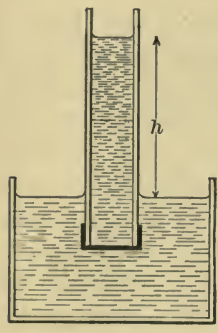


FIG. 90.—Osmotic pressure: inner tube contains solution; outer vessel, pure solvent.

porous partition is called "osmosis," as has been already stated; and the above pressure, dgh, is said to measure the "osmotic pressure" of the solution. Experiments show that, as the solution is made more and more concentrated, the osmotic pressure increases. If m grams of a substance are dissolved in m_1 grams of a solvent, the ratio $\frac{m}{m_1}$ is called the "concentration"; and in certain simple cases the osmotic pressure varies directly as this, while in others it varies more rapidly. This law is the same as that for a gas; viz., the pressure varies directly as the density of the gas, *i.e.* its concentration; but if the gas changes its character by its molecules dissociating into parts, then the pressure varies more rapidly than the density. (See page 200.) We are thus led to believe that this abnormal osmotic pressure is due to a dissociation of the dissolved molecules into simpler parts.

Liquids in Motion

Efflux of Liquids. — If a liquid is contained in a large vessel which has thin walls, and if a small opening is made in either the bottom or side, the liquid will escape. This motion is called "efflux" or "effusion." The velocity of escape may be at once calculated, because, since the wall is assumed to be thin, there is no friction, and since the opening is small, we may neglect any motion of the liquid except that of the escaping stream. Thus the phenomenon is the same as if a drop of the liquid disappeared off the surface and reappeared lower down with a certain speed. If the opening is at a depth h below the surface, and if the speed of efflux is s, each drop of mass m loses an amount of potential energy mgh and gains an amount of kinetic energy $\frac{1}{2} ms^2$. Therefore these two quantities are equal, or $s = \sqrt{2gh}$. This is, of course, the formula for a particle falling freely toward the earth; and therefore, if the jet were turned upward, it would rise to the height of the level of the liquid in the vessel, were it not for the opposing action of the air. (This formula was first deduced by Torricelli.) The pressure in the liquid at the opening is $P + dgh$, while that on the outside is P; so the difference in pressure causing the flow is dgh. Calling this p, the speed of efflux may be expressed in terms of it, viz., $s^2 = 2\frac{p}{d}$, or $s = \sqrt{\frac{2p}{d}}$. The direction of the jet depends, of course, upon the position of the opening; and, unless this is on the bottom, the path of the jet is a parabola.

Other cases of motions of liquids will be discussed in Chapter IX.

Capillarity and Surface Tension

Fundamental Principle. — If a liquid is left to itself, free from external forces, it assumes the shape of a sphere; and this is approximately the condition with falling drops of rain or of molten metal (like shot) and with soap bubbles. It is rigorously so if a small quantity of a liquid is immersed in another liquid of the same density with which it does not mix — Archimedes' principle. Of all solid geometrical figures having the same volume the sphere has the least surface; so this fundamental property of a liquid surface is that it becomes as small as it can. Thus the surface of a drop contracts until the resulting pressure in the liquid balances the contracting force; it requires a force to blow a soap bubble, and, if one is left attached to the pipe and the lips are removed, it will contract. Again, if a glass plate is dipped in water (or any solid is dipped in any liquid that wets it), and is then raised slightly, the surface of the liquid near the plate is curved with the concavity upward. It has contracted from a rectangular shape, in doing which some of the liquid is raised above the horizontal surface; and the liquid comes to rest when the weight of this elevated portion balances the contracting force of the surface. Similarly, if a glass plate is dipped in mercury (or any solid is dipped in any liquid that does not wet it), the surface of the liquid near the plate is curved so as to be convex. Since the liquid does not wet the plate, its surface continues around below the plate; and, as it contracts, it rounds off the corners, thus leaving a free space which the force of gravity would cause the liquid to fill were it not for the contracting force of the surface. There is equilibrium, then, when these two forces balance each other. This phenomenon in the neighborhood of a solid dipping in a liquid is most marked when the former is a tube with a small bore. If such a tube is dipped in a liquid that wets it and is then raised slightly so as to leave a liquid surface on the inner and outer walls, the whole liquid surface includes that on the walls and that of the liquid in which the tube dips. So considering the liquid surface inside the tube, it has the appearance of the inside of the finger of a glove. The liquid is then raised in the tube, owing to the contraction of the surface; and equilibrium is reached when this contracting force is balanced by the effect of gravity on the raised portion of the liquid. Similarly, when a glass tube is dipped in mercury, the surface in the tube is depressed.

FIG. 91. — Capillary action when a solid plate is dipped in a liquid: (1) when the liquid wets the solid; (2) when the liquid does not.

Surface Tension. — There is, therefore, a force produced by a liquid surface; and the simplest manner of defining it is to consider the force acting across a line of unit length in the surface. This is a molecular force and is evidently due to the fact that a molecule in the surface of a liquid is in a different condition from one in the interior. For a surface of a given liquid in contact with a definite medium, then, no matter whether its area is large or small, this force is a constant quantity; and unless it is stated otherwise, the surrounding medium is always understood to be ordinary air. The force acting across a line of unit length of the surface of a given liquid in contact with a definite medium is called its surface tension with reference to the medium, and has the symbol T. A simple direct experiment showing the amount of this tension is to construct a rectangular frame of wire, one side of which is movable, and to make a film of liquid fill the area. (This may be done by dipping the frame for a moment into soapy water and then removing it.) It will be found that a force must be exerted on the movable wire to keep the film from contracting. Calling the length of the movable wire l, this force equals $2\,Tl$, because the film has *two* surfaces. If under the action of the force the wire is moved so as to make the film larger, work is done. If the distance the wire moves is called x, this work is $2\,Tlx$; but the increase in area on the *two* sides is $2\,lx$, and therefore the work done per unit increase of area is T. In other words, to increase the surface of a liquid by one unit of area requires an amount of work equal to T; or, the potential energy of a surface of area A is numerically equal to TA. As the surface of a liquid is increased, it is not stretched, as a rubber bag or toy balloon is, but *new* surface is formed by molecules coming up into the surface from the interior; the new and the old surfaces being identically alike. (If in the experiment just described the movable wire be drawn out too far, or if a soap bubble be blown up too large, a point is reached when the two surfaces of the film come so close together that, if their area is

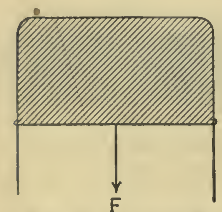


FIG. 92. — A soap-film stretched on a wire frame, one side of which is movable.

further increased, the interior molecules are no longer in the condition in which they are when the film is thicker; and all the properties of the film are changed.)

Connection between Pressure and Surface Tension. — Let us now examine more closely the illustrations of the contracting force of a liquid surface that were given above. A spherical drop may be considered as made up of two halves touching at an equatorial section; they are held together by the tension *in the surface*, acting *across* the equator; and there is a reaction, as shown by the pressure in the liquid, acting over the equatorial section in which the two halves touch. If r is the radius of the sphere, the length of the equator is $2\pi r$, and the force of contraction due to the surface tension across it is therefore $T 2\pi r$; the *area* of the equatorial section is πr^2, and, if p is the pressure in the drop, the force of reaction over this section is $p\pi r^2$. Since the drop is supposed to have ceased to contract, these two forces must be equal, or $p\pi r^2 = T 2\pi r$; *i.e.* $p = \frac{2T}{r}$. This pressure is felt, of course, throughout the drop, and is in addition to the pressures due to the atmosphere or to gravity. The formula states that, if a liquid whose surface tension is T has a portion of its surface (or all of it) curved in the form of a sphere of radius r, there is a pressure $\frac{2T}{r}$ toward the centre of curvature, and so, if the surface is at rest, there must be an external pressure $\frac{2T}{r}$ acting on it in order to hold it in equilibrium.

Thus, if a soap bubble is blown out, the pressure required for the gas inside, in excess of that outside, is $\frac{2T}{r_1} + \frac{2T}{r_2}$ where r_1 and r_2 are the radii of the interior and exterior surfaces. Since the film is very thin, the radii may be considered equal, and the pressure of the air inside must be greater than that out by an amount $\frac{4T}{r_1}$. If by means of a pipe with two openings two bubbles of radii R_1 and R_2 are blown simultaneously and are then left connected, the pressure $\frac{4T}{R_1}$ will be required in one in order to maintain equilibrium, and $\frac{4T}{R_2}$ in the other. So there will not be equilibrium. If $R_1 > R_2$, $\frac{4T}{R_1} < \frac{4T}{R_2}$; therefore the bubble with the larger radius requires the less pressure, and it will become larger while the other contracts.

It is seen that the smaller the drop or the bubble the greater is the pressure that is required to maintain it in equilibrium. In the limit, as the radius r approaches zero, the corresponding pressure p becomes infinitely great; it is therefore impossible to form a drop or a bubble with an infinitely small radius. Thus, drops of rain or dew always form around nuclei or points; and bubbles, like those seen in a boiling liquid, start from points or small bubbles of dissolved gases. (If the nucleus is electrically charged, a drop can be formed with a diameter smaller than would be possible with one that is uncharged.)

Capillary Phenomena. — Having deduced this value for the pressure corresponding to a certain radius, we can at once find the connection between the height to which a liquid is raised in a tube and the dimensions of the tube. Let us consider the case of a tube that is wet by the liquid, and let its cross section at the top of the column of the liquid be a circle of radius r; then the shape of this upper surface of the liquid is that of a hemisphere of radius r. The pressure in the liquid at a point inside the tube on a level with the horizontal surface outside is the same as that at

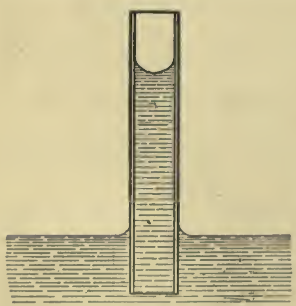


Fig. 93. — Liquid in a capillary tube which it wets, *e.g.* water in glass. (Greatly magnified.)

a point in this surface, viz., P, the atmospheric pressure; but it is due to three independent causes: 1, the atmospheric pressure P on the top of the column; 2, the vertical height of the column (this pressure amounts to dgh); 3, the contracting force of the curved surface (this is equivalent to a pressure $\frac{2T}{r}$, and *opposes* the other two). Thus the total pressure is $P + dgh - \frac{2T}{r}$. This, as has just been said, must be equal to P; and therefore $dgh = \frac{2T}{r}$ or $h = \frac{2T}{dgr}$.

(If h is the distance from the bottom of the curved surface to the free horizontal surface, the above formula is not correct, since allowance must be made for the portion of liquid above the bottom of the curved surface. It may be shown without difficulty that the correct formula is $h + \frac{1}{3}r = \frac{2T}{dgr}$; but, since as a rule r is small in comparison with h, the approximate formula is generally used.) It is thus seen that the smaller the radius of the tube the greater is h, or the higher does the liquid rise. (Phenomena dealing with surface tension are usually called "capillary," because the bore of tubes which show marked effects as just described is comparable with the size of a hair, the Latin name for which is *capillus*.) Illustrations of this action are given by a lump of sugar when dipped in water, a blotting paper absorbing a drop of ink, etc.

It should be noted particularly that, in the above discussion and formula, no reference is made to the cross section of the tube except at the point where the top of the column of liquid comes; so the tube elsewhere can have any size. The liquid will not rise in the tube *of itself* unless the bore is small throughout and the inner wall is wet with the liquid. But if the liquid is sucked up in the tube and then allowed to fall, it will come to rest at the height given by the formula.

In a perfectly similar manner it may be shown that the depression of the surface of mercury in a glass tube (or of any liquid in a tube which it does not wet) is given by the same formula.

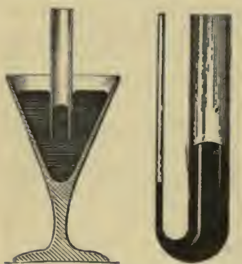


FIG. 94. — Capillary action between mercury and glass.

This formula can be used to measure the surface tension of a liquid; for h, d, g, and r can all be measured with accuracy. There are, however, other methods which in some respects are more satisfactory. The values of the surface tension of a few liquids in contact with air, in dynes per centimetre, at about 30° C., are given in the following table:

TABLE

Water	72.8	Olive oil	34.6
Mercury	513.	Turpentine	28.8
Alcohol	22.	Petroleum	29.7

Another mode of considering the curvature of a liquid surface near a solid wall is as follows: imagine the liquid to have a horizontal surface clear up to the wall; a particle of the liquid surface near the wall will be under the action of three forces (apart from gravity), viz.: 1, one owing to the molecular forces of the rest of the liquid; it is represented in the cut by F; 2, one owing to the molecular forces between the liquid and the solid; provided the latter is wet by the former, it is represented by F_1; 3, one owing to the forces between the upper medium — the air, in general — and the particle of liquid; this is neglected here. The resultant of these is represented by R; and therefore the surface of the liquid must be so curved as to be at right angles to its direction.

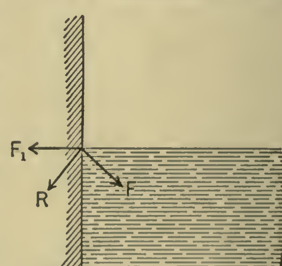


FIG. 95. — Forces acting on a particle of a liquid in its surface at a point near a solid wall.

If the surface of the liquid is not spherical, the formula for the pressure may be shown to be $p = T\left(\frac{1}{R_1} + \frac{1}{R_2}\right)$, where R_1 and R_2 are the so-called principal radii of curvature. Thus, if the surface is a cylinder of radius r, $R_1 = 0$ and $R = r$; so $p = \frac{T}{r}$. This formula may be proved directly by considering a cylindrical drop instead of a spherical one as was done on page 185. It may be shown further that, if a cylindrical portion of liquid (as, for instance, a jet of water issuing under considerable pressure) has a small radius, it breaks up into spherical drops. The exact statement is that if l is the length of the cylinder and r its radius, the surface is unstable when $l > 2\pi r$.

Effect of Surface Impurities. — If a liquid is homogeneous, the surface tension of all points of its surface must be the same; and therefore it is impossible to blow a bubble of such a liquid or to stretch a film of it on wire frames; for it is evident that, if a vertical film exists, the tension in the upper portions of the surface must be greater than in the lower, because the former have to support the *weight* of the latter, and such inequalities in tension cannot exist in a homogeneous surface. If bubbles or films are to be formed, then, the liquid surface must be made heterogeneous, *e.g.* soapy water may be used. In these films the upper portions are observed to be purer than the lower ones, showing that a pure film has a greater surface tension than a contaminated or heterogeneous one.

If minute portions of camphor are dropped upon a *clean* surface of water, they are seen to dart backward and forward with a most violent motion; this, however, is stopped by an extremely small trace of oil on the surface. A piece of camphor is never exactly symmetrical and therefore dissolves in the water more rapidly at one point than at another, and pollutes the surface more rapidly there. This makes the surface tension less at this point than elsewhere; and so the camphor is drawn away by the contraction of the purer surface on the opposite side. When a drop of oil is put on water, it spreads over the surface under the action of the surface tensions of the various surfaces. There are three of these forces acting on any portion of the edge of the drop: the tension T_3 between the water and the air, T_2

FIG. 96. — The surface tensions acting along an edge of a drop of oil when first placed on the surface of water.

between the oil and the air, T_1 between the oil and the water. The oil spreads because T_3 is greater than the resultant of T_2 and T_1. This thin layer of oil will prevent the dissolving of the camphor, and its motions will cease.

If a drop of alcohol is poured on a glass plate that is covered with a thin layer of water, the tension of the surface of the solution of alcohol is so much less than that of the pure water that the latter surface contracts, tearing apart the former and leaving the glass quite dry.

The surface tension of a liquid varies with the temperature, decreasing as the latter rises. This may be shown by many obvious experiments.

Ripples; Effect of Oil upon Waves. — It requires work to increase the area of a liquid surface, and so if the surface is increased slightly by some disturbance, there is a force of restitution, and waves will be propagated over the surface, which are quite distinct from those due to gravity. These are due to surface tension, and the crests come so close together that they are called "ripples." These may be seen if a fine wire in a vertical position and dipping in a liquid is moved sidewise. They will be discussed later.

As the wind blows over a liquid surface it will soon magnify ripples into regular gravitational waves, and it is evident that the less surface there is exposed to this action of the wind so much the less is its effect. If a thin layer of oil is spread over a liquid surface, the wind blowing over it will tend to gather the oil in the same regions where it would heap up the water; this excess of oil over one region produces a scarcity of it over others, and the surface tension in the latter is therefore greater than in the former, and so it opposes the action of the wind in causing the water to form waves.

CHAPTER VIII

GASES

General Properties

Dalton's Law. — The fundamental properties of a gas as distinct from a liquid are explained by assuming that its molecules are so far apart and have such freedom of motion that it distributes itself uniformly throughout a vessel of any shape and may be compressed with ease into a much smaller one. The actual space occupied by the particles of the gas must be extremely small, because if two different gases are inclosed in the same vessel, they mix uniformly, and each behaves almost perfectly as if the other were absent; the pressure at any point (or on the walls) is the sum of the pressures which each gas by itself would produce. This is known as Dalton's Law of Gases, having been discovered by him in 1802; but careful experiments show that it is not absolutely exact.

Effusion. — If any gas is inclosed in a vessel that is surrounded by a different gas, and is allowed to escape into the latter, it is for all practical purposes as if it were escaping into a vacuum, for each gas acts independently of the other, and the only difference enters from the fact that a longer time is required for a gas to diffuse into another gas than for it to distribute itself in an empty space. If the opening in the vessel from which the gas is escaping is small, and if its walls are thin, the same formula applies as for the efflux of a liquid, viz., $s=\sqrt{\frac{2p}{d}}$, where s is the speed of the escaping gas, d is its density, and p is its pressure. (p is really the difference in pressure of *the escaping gas*, inside and out; but if the outside space is very large compared with the vessel, the pressure outside may be neglected.) While the one gas is escaping, the surrounding gas is entering, and the two processes go on independently. This relation between the velocity of escape of a gas and its density suggested to Bunsen a method for determining the relative densities of two gases, because the quantity of gas that escapes in a given time can be measured with ease. If the openings through which the gas escapes are extremely fine, a different law is obeyed, which was investigated by Graham.

Sensitive Flame. — If a gas escapes through a tube with a small opening under considerable pressure, it forms a "jet," which preserves its identity for some distance without diffusing into the surrounding gas. The jet is inclosed, as it were, in an envelope which prevents the two gases from mingling. This envelope is made up of particles of the escaping gas which are given a rotation owing to their rapid motion over the particles of the surrounding gas. If this envelope can be broken into, the two gases mix and the jet loses its character. If a jet of illuminating gas is formed in this way and is lighted, it forms a tall, narrow flame nearly cylindrical; but if a whistle is blown, or a bunch of keys rattled near by, the jet breaks down into an ordinary fan-shaped gas flame. This is owing to the disturbances sent out by the sounding body in the form of short waves or pulses in the air, which disrupt the gaseous envelope of the flame. For this reason a jet like this is called a "sensitive flame." It is used to study the many wave phenomena in connection with Sound.

Compressibility of Gases. — If a gas is inclosed in a cylinder provided with a movable piston, this may be forced in or out, and the corresponding values of the pressure and volume may be measured. Or, if a tube is bent in the form of a letter **J**, and its lower end sealed, it may be placed vertically, and some liquid like mercury may be poured in at the open end, thus trapping a quantity of the air (or other gas) in the shorter arm. As more and more liquid is poured in, the volume and pressure of the gas vary, and both may be measured. (If h is the vertical distance between the two surfaces of the liquid, d its density, and P the atmospheric pressure, the pressure of the gas is $P + dgh$.)

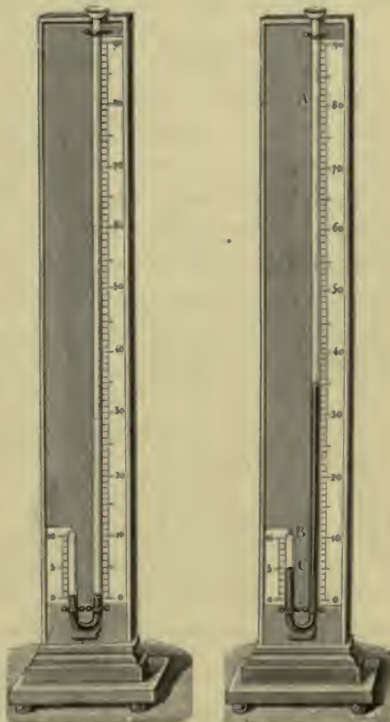


FIG. 97. — Apparatus for observing Boyle's Law: a small quantity of air is trapped in the closed end of the tube.

The coefficient of elasticity of a gas is, as explained before, $\frac{v\Delta p}{\Delta v}$, where Δv is the change in the volume v produced by a change in pressure equal to Δp. Experiments prove that this coefficient depends greatly upon how rapidly the compression (or expansion) is produced. The reason is that, if the volume of a gas is made smaller suddenly, its temperature is raised, and this of itself produces an increase in pressure quite independent of that which would accompany the same decrease in volume if brought about slowly. Two extreme cases are ordinarily considered: one, when the decrease in volume is made so slowly that there is no change in temperature; the other, when the change is made as rapidly as possible, so that there is no time for the temperature effect to become weakened by diffusion. The former is called an "isothermal" change; the latter, an "adiabatic" one.

Boyle's Law and its Consequences

Boyle's Law. — The first philosopher to study experimentally the exact properties of gases was Robert Boyle, who, in 1660, carried out a most careful series of experiments on the variation in the volume of a gas as its pressure is changed. He discovered that, to a high degree of approximation, the pressure and volume of a given quantity of a gas are connected by the following relation: the product of the values of the pressure and volume remains constant during all changes, *provided the temperature is unchanged.* In symbols, that is, $pv =$ constant, if the temperature is constant, where p is the pressure and v the corresponding volume of a given quantity of the gas. This means that if the volume is decreased to one half its value, the pressure is doubled, etc. This is known as "Boyle's Law for a Gas." Naturally, if there is twice the mass of the gas at the same pressure, its volume is twice as great, and writing m for the mass of the gas and k as a constant factor of proportionality, Boyle's law becomes

$$pv = km,$$

or, writing as usual, d for the density,

$$p = kd.$$

k is, then, a constant for a given kind of gas at a definite temperature; if either the gas or the temperature is changed, k takes a different value.

As stated above, Boyle's law is not exact except for small variations in pressure. If the pressure is increased greatly, the product pv, instead of remaining constant, increases also. This fact was recognized by Boyle himself and has been confirmed by more recent investigators, notably Regnault and Amagat.

Dalton's law may be expressed quite simply in terms of Boyle's law. If several gases are inclosed separately in different vessels which have the same volume, but are at the same temperature, let their pressures be p_1, p_2, etc.; then, if these gases are all put in one of the vessels, the pressure will be

$$p = p_1 + p_2 +, \text{etc.},$$
$$= k_1 d_1 + k_2 d_2 +, \text{etc.}$$

Coefficients of Elasticity. — The value of the isothermal coefficient of elasticity may also be expressed in a simple form, because Boyle's law applies to the changes. Let p and v be the initial pressure and volume of a gas; then, if the temperature is kept constant and the pressure is increased to $p + \Delta p$, the corresponding volume $v - \Delta v$ must be such that

$$pv = (p + \Delta p)\ (v - \Delta v)$$
$$= pv + v\Delta p - p\Delta v - \Delta p \Delta v.$$

Therefore, $$v\Delta p - p\Delta v - \Delta p\Delta v = 0.$$

But if the changes are extremely small, the last product may be neglected, and hence

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

or $$v\frac{\Delta p}{\Delta v} = p.$$

But Δp is the change in pressure corresponding to the change in volume Δv, and the coefficient of elasticity therefore equals p. So the isothermal coefficient of elasticity of a gas when under a pressure p numerically equals p.

It will be shown later that the adiabatic coefficient of elasticity for any gas is the product of the pressure by a constant whose value is about 1.41 for the gases hydrogen, oxygen, and nitrogen, but is different for others.

Isothermal of a Gas. — Boyle's law may be expressed graphically by using axes of pressure and volume as has been done in previous articles. Choose arbitrarily some point P in the plane of the two axes and draw through it a line $\overline{PQ}$ parallel to the axis of pressure; the corresponding pressure is $\overline{QP}$, and the volume is $\overline{OQ}$. To represent Boyle's law, a series of points must be chosen such that the products of the value of their pressures and volumes equal the product of $\overline{QP}$ and $\overline{OQ}$. These points make up a curve which is called an "equilateral hyperbola," and it is shown in the cut.

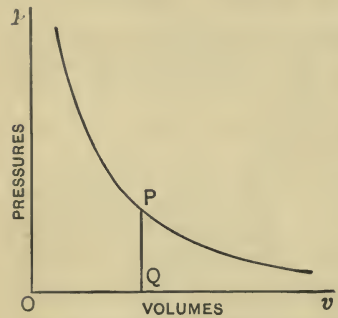


FIG. 98. — The curve is an equilateral hyperbola, the equation for which is pv = constant.

The formula pv = constant is known as the "equation of condition" for an ideal gas at constant temperature, or as the equation of an "isothermal" for a gas; and, as has been said, it is only approximately true for an actual gas. Other formulæ have been proposed which apply more exactly to ordinary gases over wider ranges of pressure. The most satisfactory of these is due to van der Waals, and has the form

$$\left(p + \frac{a}{v^2}\right)(v - b) = \text{constant.}$$

In this p and v have their usual meaning, and a and b are constant quantities for any one gas. This equation agrees fairly well with experimental results, when a gas is compressed from its ordinary condition until it is a liquid, as is explained in more advanced textbooks. (See Edser, HEAT.)

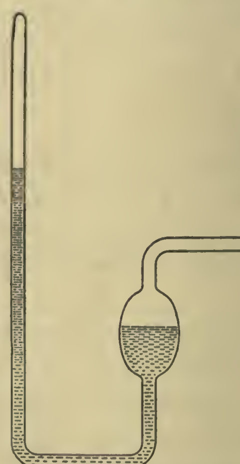


FIG. 99. — Closed manometer.

Closed Manometer. — A convenient method for the measurement of high pressures is afforded by Boyle's law. Some gas, such as air, is trapped in a closed tube by means of mercury; its volume is measured under atmospheric pressure; the pressure to be measured is then applied to the mercury, thus compressing the confined gas, and the resulting volume is measured. The ratio of these two volumes equals the reciprocal of the pressure expressed in "atmospheres." This instrument is called a "closed manometer." (Such an instrument is used often in connection with a piezometer in order to measure the pressure.)

Kinetic Theory of Gases

Fundamental Phenomena; Temperature. — The pressure that a gas exerts on the walls of the vessel containing it is at once explained if we assume that it consists of a great number of minute particles which are in rapid motion. As any one particle strikes the wall, it has its momentum perpendicular to the wall reversed, and therefore it exerts an impulse on it. The total force on the wall is the change in momentum produced in a unit of time; and the pressure is the force per unit area.

If one portion of the wall is movable, we can imagine it yielding to these impulses, provided the external force is not sufficient to withstand the bombardment; but as it yields, it is evident that the linear velocity of any particle rebounding at that instant is less than if the wall did not move; and so the kinetic energy of translation of the particles of the gas is decreased while work is done by the gas in overcoming the external force. Similarly, if the movable portion of the wall is forced in, work is done on the gas; and it is evident that the linear velocity of a particle rebounding at that instant is increased, and so the kinetic energy of translation of the particles of the gas is increased. Experiments on actual gases show that if one is allowed to expand against external forces, its *temperature* decreases, while if it is compressed, its temperature increases. Thus it is seen that the temperature of a gas varies directly as the kinetic energy of translation of its particles.

Pressure. — We know nothing about the actual size or shape of a molecule; but we can prove that, if we had inclosed in a vessel with rigid walls a great number of small, perfectly elastic spheres, moving at random but with great speed, this collection of particles would have many properties similar to those of a gas. For ease of calculation, let us assume that the vessel is a rectangular one, having edges of length a, b, and c. At any instant a definite particle has a certain velocity, but owing to impacts with other particles and with the walls, this changes frequently, and this is true of all the particles; so, apparently, there is no regularity. But if things are in a steady state, there is a certain unvarying proportion of the particles — not the same individual ones, however — that have a given component velocity parallel to any one edge of the rectangle. Let N be the number of particles *per unit volume* that have the component velocity v parallel to the edge whose length is a; then the total number in the vessel that have this component velocity is $Nabc$. If each of these particles has the mass m, the momentum of each parallel to the edge referred to is mv; and therefore as each strikes the wall at the end, and its velocity is changed from v to $-v$, its momentum is changed from mv to $-mv$, or by an amount $2\,mv$. The time taken for a particle with the velocity v to pass from this wall across to the other end and return is $\frac{2\,a}{v}$. (This is the time taken for the *effect* of the particle to be again felt at the wall, if, instead of moving over the whole distance and back again, it impinges on another particle and so hands on its momentum.) Therefore the number of impacts each particle makes on the wall in a unit of time is $\frac{v}{2\,a}$. Since at each impact of each particle there is a change in momentum

$2\,mv$, the change in momentum at the wall owing to one particle during a unit of time is $\frac{v}{2\,a}\cdot 2\,mv$ or $\frac{mv^2}{a}$; and, as there are $Nabc$ particles having this component velocity v, the change in momentum in a unit of time, owing to the particles whose component velocity is v, is $Nabc\cdot\frac{mv^2}{a}$, or Nmv^2bc. This is, then, the *force* on the end wall owing to the above particles; and since the area of this wall is bc, the *pressure* is Nmv^2. The total pressure is found by adding up similar terms for the particles having other velocities. The number of particles per unit volume, however, which have the component velocity $-v$ must also equal N on the principle of probabilities, and so the number which have the same value v^2 is $2\,N$. Therefore, writing v_1, v_2, etc., for different velocities and N_1, N_2, etc., for corresponding numbers per unit volume, the total pressure is

$$(2\,N_1v_1^2 + 2\,N_2v_2^2 + \text{etc.})\,m.$$

Further, the mean value of v^2 is

$$\frac{2\,N_1v_1^2 + 2\,N_2v_2^2 + \text{etc.}}{2\,N_1 + 2\,N_2}.$$

(See page 30.) But $2\,N_1 + 2\,N_2 +$ etc. is the *total* number of particles in a unit volume; call its value N and write for the mean value of v^2, $\overline{v^2}$. Then the total pressure equals $mN\overline{v^2}$. The total velocity V of any particle expressed in terms of its components parallel to three directions at right angles to each other, u, v, w, is given by $V^2 = u^2 + v^2 + w^2$. But, since the motion of the particles is a perfectly random one, the mean values of u^2, v^2, and w^2 must be equal; so writing $\overline{V^2}$ for the mean value of V^2, $\overline{V^2} = 3\,\overline{v^2}$; and the final form for the pressure owing to the particles is $p = \frac{1}{3}\,mN\overline{V^2}$. Since N is the total number of particles in a unit volume, and since each particle has a mass m, the *density*, d, or the mass in a unit volume, is mN; hence the pressure may be written

$$p = \frac{1}{3} d \overline{V^2}.$$

This states that, if the mean kinetic energy of translation of the particles does not change, the pressure varies directly as the density. This is Boyle's law, assuming that the temperature of a gas corresponds to the mean kinetic energy of translation of its particles.

If this formula can be applied to an actual gas, the mean squared velocity of its molecules may be at once calculated, because $\overline{V^2} = \frac{3p}{d}$, and the density of a gas at a certain pressure may be determined by experiment. The density of a gas varies with its temperature as well as with the pressure, and so does therefore $\overline{V^2}$. At the temperature of melting ice, V for hydrogen is calculated to be 1843 metres per second; and for carbonic acid gas, 392 metres per second. At the temperature of boiling water, each of these is $1\frac{1}{4}$ greater.

Avogadro's Hypothesis. — Referring to the previous formula for the pressure, viz., $p = \frac{1}{3} mN\overline{V^2}$, it is seen that, if there are several sets of particles inclosed in the same space, and if we can assume that they act independently of one another, the total pressure is $p = \frac{1}{3}(m_1N_1\overline{V_1^2} + m_2N_2\overline{V_2^2} + \text{etc.})$. It may be shown that, if two or more sets of particles are in equilibrium together, their mean kinetic energies of translation are equal; hence, in this case, $m_1\overline{V_1^2} = m_2\overline{V_2^2} = \text{etc.}$, and therefore $p = \frac{1}{3}(N_1 + N_2 + \text{etc.})\, m\overline{V^2}$, showing that the pressure depends upon the *total number* of particles per unit volume, not upon their masses. The same statement is *assumed* to be true of gases, and is equivalent to "Avogadro's hypothesis" (page 201); and no known fact contradicts it, provided the gas is not too dense. It often happens that when a complex gas is raised to a high temperature, the pressure increases abnormally, which always corresponds to a dissociation of the molecules of the gas into simpler parts;

the extent of the dissociation is calculated from measurements of the pressure, assuming the truth of the above formula.

Again, if there are two sets of particles that have the same pressure and the same mean kinetic energy of translation, *i.e.* if

$$\tfrac{1}{3} m_1 N_1 \overline{V_1^2} = \tfrac{1}{3} m_2 N_2 \overline{V_2^2}$$

and

$$\tfrac{1}{2} m_1 \overline{V_1^2} = \tfrac{1}{2} m_2 \overline{V_2^2},$$

it follows at once that $N_1 = N_2$; that is, that the two sets have the same number of particles in each unit volume. This statement when applied to gases would be: in two gases at the same pressure and temperature there are the same number of molecules in each cubic centimetre. This is known as "Avogadro's hypothesis," having been proposed by this Italian chemist in 1811, and it serves as the basis of the standard methods for the determination of "molecular weights" in chemistry. The "molecular weight" of a gas is a number that is proportional to the mass of one of its molecules. Thus, if M_1 and M_2 are the molecular weights of two gases, $\frac{M_1}{M_2} = \frac{m_1}{m_2}$. Then, since $N_1 = N_2$, *under the above conditions*, the densities of the two gases, $d_1 = m_1 N_1$, $d_2 = m_2 N_2$, are in the same ratio as the masses of the individual particles; and, if an arbitrary number is assigned as the molecular weight of one gas, that of the other is found by measuring the ratio of the densities of the two gases at the same temperature and pressure, and multiplying by the arbitrary number. For

$$\frac{M_1}{M_2} = \frac{m_1}{m_2} = \frac{d_1}{d_2}, \text{ or } M_1 = M_2 \frac{d_1}{d_2}.$$

The accepted system of molecular weights is based upon 32 being the number assigned to oxygen gas. (It may be seen at once that Avogadro's hypothesis is not true if a gas is greatly compressed, because, as has been said in speaking of Boyle's law, if the pressure is very great, the gas is not compressed according to this law, and so there is a smaller number of molecules in a unit volume than is required by Avogadro's hypothesis. But, if the pressure is small, there are no known facts contrary to it.)

Viscosity and Diffusion. — As has been shown before, we can explain the viscosity of a gas and the diffusion of one gas into another, and we shall soon see that we can explain the method by which a gas "conducts heat" by assuming that its particles have such freedom of motion that they can move about uninfluenced by other particles except when they come very close together, *i.e.* when they have what may be called an "encounter." In the interval of time between two encounters the particle is moving in a straight path with a constant speed; the length of this path is called the "free path," and its average value for all the particles is called the "mean free path" of the gas. When two particles have an encounter, their centres come within a certain distance of each other and then separate; one half of this minimum distance is called the "radius of the particle."

Similarly, if we consider a set of minute elastic spherical particles, we can explain its viscosity and the manner in which any increase in the kinetic energy of one portion is distributed throughout the whole set; and if we have two sets of such particles, we can explain the diffusion of one into the other. Further, we can calculate what the force of viscosity, the rate of distribution of kinetic energy, *i.e.* of conductivity of "heat," and the rate of diffusion of such sets of spheres are, expressing these quantities in terms of the mass of a particle, its mean energy, its mean free path, its radius, and the number of spheres in a unit volume. Then, if we assume that an actual gas behaves approximately like a set of spheres, we can measure its pressure and density at a given temperature, its viscosity and conductivity for heat, and its rate of diffusion, and, by comparison with the mechanical formulæ deduced for a set of spheres, obtain approximate values for the various properties of a gas molecule. A few of these may be mentioned. At a pressure of 76 cm. of mercury and a temperature of about 20° C. (*i.e.* 70° F.), the mean free path of a hydrogen molecule is 0.0000185 cm., and the number of impacts it makes in a second is 9480 million; for oxygen, these figures are 0.0000099 cm. and 4760 million; for carbonic acid gas, 0.0000068 cm. and 5510 million. By various processes the dimensions of a molecule and the number in a unit volume may be approximately determined; the "radius" of a molecule is found to be of the order of a ten-millionth of a millimetre, and the number in a cubic centimetre is of the order of 2×10^{19}, *i.e.* twenty quintillions.

Fourth State of Matter. — If a gas is inclosed in a glass bulb which can be gradually exhausted by means of an air pump, as will be explained later, the most evident change produced is the decrease in density and the consequent increase in length of the mean free path. (If the exhaustion is carried so far that the pressure in the bulb is that of one thousandth of a centimetre of mercury, the mean free path is 7630 times as great as it is at the pressure of 1 cm., or about 1.5 cm. for hydrogen.) The properties of matter in this condition are quite different from those of ordinary gases; and for this reason the matter is now said to be in a "Fourth State." Its chief properties were investigated by Sir William Crookes, and they will be described later when electrical phenomena are discussed. One purely mechanical property should, however, be mentioned here. It is illustrated by the following experiment: a framework is made consisting of two or more cross arms, which carry at each end a small piece of mica blackened on one face and not on the other; the plane of each mica vane is perpendicular to that of the cross arms, but includes the line of direction of the arm which carries it, and the blackened face of one vane is turned toward the polished face of the next one. This little wheel is suspended in a bulb in such a manner as to be free to turn about an axis perpendicular to the plane of the cross arms. If a hot body, like a burning match, is brought near the bulb, nothing noticeable happens if the gas inside the bulb is at ordinary pressure; but, if the gas is exhausted to a few thousandths or hundredths of a centimetre of mercury, a stage is reached when the framework begins to rotate on its axis in the direction which it would move if the blackened face of each vane were repelled by the hot body. (If the gas is exhausted as completely as possible, this motion does not arise; and at slightly higher pressures there are complications in the phenomena which need not be discussed here.) The explanation of this action is as follows: a blackened surface becomes hotter than a polished one when exposed to a source of heat, as is known to every one, and therefore those molecules in the gas that rebound from or leave the blackened face of a mica vane do so with a greater velocity than those that rebound from the polished faces. If the gas, however, is at ordinary pressure, the impacts are so numerous that the condition throughout the bulb is nearly uniform, and there is no mechanical action on the vanes; but if the gas is so rarefied that the mean free path is a centimetre or more, the molecules leaving the blackened face give the vane a backward push which is not counterbalanced by that given the other face of the vane, and consequently the framework rotates in the direction described above. This instrument is called a "radiometer"; and its action was discovered by chance by Sir William Crookes in the course of a research which involved weighing small amounts in a vacuum. It has been modified recently so as to serve as an instrument for measuring certain quantities that are of importance in the discussion of heat phenomena.

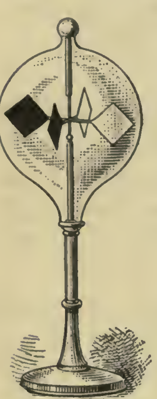


FIG. 100.—Crookes' radiometer.

CHAPTER IX

HYDRAULIC MACHINES: PUMPS, ETC.

Barker's Wheel. — This is a simple instrument, consisting of a fixed tube to one end of which is attached by a pivot a framework consisting of a number of cross tubes, whose ends are bent at right angles to their length and to that of the fixed tube. Connection is made through the pivot from the fixed tube into the movable ones. If now a stream of some *fluid*, either liquid or gas, is forced through the fixed tube and out the bent ends of the cross tubes, the latter will rotate rapidly in a direction opposite to that in which the fluid escapes. This is then simply an illustration of the conservation of momentum. These wheels are often seen connected with devices for sprinkling lawns with water.

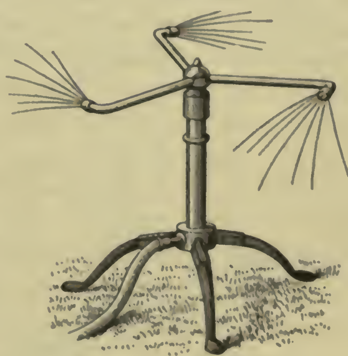


FIG. 101. — Barker's wheel.

Turbines. — A turbine consists of a wheel that can rotate on an axis and that has for its spokes curved flanges so arranged that as a fluid presses against them it exerts a moment around the axis. (Compare the action of the wind on the fans of a windmill.) In practice, a water turbine is placed at the bottom of a deep well (or high cylindrical tube) which is kept filled with water so that the wheel is under great pressure; the water is admitted through the turbine near its axis, flows out along the flanges, and escapes at the edges, so that the wheel is set in rotation by the pressure. In a similar manner steam can be used to drive a turbine, as is done in the so-called "turbine boats," in which there are several turbines fastened directly to the shaft of the boat.

Hydraulic Ram. — In this instrument, which was invented by Montgolfier, in 1796, and is in such general use for forcing water from springs into tanks at a considerable elevation above them, the principle made use of is that a large quantity of water falling through a small distance may raise a small quantity through a great distance. A simple form is shown in the cut. The essential features of the machine are a large tube, down which the water flows and which is closed by the escape valve *P* opening *inward*, and by another valve *Q* opening *outward* into an air-tight reservoir called the "air chamber." This contains some air; and into it enters for some distance the outlet pipe which carries the water to the tank. The escape valve *P* has a weight which exceeds slightly the upward force against it due to the water when there is no flowing — this upward force is the excess of that on the lower side of the valve over the downward force on its upper side. Therefore, at the beginning of the operation the valve drops; as it does so, the water escapes, and, since in moving water the pressure is less than in water at rest, the downward force on the upper side of the valve, over which the water is flowing, is diminished, and the upward force is

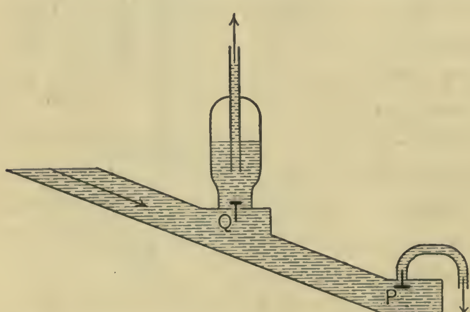


Fig. 102. — Hydraulic ram. *P* and *Q* are valves opening automatically.

sufficient to raise the valve and close it, thus stopping the flow; the valve therefore again drops owing to its weight, and the operation is repeated automatically. When the water is at rest at the beginning of the operation, the other valve Q is down, closing the opening; it remains so as long as the escape valve is open, for the pressure on its lower side is now small since the water is flowing. When the escape valve closes, there is an immediate increase in pressure throughout the whole tube, the outlet valve Q is pushed up and some water enters the air chamber; then the valve drops as the pressure is thus relieved, and the operation is repeated. As more and more water enters the air chamber, a time is reached when the level of the water covers the open end of the outlet pipe which connects the chamber with the tank; after this time, as the water enters, the air trapped above it is compressed and has its pressure increased. Water is thus forced up the pipe into the tank. This operation is more or less continuous; for, as the water enters the air chamber rapidly, the air is compressed and some water flows up into the tank; and then, during the interval of time which passes before some more water enters, the compressed air expands and continues the flow.

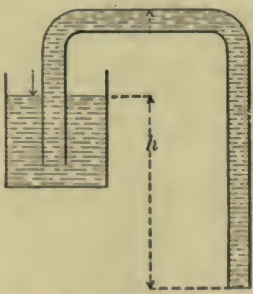


FIG. 103. — Siphon.

Siphon. — This consists of a large tube or pipe bent into the form of a **U**, but with its two arms of unequal length. This is placed in a vertical position, with its shorter arm dipping below the surface of a liquid in a vessel, and its longer arm outside. By means of suction applied to the open end, the siphon is now filled with liquid; and, if left to itself, the liquid in the vessel will flow out through the siphon until its surface comes below the end of the short arm. The explanation is evident if one considers the conditions that exist when the siphon is full, at the instant before the flow starts. The pressure in the liquid in the longer arm at a level with the liquid surface in the vessel is, of course, that of the atmosphere P; so, if the open end of the siphon is at a depth h below this, the pressure in the liquid at this point is $P + dgh$ if its density is d. But the opposing pressure is simply that of the atmosphere P, and the difference of pressure dgh forces the liquid out.

The shorter arm of the siphon must not be too long; for if it is greater than the height to which the liquid would rise in a barometer, the pressure on the free surface of the liquid will not be sufficient to force the liquid up to the turn of the siphon. (If the tube is of fine bore, other actions than gravity and atmospheric pressure come into play.)

Liquid Pumps. — These are instruments devised for the purpose of raising liquids from one tank or well into another at a greater height, or for forcing a liquid through a long pipe against friction. There are two types: the "lift pump" and the "force pump." The former consists of a cylinder in which fits an air-tight piston provided with a valve B opening upward, and whose lower end is closed by a valve A, also opening upward, where the pipe leading to the tank containing the liquid to be raised is attached. The vertical distance from the lower end of the cylinder to the level of the liquid in the well or tank must not exceed the barometric height of that liquid. Then if the piston is raised, some liquid is forced up through the lower valve into the cylinder by the pressure of the atmosphere on the surface of the liquid in the tank; if now the piston is brought to rest and then pushed down, the lower

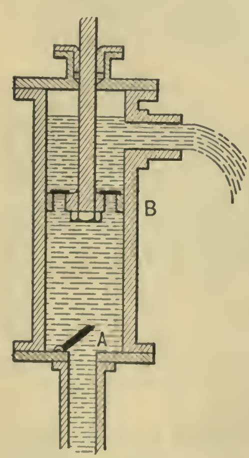


Fig. 104. — Lift pump. A and B are valves opening upwards.

valve drops and the one in the piston is lifted, the liquid passing through it from below the piston to the space above. When the piston is again raised, the liquid on top of it is lifted and may escape through a side outlet into a tank; at the same time more liquid is being drawn up through the lower valve into the cylinder, and the process may be repeated indefinitely.

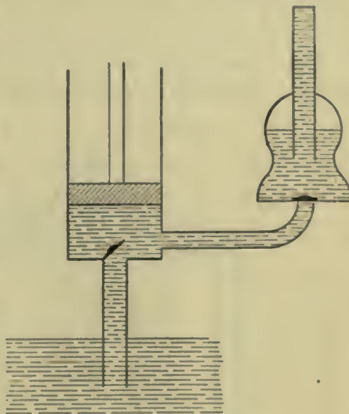


FIG. 105. — Force pump.

In the force pump the piston has no valve, and an air chamber, like that of the hydraulic ram, is attached to one side of the cylinder. The explanation of its action is self-evident. This pump is as a rule placed near the surface of the liquid which is to be pumped, and the upper tank may be as high as is necessary.

Air Pumps. *a. Mechanical.* — These are instruments devised either to force more and more gas into a given space, or to withdraw as much gas as is desired from a closed vessel; in other words, to increase or to decrease the pressure inside the vessel. The former are called "compression" pumps; the latter, "exhaust" pumps.

The simplest form of exhaust pump is illustrated on page 210. Its mode of action is essentially that of the lift pump already described, the main point of difference being that in the latter the valves open and close automatically, while in the air pump they must be operated by mechanical means, since the difference in pressure of the gas on the two sides of the valves is not sufficient to move them. Such pumps as this are called "mechanical" ones. Other forms in general use are the Sprengel and the Geissler-Toepler.

b. Sprengel pump. — The action of this pump consists in having drops of mercury so fall as to trap the gas between them and thus carry it away. There is an elongated glass bulb, to the side of which is joined a long tube, as shown in the cut, whose lower end is connected by a rubber tube with a reservoir, so that the mercury may be thus forced

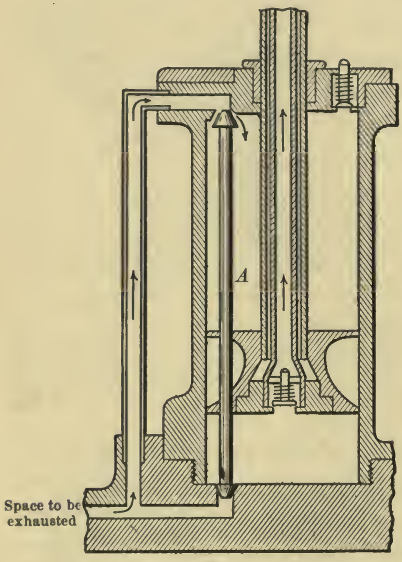


FIG. 106. — Mechanical air pump. The vertical rod, *A*, is held by the moving piston with sufficient friction to move it up or down until brought to rest by the conical ends entering their sockets; then the piston slips along the rod. In the cut the piston is moving down.

FIG. 107. — Sprengel air pump.

slowly into the bulb. At the lower end of the bulb is joined a glass tube of narrow bore and at least 80 cm. long, and at the upper end of the bulb is a connection with the space to be exhausted. The tubes at the side and bottom of the bulb are so arranged that, as the drops of mercury break off and fall, they hit the opening of the lower tube and pass down it in the form of short cylinders. The space between these cylinders thus formed is occupied by small amounts of the gas, drawn in from the connected vessel; and so these drops act like a succession of small pistons forcing out the gas. The lower end of the long tube may dip into a basin of mercury, and the gas will bubble out at the surface, or it may be bent so as to form a "trap." As the exhaustion continues, the mercury will rise in the long tube, and will finally stand at the barometric height when the vacuum is as complete as it can be made.

c. Geissler-Toepler pump. — In this pump there is a large bulb to which are joined two tubes, — one at the top, the other at the bottom. The lower one is at least 80 cm. long and is connected at its lower end to a large vessel of mercury by means of a long rubber tube. The upper tube is bent over into a vertical direction downward, and dips into a basin of mercury, or forms a trap. Around the large bulb there is a branch tube connecting the upper and lower tubes just as they leave the bulb; and into this branch is joined a long vertical tube leading to the vessel which is to be exhausted. (This tube is replaced often by a short vertical one containing a glass valve.)

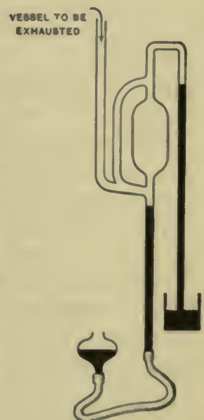


Fig. 108. — Geissler-Toepler air pump.

If the large vessel of mercury is now raised, as it can be owing to the flexible rubber tubing, the mercury will rise into the bulb and the connecting tubes, shutting off connection with the vessel to be exhausted, and will drive out all the gas in the bulb through the tube in the top, so that it will bubble out through the mercury in the basin at its end. If, now, the movable vessel of mercury is lowered, no air can enter through the tube at the top of the bulb, because it is "sealed" by the mercury in the basin, which will rise in the tube; but as soon as the mercury falls below the opening to the long vertical tube, the gas in the vessel to be exhausted will expand and fill the bulb and the connecting tubes. When the movable vessel of mercury is again raised, it drives out the gas in the bulb; and as the process continues, the exhaustion of the vessel proceeds rapidly. The tube leading from the top of the bulb around to the basin of mercury must be at least 80 cm. high, and the long vertical tube leading to the vessel to be exhausted must be still longer.

BOOKS OF REFERENCE

KIMBALL. The Physical Properties of Gases. Boston. 1890.

BARUS. The Laws of Gases. New York. 1899.
This contains Boyle's original paper and also Amagat's memoirs on the variations of a gas from Boyle's law.

BOYS. Soap Bubbles. London. 1890.
This is a description of many most interesting capillary phenomena.

GREENHILL. Hydrostatics. London. 1894.
A standard advanced text-book.

TAIT. Properties of Matter. Edinburgh. 1885.
A most interesting and useful text-book.

RISTEEN. Molecules and the Molecular Theory of Matter. Boston. 1895.
A popular, yet accurate, description of the kinetic theory of matter.

HOLMAN. Matter, Energy, Force, and Work. New York. 1898.
This is a philosophical discussion of the properties of matter.

POYNTING AND THOMSON. Properties of Matter. London. 1902.
This is an advanced text-book, and will be found most useful for reference.

JONES. Elements of Physical Chemistry. New York. 1902.
This contains an excellent description of the properties of solids, liquids, and gases.

HEAT

INTRODUCTION

The properties of matter that have been discussed in the previous pages are mass, weight, shape, size, elasticity, pressure, etc. The mass of a body cannot be changed by any mechanical means, nor can its weight at any one point on the earth's surface; but the other properties may be changed at will. One of the simplest methods of doing this is to alter the temperature of the body; and this process will be discussed in the following pages.

Molecular Energy. — We have proved, in the discussion of diffusion, of viscosity, and of the properties of gases, that matter consists of minute parts which are in motion, the extent of the freedom of this motion varying with the condition of the matter. In a solid these minute particles as a rule only make oscillations; while in fluids they can move from one portion of space to another. Thus these particles have both kinetic and potential energy, — the former owing to their motion, the latter owing to the fact that work is required to bring two molecules into a certain definite position with reference to each other. The molecules themselves consist of parts, and these have energy in both the kinetic and the potential forms.

This *internal* energy of a body is quite apart from its energy owing to its motion as a whole, or to its position with reference to the earth, and may be increased, as is obvious, by doing work against forces that act in connection with the molecules, *e.g.* by overcoming friction or by compressing a gas.

It may be varied in the case of an elastic body by setting it in vibration, or by sending waves or pulses through it; for under these conditions the kinetic and potential energies of the molecules are altered. Thus a bell if struck by a hammer vibrates, and as a result waves are produced in the surrounding air, the particles of which therefore are set in vibration.

We may, therefore, consider two kinds of internal molecular motions: one corresponding to a state of wave motion when all the particles or molecules are in similar vibrations; the other, to a condition in which there is no regularity in the vibrations or motions. This last condition exists when a body is in its ordinary undisturbed state, and is altered when friction is overcome, when a gas is suddenly compressed, etc. The phenomena associated with variations in this internal energy of bodies, owing to their *irregular* molecular motions, belong to that branch of Physics which is called "Heat."

CHAPTER X

HEAT PHENOMENA

Preliminary Ideas. — In describing and discussing mechanical phenomena the sequence of ideas was somewhat as follows: by means of our muscular sense we experience certain sensations which we associate with matter, and we are led to distinguish certain properties of matter which we consider as independent of each other for the time being, viz., mass, weight, elasticity, etc.; we study these at first by means of our own muscles, but later discover physical methods for the same purpose. By means of our other senses we can also investigate other properties of matter and the corresponding "laws of nature." Every one knows what is meant by the words "hot" and "cold"; and if a man dips his hand in turn into two basins of water, he is as a rule able to distinguish between them by means of his temperature sense, and so can say one is hotter or colder than the other. We experience this sensation of hotness when we stand in the sunshine or near a fire, when we put our hands in a basin of water that is on a stove, when we touch a body that has been rubbed violently against another, etc.; and we feel the sensation of coldness when we touch or stand near a block of ice, when we wet our hands and allow the water to evaporate, etc. If we expose inanimate objects to the same conditions, they undergo changes; and, in fact, in general all their physical properties with the exception of their mass and weight change.

Thus, if a piece of iron is exposed to the sun or put on a stove, its volume increases, its elasticity changes, it feels hot to our fingers. If a piece of ice is put in a basin on a stove, it changes its state, becoming a liquid ; this water gradually feels hotter and hotter to our fingers, and finally boils away in the form of a gas. If a gas is inclosed in a glass bulb and exposed to a flame, both its pressure and volume change, etc. All these changes could be produced equally well by friction. Similarly, if a piece of iron is put on a block of ice, its volume becomes smaller and it feels cold to our hands ; water can be frozen by making some of it evaporate rapidly, etc. These changes are called "heat effects."

Nature and Cause of Heat Effects. — If we investigate the conditions under which these changes occur, we see that in them all work is being done either on the minute portions of the body or by them. We shall consider one or two of these conditions in detail. When two solid bodies are rubbed together, the force of friction is overcome, and the changes produced, in general, are increase in hotness, increase in volume, etc. In friction, however, the force owes its origin to minute inequalities in the two surfaces, which are leveled off or altered as the work is done. Similarly, in all cases of fluid friction, the work done in maintaining the motion is clearly spent, as has been shown, in giving energy to the molecules or minute moving parts. An illustration of the heating effect produced by friction is furnished by meteors and "shooting stars." As these pieces of matter enter the atmosphere of the earth, they are heated to incandescence by friction against the air.

Again, when a gas is compressed, it becomes hot and its pressure increases; but, as we have shown on page 197, in this case work is done in increasing the kinetic energy of translation of the particles of the gas. If the gas is allowed to expand, doing work against some external force, it becomes cold and its pressure decreases ; but it has been shown that while this happens the molecules of the gas lose kinetic energy. In a flame, or any process of combustion, there are molecular changes going on which must necessarily involve the energy of the molecules. It will be shown later that all bodies are emitting waves in the ether; and, if those whose lengths are comparable with the size of molecules are absorbed by bodies, they become hot, expand, etc. This is the explanation of the effect of exposure to sunlight, etc.

Again, if we consider the effects themselves that are produced in bodies under the conditions described, it is seen that they correspond to what we would expect if they are due to work being done on or by molecules. When a body expands, the molecules are separated, and work is done against the molecular forces; similarly, if a body contracts, these forces do work. When a solid is melted, the molecules are given greater freedom of motion, and therefore work must be done on them; similarly, when a liquid solidifies, the molecules lose their freedom of motion, and in this change lose energy. As a gas becomes hot, owing to compression, its molecules gain kinetic energy of translation; as the gas cools, owing to expansion against external forces, its molecules lose kinetic energy of translation; so that in the case of a gas, at least, changes in temperature correspond to changes in kinetic energy. Similarly, changes in elasticity, hardness, viscosity, etc., may be explained if we think of the energy of the molecules as being affected.

It is natural, therefore, to believe that all these changes are due to the fact that the energy of the molecules has been increased by the result of external forces or that it has been decreased by the molecules doing work against external forces. Then two questions arise: Do these changes depend simply upon the *quantity* of energy that the bodies receive or upon the *manner* in which the work is done? And is the energy that the molecules receive exactly equal to the work done on them? The experiments that have been performed in order to answer these questions will be described later;

but it may be stated here that there is every reason for believing that the changes depend upon the quantity of energy received by the molecules, not upon the manner in which it is delivered, and also that there is neither loss nor gain of energy in the process.

Application of the Principle of the Conservation of Energy. — During most of these changes the volume of the body that undergoes them alters; and, in this process, work is done on or by whatever external force or pressure is acting on the body maintaining its volume, and, in general, by the force of gravity also. Thus, if a pillar supporting a building expands, the building is raised and work is done; the air presses against the sides of the pillar, and this force is also overcome as it expands; again, the centre of gravity of the pillar itself is raised, and thus more work is done against gravity. Consequently, when work is done against the molecular forces of a body so that it undergoes changes in temperature, in size, in state, etc., a definite amount of energy is given the body and this is spent in two ways: (1) in increasing its internal energy; (2) in doing external work as just described. (During these changes the molecules are affected, and some of their kinetic energy may become potential or *vice versa;* but in these *internal* changes there is neither loss nor gain of energy.) Similarly, when reverse changes take place and the body becomes cool and contracts, the external forces of pressure and gravity do work on the body, the internal energy decreases, and the molecules of the body do work in such a manner as to give energy to external bodies; the amount of this last must therefore equal the sum of the work done on the body by the external forces and the amount of the decrease in the internal energy. Thus, when a piece of iron is placed in a basin of hot water, the latter loses a certain amount of internal energy, and, since it contracts, the atmospheric pressure does work on it; and in return for these two supplies of energy the internal energy of the iron increases, and as it expands it pushes back the atmosphere and so does work. (We are neglecting purposely all losses of energy by radiation and conduction.) In any change, then, we may write the equation:

quantity of energy received by work done on the molecules = increase in internal energy + work done against external forces, such as surface pressure and gravity,

or the equivalent one:

quantity of energy given up by the molecules doing work against external forces = decrease in internal energy + work done on the body by external forces, such as surface pressure and gravity.

Heat Energy. — The energy that is given a body when work is done on it against molecular forces is called "heat energy," and the effects that bodies experience when they gain or lose this energy are called "heat effects." In ordinary language the word "heat" is often used in place of heat energy, and we speak of "adding heat to a body," or withdrawing it, of a "source of heat," etc., where the meaning is obvious. As has been said, all the properties of a body except its mass and weight in general change when heat energy is added to it or taken from it. The most obvious of these changes are the following:

1. Change in hotness, as perceived by our temperature sense.

2. Change in volume, if the external pressure is kept constant.

3. Change in pressure, if the volume is kept constant, in case the body is a fluid.

4. Change in state, such as fusion, evaporation, etc.

5. Change in electrical or magnetic properties, such as electrical conductivity, magnetic strength, etc.

Heat Quantities. — In the discussion of heat effects two physical quantities enter that have not hitherto been described with exactness. In the first place, we must explain what is meant by the word "temperature," which is used ordinarily as giving an idea of the hotness of a body, and must describe a method by which a numerical value may be assigned it. Further, in all heat effects we are concerned primarily with quantities of energy entering or leaving bodies; and some convenient unit must be defined in terms of which this energy may be measured. Changes in volume and pressure can be measured by means already described, and may be expressed in terms of the ordinary units — the cubic centimetre and dynes per square centimetre or centimetres of mercury. In changes of state we have to consider alterations in volume, in elastic properties, etc.; but these require no new definitions or units.

Temperature

Preliminary Ideas in Regard to Temperature. — We have used, whenever convenient, the word "temperature" in its everyday meaning as a quantity describing the hotness of a body, and have said that a body which felt hot to us had a higher temperature than one which felt cold. This sensation is due to some property of the molecules of the body, in virtue of which they affect our nerves. We saw in speaking of gases, page 197, that the mean kinetic energy of translation of its molecules obeys the same laws as does its temperature; or, in other words, the temperature or hotness of a gas is due to and is measured by the kinetic energy of translation of its molecules. There are many reasons for believing that this is true also of solids and liquids; and so, when the temperature of a body is raised, we believe that the kinetic energy of its molecules is increased.

Temperature Scales; Thermometers. — Evidently this property of a body cannot be *measured*, because it is impossible to conceive what is meant by a unit of hotness; but we can assign it a numerical value. For when the temperature of a material body is changed, its physical properties all change; and so, instead of using our hands or bodies as instruments for investigating temperature, we can employ any material body and observe those properties which change as the temperature is changed. This body would then be called a "thermometer." Thus we might use a homogeneous metal rod and observe its length. If the rod had the same length when immersed in two different baths of oil, we should say that their temperatures were the same with this thermometer; whereas, if the length were greater when the rod was in one bath than when in the other, the former would be said to have the higher temperature.

Experiments show that, if two bodies at different temperatures are placed together in such a manner that heat energy can pass between them, *e.g.* if two liquids are stirred up together, if a solid is immersed in a fluid, etc., they finally come to the same temperature intermediate between their two initial temperatures; the body at the lower temperature must therefore gain energy, and the one at the higher temperature must lose it. (The former also loses energy in general, but it receives more than it loses, and so on the whole gains. Similarly, the latter in general gains energy, but its loss exceeds its gain.) Thus, from a physical point of view, the difference of temperature between two bodies determines which is to lose energy, or it determines the "direction of the flow of heat energy." The temperature of a body, then, is a property defining its thermal relations with neighboring bodies.

The general method of assigning a number to the temperature is as follows: two standard thermal conditions are selected, and numbers are given them arbitrarily — let these be t_1 and t_2; then some definite property of a definite body, which can be measured and which changes with the temperature, is selected; its numerical value is determined in the two standard conditions and in the one to which a number is to be assigned — let these values be a_1, a_2, and a; finally, the number for the temperature is obtained from those arbitrarily given the standard conditions by simple proportion between the numbers thus obtained, viz., calling it t,

$$t - t_1 : t_2 - t_1 = a - a_1 : a_2 - a_1.$$

Experiments show that the temperatures of a mixture of pure ice and water when in equilibrium, and of steam rising from boiling water, are constant and the same the world over, provided the external atmospheric pressure is the same; that is, a definite metal rod always has the same length if it is put in a bath of water and ice no matter when or where it is done, a given quantity of mercury has the same volume, etc. For this reason, and because they are easily obtained and include the ordinary temperatures, these two thermal conditions when the external pressure is 76 cm. of mercury are selected as the standard ones; and the numbers 0 and 100 are assigned them on the "Centigrade" or Celsius scale. The quantity agreed upon by physicists, the change of which is measured, is the pressure of a definite amount of hydrogen gas whose volume is kept constant. Let the values of the pressure of this gas at the standard temperatures be p_0 and p_{100}, and that at a temperature for which a number is desired p. Then, calling this number t, it is given by the equation:

$$t - 0 : 100 - 0 = p - p_0 : p_{100} - p_0,$$

or

$$t = 100 \frac{p - p_0}{p_{100} - p_0}.$$

This is the "temperature on the constant volume hydrogen thermometer," using the Centigrade scale; and, *whenever hereafter the temperature of a body is referred to, its value on this instrument and scale is meant.* This number is always expressed as a certain number of "degrees," and is written $t°$ C.

Several other thermometers are in more or less common use. In one the volume of a definite quantity of nitrogen (or of air), the pressure being kept constant, is the property measured as the temperature changes; in another it is the *apparent* volume (see page 234) of a quantity of mercury inclosed in a glass bulb with a fine stem that is measured. The former is called a "constant pressure nitrogen (or air) thermometer"; the latter, a "mercury-in-glass thermometer"; in them both the number for the temperature on the Centigrade scale is given by

$$t = 100 \frac{v - v_0}{v_{100} - v_0},$$

where the symbols have obvious meanings.

Another form of instrument is one in which the electrical resistance of a piece of platinum wire is measured. If R_0, R_{100}, and R are the values of this quantity at the three temperatures, the number for the temperature is given by

$$t = 100 \frac{R - R_0}{R_{100} - R_0}.$$

Such a thermometer as this is called a "platinum resistance" instrument.

The numbers obtained by these last three instruments for the temperature of the same body will not agree with each other, or with that obtained with the standard thermometer, except at 0° and 100°; but careful comparisons have been made by different observers of the readings on all four instruments at the same temperatures through wide limits; so that, if a number is obtained for the temperature of a body using a mercury thermometer, the value which would have been obtained if the standard hydrogen instrument had been used is known: this is the temperature of the body. (The correction to be applied to the scale reading of an ordinary mercury thermometer, such as is used in laboratories, is small.)

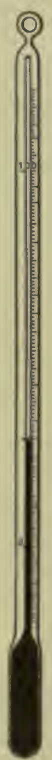
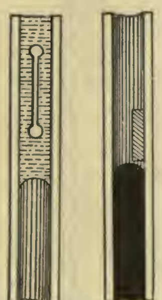


FIG. 109.—Mercury-in-glass thermometer.

Other scales than the Centigrade are in use for non-scientific purposes. In the Fahrenheit system the numbers 32 and 212 are assigned the standard temperatures of melting ice and steam, and in the Réaumur the numbers 0 and 80 are used. (Thus, 20° C. = 68° F. = 16° R.)

The ordinary laboratory thermometer is a mercury-in-glass instrument, whose stem is divided into numbered parts. The maker of the instrument aims to have these numbers come at such points that, when the division corresponding to the top of the mercury column is read, it gives the temperature on the mercury scale, *i.e.* the numbers correspond to equal increments in apparent volume. If the instrument is to be used for accurate measurements, however, the readings at the standard temperatures must be noted, and the volume of different portions of the stem must be measured, in order to determine exactly the error of each division as marked by the maker. Moreover, a glass thermometer is subject to an error due to two facts: a *glass* bulb whose temperature is raised from one value to another, and then lowered again to the former value, has a larger volume at the end than it had at the beginning; and this increase is not permanent, but disappears gradually after the lapse of weeks or months. This is owing to the heterogeneous character of glass; the molecular changes produced by raising the temperature persist after it is again lowered. Thus, if a glass thermometer reads 0°.02 C. when put in melting ice and is then, after being heated to 80° or 90°, again put in melting ice, it may read − 0°.01 C., showing that the volume of the glass bulb has increased. This is known as the "depression of the zero point." If the thermometer is kept in ice for some months, the readings will gradually rise. There are numerous other defects in the mercury thermometer which must be carefully guarded against.



To give a number to extremely low temperatures some substance should be used whose properties are the same in kind as at ordinary temperatures. Thus, mercury should not be used, for it solidifies at about − 39° C., and the changes in volume of the solid mercury cannot be compared with the similar changes of liquid mercury at ordinary temperatures. Hydrogen gas at a small pressure may be used, or a platinum resistance thermometer.

Similarly, to give numbers to extremely high temperatures special precautions must be taken. The best

FIG. 110. — Rutherford's maximum and minimum thermometers. The former contains mercury, the latter alcohol.

practical methods depend upon certain laws of radiation which will be discussed later; but for standardizing purposes a hydrogen thermometer must be used.

There are many special types of thermometers devised for particular purposes. Among these it may be worth while to describe briefly one that registers the extreme temperatures which occur during a certain interval of time, and one that is used by physicians for clinical purposes. Rutherford's "maximum and minimum thermometers" are two instruments, as shown in the cut, which are supported with their stems horizontal: one contains alcohol, the other mercury. Just *inside* the end of the alcohol column is a small glass rod with rounded ends, as shown on an enlarged scale in a portion of the cut; so, as the temperature falls and the alcohol contracts, this rod is drawn back by the curved surface of the alcohol; and, when the temperature rises again, it remains stationary. Similarly, just *outside* the end of the mercury column is a small iron rod which is pushed forward by the mercury as it expands and remains stationary when the mercury contracts. Thus both the lowest and the highest temperatures reached are recorded. The glass index can be jarred back in place again, and the iron one can be drawn back by a magnet. (In Six's form of instrument these two thermometers are combined in one.)

FIG. 111. — Clinical thermometer.

A clinical thermometer is shown in the cut. Its stem has a fine bore with a swelling at its lower end separating it from the bulb containing the mercury; so that, as the temperature rises, the mercury does not reach the divided stem until a temperature of 90° F. is reached. At the swelling there is a constriction in the tube; and if the temperature falls when the mercury is above this, the column breaks off, leaving the mercury in the stem. Thus, if the instrument is inserted in the mouth and the mercury rises to a certain division, it will still stand at this point after the thermometer is withdrawn. The mercury may be driven back into the bulb by shaking the instrument in the proper manner. (Centrifugal force is usually applied.)

The first thermometer was made by Galileo, as early as 1593. It depended upon the expansion of air, and was not devised in such a manner as to be independent of barometric changes. Boullieau, in 1659, made the first mercury-in-glass instrument. The credit of showing that melting ice and boiling water furnish "fixed points," and of proposing that they be adopted in a thermometric scale, belongs to Huygens (1665). Many improvements were made by Fahrenheit (1724).

Units of Heat Energy. — In all heat effects, as has been said, we are concerned with the quantity of energy that must be added to or withdrawn from a body in order to produce a given change; and so a convenient unit must be chosen, and suitable methods of measurement must be devised. The scientific unit for the expression of amounts of all kinds of energy, including therefore heat energy, is the erg or the joule, *i.e.* 10^7 ergs (see page 112); but heat effects are not as a rule produced by direct mechanical processes in which the amount of work done can be measured by a dynamometer. The standard method of producing a heat effect in a body is to immerse it — if it is a solid — in a quantity of water at a different temperature; the temperature of the water falls or rises because it loses or gains the heat energy that enters or leaves the body, allowance being made for external work and for the influence of surrounding bodies. The natural unit of heat energy is, then, either the amount required to raise the temperature of a unit mass of water through 1°, or one nth the amount required to raise its temperature through n°. (These two quantities of energy are not in general the same.) By stirring a paddle rapidly and continuously in a known quantity of water, the amount of work (measured in ergs) required to raise its temperature through a known number of degrees (on any scale) may be determined by a dynamometer; and so the value of the practical unit of heat energy may be expressed in ergs. Experiments show that the number of ergs required to raise the temperature of a definite quantity of water through 1° is different for different temperatures, *i.e.* it is not the same when the temperature is raised from 5° to 6° as if the limits were 10° and 11°, etc.; but the difference is very small. The work required, however, to raise the temperature of a definite quantity of water from 0° to 100° C. is almost exactly 100 times as much as that required to raise it from 15° C. to 16° C. For this reason the practical unit of heat energy is defined to be the "amount required to raise the temperature of 1 g. of water from 15° to 16° C."; this is called a "gram calorie at 15° C.," or, simply, the calorie. Its value in ergs, as determined by Rowland, Callendar and Barnes, and others, is 4.187×10^7; that is, it is 4.187 joules.

The great disadvantage in having as a "heat unit" one that depends upon a range of temperature (other than from 0° to 100° C.) lies in the difficulty of determining temperature accurately, and in the fact that so many arbitrary quantities and ideas enter into the definition of a temperature scale. If it were practicable, it would be much better to take as a heat unit the amount required to melt 1 g. of ice at 0° C., or to produce some other change in state, because during these changes the temperature does not vary.

Transfer of Heat Energy. — Before discussing the various heat effects in detail, a few words should be said in regard to the various methods by which heat energy is added to or taken from a body. These are three in number, and are illustrated in the following experiment: if one's hand is held above a heated stove, it feels hot, and at the same time one is conscious of an ascending current of air. Similarly, if a body is held in the upper portion of any fluid whose lower portion is maintained at a high temperature, it will receive heat energy from the ascending currents of heated fluid. This process is known as "convection." If one end of a metal rod is put into a fire, its temperature rises, and that of other neighboring portions of the rod also. In this process there is no actual displacement of the matter, and therefore there is no convection; but the energy is handed on from molecule to molecule down the rod. This is called "conduction." Again, if a body is exposed to the sun or is held at one side of a hot stove, its temperature in general rises, it is receiving heat energy — not, however, by convection or conduction. This process is called "radiation," and will be shown later to consist in the absorption by the body of waves in the ether. All these processes will be described in detail in a later chapter.

FIG. 112. — Dewar flask.

Convection and conduction cannot take place through a vacuum, and radiation is almost entirely prevented by having the surface of the body or vessel covered with a highly polished metallic layer. In his experiments on liquid air and hydrogen, Dewar has used a flask, called by his name, which consists of a double-walled glass vessel, the space between the walls being exhausted as completely as possible. Traces of mercury vapor are left in this space; and at low temperatures this freezes, forming a metallic surface over the glass walls.

CHAPTER XI

CHANGES IN VOLUME AND PRESSURE

Introduction. — The fact that all bodies, with the exception of water below 4° C. and one or two unimportant substances, increase in volume when their temperature is raised, is most familiar to every one; and numerous measurements have been made of these changes. Naturally the amount of the increase in volume depends upon the external force acting, and upon whether this is constant or not. The mechanical force required to influence the expansion of a solid or a liquid is so great that ordinary changes in the atmospheric pressure have no measurable effect; but this is not so in the case of a gas. It is therefore necessary, in studying those variations in volume of a body which accompany changes in temperature, to describe the external conditions, if one wishes to be definite. The condition which is always assumed, unless the contrary is stated, is that of constant external pressure.

Solids

Linear and Cubical Expansion. — In measuring the change in volume of a solid it is, as a rule, easier to measure the changes in length of certain linear dimensions of the body and from these to calculate the change in volume. If the body is isotropic, *i.e.* has the same properties in all directions, this calculation is most simple. Imagine the body in the form of a cube, the length of whose edges at any one temperature $t_1°$ is l_1 and at the temperature $t_2°$ is l_2, then the volume at $t_1°$ is l_1^3 and at $t_2°$ is l_2^3; so the change in volume is $l_2^3 - l_1^3$. If the body is not isotropic, but has different properties in different directions, three directions in it (which depend upon the arrangement of the molecules) may be determined, such that the changes along these are independent of one another — these are called "axes." (In crystals they are the crystallographic axes.) If, then, a rectangular solid is made of this body with its edges parallel to these directions, and if l_1, m_1, n_1 are the lengths of the edges at $t_1°$ and l_2, m_2, n_2 their lengths at $t_2°$, the corresponding volumes are $l_1m_1n_1$ and $l_2m_2n_2$; and the change in volume is $l_2m_2n_2 - l_1m_1n_1$.

The change in length of any straight line in the surface of a solid or of any straight edge may be measured by various means. The body is immersed in a bath of some fluid, whose temperature may be varied; and the lengths are determined by a comparator of some kind. (Reference may be made to any laboratory manual.) Experiments show that, to a sufficient degree of accuracy, the change in length varies directly as the original length and as the change in temperature, but is different for bodies of different materials. If, as above, l_1 is the length of a certain line of the solid body at $t_1°$, and l_2 that at $t_2°$, these facts may be expressed by the formula

$$l_2 - l_1 = al_1(t_2 - t_1),$$

in which a is a factor of proportionality, which is different for different substances. a is called the "coefficient of linear expansion of the body referred to the temperature $t_1°$." Ordinarily, the temperature to which these coefficients refer is 0° C.; and if l_0 is the length at 0° and l that at $t°$, the relation is $l - l_0 = a_0l_0t$, where a_0 is the coefficient of linear expansion referred to 0° C. This formula may be written $l = l_0(1 + a_0t)$.

(It is evident, then, that a in the first formula is connected with a_0 by the relation $a = \frac{a_0}{1 + a_0t}$; and if a_0 is an extremely small quantity, as it is for all solids, we may neglect the term a_0t in comparison with 1, and write $a = a_0$ approximately.)

If the solid is isotropic, and if we write v_0 for the volume at 0° C. and v for that at t°,

$$v_0 = l_0^3 \text{ and } v = l^3 = l_0^3(1 + a_0t)^3 = v_0(1 + a_0t)^3.$$

But $(1 + a_0t)^3 = 1 + 3a_0t + 3a_0^2t^2 + a_0^3t^3$; and if a_0 is a small quantity, the last two terms may be neglected in comparison with 1. So $v = v_0(1 + 3a_0t)$; or, finally, writing b_0 for $3a_0$

$$v = v_0(1 + b_0t).$$

b_0 is called the "coefficient of cubical expansion referred to 0° C."; and its numerical value is, from what has just been proved, three times that of the linear coefficient.

Similarly, if the body is not isotropic, and if a, a', a'' are the coefficients of linear expansion along the three axes, we may write for the lengths of three lines parallel to these at t° C.:

$$l = l_0(1 + at),$$

$$m = m_0(1 + a't),$$

$$n = n_0(1 + a''t).$$

Consequently, multiplying these three equations, we have the formula,

$$v = v_0(1 + at)(1 + a't)(1 + a''t)$$

$$= v_0[1 + (a + a' + a'')t],$$

if the quantities a, a', and a'' are all small. If we write $b_0 = a + a' + a''$, $v = v_0(1 + b_0t)$, as before. In certain crystals one of the coefficients of linear expansion is negative, *i.e.* lengths parallel to one of the axes diminish as the temperature rises; and, if its value exceeds numerically that of the sum of the other two coefficients, the volume will contract with increased temperature.

If a hollow body has its temperature changed, it expands or contracts exactly as if it had no cavities; these spaces increase in size or diminish according to the same law as the solid portions. This is at once evident if we regard the body as built up of solid blocks, and consider the change in size of these blocks as the temperature varies. An iron tire is shrunk on a wheel, or the various coatings on a large gun, by first raising their temperatures until they will just slip in place, and then cooling them.

The following table contains the values of the coefficients of linear expansion of a few familiar solids :

Brass	0.000019	Steel (annealed)	0.000011
Copper	0.0000168	Steel alloyed with 36% nickel	0.00000087
Glass	0.0000083	Zinc	0.0000292
Iron	0.0000121		
Platinum	0.0000090		

Illustration of Expansion. — The fact that the coefficients of expansion of different bodies are different is often made use of to neutralize the expansion which ordinarily follows rise in temperature. Thus, the period of a pendulum clock would naturally increase with rise in temperature owing to the expansion of the pendulum rod; but this may be avoided in several ways. One is to make the pendulum as shown in the cut, which illustrates Harrison's gridiron compensated pendulum. In it the rods which are shown as single black lines are of iron and the others of brass. It is seen that, as the temperature rises, the expansion of the iron rods lowers the pendulum bob, but that of the brass ones raises it. The linear expansion of brass is about $1\frac{1}{2}$ times that of iron; and so if the combined length of the two iron rods on each side and the middle one is $1\frac{1}{2}$ times that of the two brass ones on each side, there will be no change in length as the temperature varies. (A simpler method is used in the clock in the tower of the Houses of Parliament, London. The pendulum consists of an iron rod surrounded by a zinc cylinder, which in turn is surrounded by an iron one carrying the pendulum bob. The lower ends of the iron rod and the zinc cylinder are attached, and the upper ends of the two cylinders. Since the linear expansion of zinc is about $2\frac{1}{2}$ times that of iron, the combined lengths of the iron rod and cylinder must equal $2\frac{1}{2}$ times that of the length of the zinc cylinder.) Graham's compensation pendulum consists of an iron rod whose lower end screws into a cast-iron cylindrical vessel partially filled with mercury. The quantity of mercury is so chosen that by its expansion combined with that of the iron rod and cylinder the centre of gravity of the whole does not move when the temperature is varied.

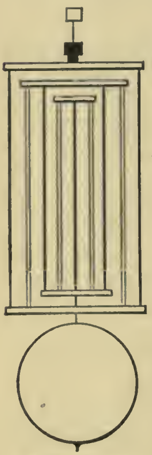


FIG. 113. — Compensation "gridiron" pendulum. The single black lines represent iron rods; the double lines brass ones.

Again, the period of a watch or of the ordinary spring clock is regulated by the vibrations of the balance wheel. This consists of a small wheel attached to a flat coiled spring, which coils and uncoils, making harmonic vibrations. If the temperature rises, two changes occur: the elasticity of the spring is decreased, and the wheel expands. Owing to both these effects, the period of the spring would be increased were it not for the peculiar construction of the rim of the wheel. As is shown in the cut, this consists of two (sometimes three) parts or sections. One end of each is fastened to a spoke, but the other is free. Near this latter end are screws or movable weights. Each section of the rim is double, consisting of iron on the inside and brass on the outside. Consequently, owing to the greater expansion of the brass, when the temperature is raised each section of the rim is made more curved, and the weights are brought in nearer the centre. This decreases the moment of inertia, and thus tends to decrease the period of vibration. Therefore, by suitably adjusting the weights or screws, the balance wheel may be made to retain a constant period, however the temperature changes.

FIG. 114. — Balance wheel of a watch; the inner portion of the rim is iron; the outer is brass.

Quite recently Guillaume has discovered that an alloy of nickel and steel in the proportion of 36 parts of nickel to 64 parts of steel has a coefficient of linear expansion equal to 0.00000087; and so its expansion may be neglected in all ordinary apparatus or work. The change in length of this alloy with change in temperature takes place very slowly if the rise in temperature is small; for a rod made of it does not reach its full expansion for one or two months.

Liquids

Apparent and Absolute Expansion. — In measuring the change in volume of a liquid, owing to any cause, a difficulty is met which has been referred to before; namely, the fact that the liquid must be contained in a solid vessel, and whatever affects the volume of the former will also, in general, affect that of the latter. Thus, when the containing vessel is a bulb with a tube attached, if the liquid fills the bulb and part of the tube, it is observed that, when the bulb is heated suddenly by immersing it in a basin of hot water or in any other way, the top of the column of liquid in the tube immediately sinks and then rises gradually, ascending finally higher than it was originally. This is owing to the fact that the first effect of the application of the hot bath is to raise the temperature of the bulb, and it therefore expands before the liquid inside is affected; as soon, however, as its temperature is raised, it expands and rises in the tube. If the bulb is chilled, instead of heated, the reverse of these changes takes place; the liquid first rises in the tube and then sinks, falling below its original position. It is evident, then, that the apparent change in volume of the liquid is less than the real change by an amount equal to the change in volume of the containing solid. So, if the coefficient of expansion of the solid is known, the direct method for determining the change in volume of a liquid that accompanies a change in temperature is to inclose the liquid in a bulb with a finely divided stem whose volumes are known, and to measure the volume of the liquid at any one temperature and the apparent change in volume when the temperature is altered. Let v_1 be the initial volume at the temperature $t_1°$, and v the *apparent* increase in volume when the temperature is raised to $t_2°$; further, let the coefficient of cubical expansion of the solid have the value b. Then the increase in volume of the solid is $v_1[1 + b(t_2 - t_1)]$; and hence the true change in volume of the liquid is $v + v_1[1 + b(t_2 - t_1)]$. Experiments show that for nearly all liquids — water is an exception, as will be explained below — the relation between the change in volume and that in temperature is of the same form as for solids, viz., $v_2 - v_1 = v_1 b(t_2 - t_1)$, where b is the coefficient of cubical expansion of the liquid, referred to $t_1°$. If the initial temperature is 0° C., this becomes, as before, $v = v_0(1 + b_0 t)$; and it is to be noted that $b = \frac{b_0}{1 + b_0 t}$, and therefore only if b_0 is small can it replace b.

The coefficient of expansion is found to be different for different liquids, as is shown in the following table. It should be noted that the expansion of liquids is, in general, much greater than that of solids.

Ethyl alcohol . . .	between	0° and 80° C.	0.00104
Ethyl ether . . .	between	– 15° and 38° C.	0.00215
Glycerine . . .			0.000534
Mercury . . .	between	0° and 100° C.	0.000182
Turpentine . . .	between	– 9° and 106° C.	0.00105

It is evident that, if the coefficient of expansion of a liquid is known, the same series of observations and measurements gives a method for the determination of the coefficient of cubical expansion of the solid that contains the liquid. This method therefore can be used for all solids that can be formed into bulbs, *e.g.* glass or quartz. Mercury is the liquid that is, in general, used, because its expansion is known, and for other obvious reasons.

Measurement of Coefficient of Expansion. — A better method, however, for the determination of the coefficient of expansion of a liquid, which does not involve a knowledge of that of the solid vessel, was devised by Dulong and Petit and improved by Regnault. It depends upon the fact that the heights at which two liquids of different densities stand when they balance each other in a U-tube is independent of the material of the tube. If h_1 is the height of the column of one liquid of density d_1, and h_2 that of the column of the liquid whose density is d_2, $\frac{h_1}{h_2} = \frac{d_2}{d_1}$ (see page 176). But the formula for expansion,

$$v = v_0(1 + b_0 t),$$

can be replaced by

$$d_0 = d(1 + b_0 t),$$

if d is the density at $t°$, and d_0 that at 0°; for the density of any body whose mass remains constant varies inversely as its volume ($m = dv$). Therefore, if the ratio of the densities of a liquid at 0° and $t°$ is known, its coefficient of expansion can be deduced at once.

The simplest form of the actual experiment is as follows: A tube is made in the form of a W, with a cross tube at the top, as shown in the cut; this last has a small opening A near its middle point and is kept horizontal; there is a branch tube joined to the apparatus at B, which is connected with a reservoir K containing air that can be compressed to a pressure greater than that of the atmosphere; a quantity of the liquid whose expansion is to be studied is poured in, just sufficient to flow out of the opening A in the upper cross tube, thus insuring the condition that the upper level surfaces of the two columns are at the same height. These two separate portions of the liquid are then surrounded by baths, one at 0°, the other at t°. When equilibrium is established, the difference in level of the two surfaces of one portion of the liquid is not the same as that of the other. Call this difference for the liquid at 0°, h_0, and that for the one at t°, h. Then $d_0gh_0 = dgh$, or $\frac{d_0}{d} = \frac{h}{h_0}$. Consequently, if h_0 and h are measured, b_0 may be calculated; for, $\frac{d_0}{d} = 1 + b_0t$. $\therefore 1 + b_0t = \frac{h}{h_0}$, or $b_0 = \frac{h - h_0}{th_0}$.

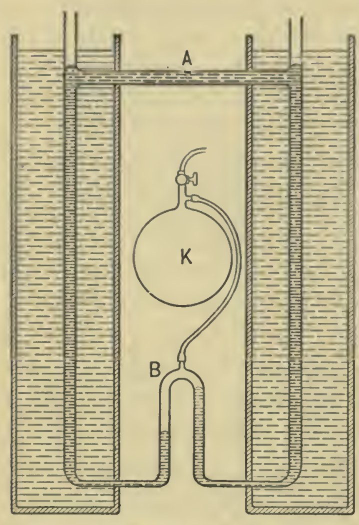


FIG. 115.—Apparatus for the determination of the coefficient of expansion of a liquid.

Expansion of Water. — When water is studied, it is found that as the temperature rises from 0° C. to 4° C., it contracts; but above 4° it expands. These changes in volume of water (or of any substance) which accompany changes in temperature may be best shown graphically. Lay off two axes, one to indicate temperatures; the other, volumes of a definite quantity of the substance. For water the curve giving the connection between v and t is as shown in the cut. (If the coefficient of expansion is a constant, the curve is a straight line sloping upward from left to right.)

FIG. 116. — Curve showing expansion of water as the temperature is changed.

This fact, that the density of water is a maximum at 4° C., and that it decreases continuously from this temperature to 0° and to 100° is of great importance in nature; for as the temperature of the water in a lake or a pond sinks in winter below 4° and reaches 0°, the water at 0° floats on top and does not sink to the bottom, as would be the case with other liquids; and consequently ice forms on the upper surface, not at the bottom.

This peculiarity of water is explained by assuming that when ice melts the liquid formed is not at first made up of molecules that are all alike, but contains two kinds, "ice molecules" and "water molecules." The former are assumed to be lighter than the latter; and, as the temperature rises, they are transformed gradually into the latter.

Gases

Expansion at Constant Pressure. — When the temperature of a gas is increased, its volume increases also, unless special precautions are taken; but, if the external pressure varies during the change in temperature, the change in volume is modified. We shall then assume that during the change the pressure is maintained constant. A gas is always inclosed in some solid vessel; and its expansion must always be taken into account although its effect is small in comparison with the expansion of the gas. When this is done, experiments show that the general formula for expansion is true, viz.,

$$v = v_0(1 + b_0t).$$

This coefficient of expansion is constant for any one gas; and, further, it is found by experiments to be practically the *same for all gases*. This is known as the "Law of Gay-Lussac." The value of this constant is very nearly 0.0036600, or $\frac{1}{273}$, using the Centigrade scale. It is to be noted, then, that gases expand more than liquids. (The apparatus used by Gay-Lussac in his investigation on the expansions of gases is shown in the cut.)

General Law for a Gas. — Having thus determined the relation between the volume and temperature of a definite quantity of gas when the pressure is kept constant, and

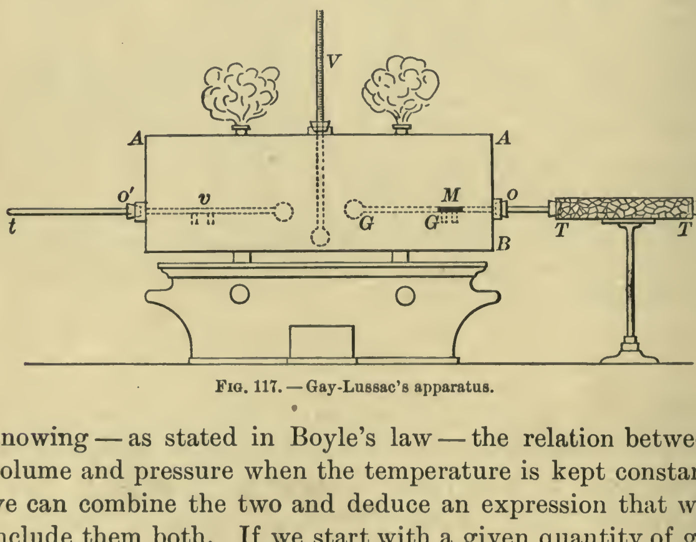


Fig. 117. — Gay-Lussac's apparatus.

knowing — as stated in Boyle's law — the relation between volume and pressure when the temperature is kept constant, we can combine the two and deduce an expression that will include them both. If we start with a given quantity of gas at 0° and at a pressure p, it will have a definite volume which we can call v_0. If now the temperature is raised to $t°$, the pressure being kept constant, the resulting volume $v = v_0(1 + b_0t)$. If the pressure is then changed to P, the temperature being kept constant, the volume will be altered and the new value V must be such that $PV = pv = pv_0(1 + b_0t)$. Thus, if the volume of a gas is measured at a temperature $t°$

and pressure P, the volume which this same quantity of gas would occupy if its temperature were made 0° and its pressure p can be calculated; for

$$v_0 = \frac{P}{p}\frac{V}{1 + b_0 t}.$$

(The "standard conditions" for a gas are a temperature of 0° C., and a pressure of 76 cm. of mercury; and when the "density of a gas" is referred to, it is always implied that the gas is in its standard condition, unless the contrary is stated.) This formula can be expressed in a simpler form, viz.,

$$\frac{PV}{1 + b_0 t} = pv_0,$$

or

$$\frac{PV}{\frac{1}{b_0} + t} = b_0 pv_0.$$

But v_0 is the volume of a given quantity of the gas at 0° and at pressure p, consequently the product pv_0 is a constant for this gas by Boyle's law. Therefore, $\frac{PV}{\frac{1}{b_0} + t}$ is a constant for any one gas, but has a different value for different gases. This may be expressed $\frac{PV}{\frac{1}{b_0} + t} = R$, a characteristic constant for any one gas. The quantity $\frac{1}{b_0} + t$, which, using the Centigrade scale, approximately equals $273 + t$, is called the "absolute gas temperature" on the Centigrade scale, and is written T. It is evident that, if different quantities of the same gas at the same temperature and pressure are taken, their volumes will vary directly as the masses. Thus, writing M for the mass of the gas, the complete law is

$$\frac{PV}{T} = RM.$$

The constant R can be found for any gas by measuring its density at a known temperature and pressure. Thus, if $t°$ C. is the temperature, p the pressure, and d the density, $R = \frac{p}{d(t + 273)}$. The value of R for hydrogen is, 4.14×10^7; for oxygen, 0.259×10^7; for nitrogen, 0.296×10^7.

If the same law were to hold for a gas as the pressure became smaller and smaller and finally vanished without the volume at the same time being made infinitely great, the temperature of this condition would be given by $T = 0$, or $t°\ C = -\frac{1}{b_0} = -273°$ C. approximately. Similarly, if the same law could be applied to a gas as its volume was made less and less and finally vanished, the pressure remaining finite, the temperature would be given by the same value, viz., $-273°$ C. approximately. Of course the above law for a gas does not apply to one if greatly compressed, and the pressure of a gas can become zero only by its volume being infinite; and so the above deductions have no physical meaning. We shall see later, however, by other reasoning, that this temperature $-\frac{1}{b_0}$, or $-273°$ C, marks the lowest temperature which any body in our universe can reach. For that reason it is called the "absolute zero." (See page 308.)

Change of Pressure at Constant Volume. — Another fact is apparent from this formula. If the changes take place in such a manner that the initial and final volumes are the same, $V = v_0$, and therefore $P = p(1 + b_0 t)$. Therefore the pressure increases at the same rate with increase of temperature when the volume is kept constant, as does the volume when the pressure is kept constant. (Actually, this is not exactly true of any gas; but this discrepancy is a consequence of the fact that Boyle's law is not exact for an actual gas nor is the coefficient of expansion a constant.)

Laws of Gay-Lussac and Charles. — The fact that all gases have approximately the same coefficient for change of pressure when the volume is kept constant was discovered by the French physicist Charles; while the corresponding one that all gases expand alike when the pressure is kept constant was discovered a few years later, in 1802, by Gay-Lussac. These statements of fact are therefore called Charles's and Gay-Lussac's laws.

Other Forms of the Gas Law. — This formula for a gas, $PV = RMT$, may be expressed differently and more simply if we assume the truth of Avogadro's hypothesis (see page 200). If m is the mass of each molecule, and N is the number in a unit volume, $M = mNV$; consequently, on substituting in the formula, we have

$$P = RmNT,$$

or $N = \frac{P}{RmT}$. But by Avogadro's hypothesis, if the pressure and temperature of two gases are the same, they have the same number of molecules in each unit volume; that is, if P and T are the same for two gases, N is also. Therefore, from the above formula, Rm must be the same for all gases; it is a constant of nature. Calling its value R_0, the formula becomes $P = R_0NT$, which states that the pressure in a gas (or a mixture of gases) varies directly as the number of molecules in a unit volume and as the absolute temperature.

This equation, in turn, may be expressed in a form which is more useful for practical purposes, because, of course, we have no accurate knowledge of the value of N. The "molecular weight" of a gas has been defined to be a number, characteristic of the gas, which is proportional to the mass of one of its molecules, and which is so chosen that the number for oxygen is 32. Thus if w is the *molecular weight* of a gas, each of whose molecules has the mass m, $w = cm$, where c is a constant for all gases and has such a numerical value as to make w equal 32 for oxygen. A number of grams of a substance equal to its molecular weight (*e.g.* 32 grams of oxygen) is called "one gram-molecule" or "one mol" of that substance. Thus, if there are N' mols of a gas in a unit volume, $V = 1$ and $M = N'w = N'cm$ in the general formula; hence, $P = RmcN'T = R_0cN'T$.

The product R_0c is a constant, the same for all gases; call it R'.

Hence $$P = R'N'T,$$

or, the pressure varies directly as the number of mols in a unit volume. The value of R' may be found by experiments on the densities of gases, because P is the pressure of a gas at temperature T when the density is such that there are N' mols in a unit volume. It is found to be 8.28×10^7 on the C. G. S. system.

Energy Relations during Expansion

Mechanical Expansion. — When the dimensions of a body are increased, either by the addition of heat energy or by mechanical forces, there is a certain amount of external work done and there are alterations in the kinetic and potential energies of the molecules. In the case of the thermal effect this is evident from what has been already said; change in relative position of the molecules, change in temperature, and external work occur together. Similarly, when a brass wire is stretched, its temperature falls; when a gas is compressed, its temperature is raised, etc.

We might have predicted that these changes in temperature would be as just stated. (See page 104.) If the temperature of a brass wire should rise when it is stretched, it would be in unstable equilibrium. If the wire hangs vertically under the stretching force of a heavy body, and if a sudden downward blow is given this weight, it will move down, stretching the wire still more, then come to rest and move up, etc., making harmonic vibrations, showing that it was in stable equilibrium. But if, owing to this stretching, as the hanging body moves down, the temperature of the wire were to rise, it would lengthen; and this increase in length would cause another rise in temperature, etc., so the equilibrium would be unstable. Therefore, only if the wire cools when stretched, is the equilibrium stable. The general law is that, if a body expands when its temperature is raised, its expansion by mechanical means will cool it, while mechanical compression will raise its temperature. The converse is true of those bodies which contract when their temperature is raised, *e.g.* water between 0° and 4° C.

As the sun radiates energy, it contracts; and owing to this cause its parts are slowly coming closer together, and therefore energy is being liberated to make up for that loss. Thus, in ordinary language, the sun owes its heat to its slow contraction. Undoubtedly also, meteoric pieces of matter are falling into the sun, and their energy is thus also added to that of the sun.

If the expanding body is a liquid or a gas, or if it is a solid immersed in a fluid, the amount of external work done equals the product of the increase in volume by the pressure. (See page 159.) The internal changes consist of change in kinetic energy and change in potential energy. The first of these is connected intimately with changes in temperature, as has been already shown. (See pages 197 and 220.) Changes in potential energy occur if there are molecular forces. The fact that these exist in solids and liquids and are large is evident from the obvious properties of these forms of matter, *e.g.* they retain definite volumes; but nothing is known in regard to the amount or cause of these forces. The case is different with a gas, for in it the forces are extremely small. (Nothing, however, can be said as to their cause.)

Internal Work in a Gas. — This fact that the internal forces in a gas are extremely small was first shown by Gay-Lussac and later by Joule, working independently and also in collaboration with Thomson (now Lord Kelvin). The early experiments of Joule are perhaps the simplest to consider. His apparatus consisted of two strong metal cylinders connected by a tube in which was a stopcock. In one of these cylinders quantities of the gas to be studied were compressed until the pressure was as high as the apparatus would permit; while from the other the gas was exhausted. The whole apparatus was then submerged in a tank containing water, and this was stirred until the temperature came to a steady state. Then the stopcock in the tube connecting the two cylinders was opened; and, when the gas had redistributed itself, occupying a greater volume and thus coming to a smaller pressure, but *doing no external work*, the temperature of the water in the tank after being well stirred was again observed. In no case was there any measurable change. If there had been molecular forces of *attraction* between the molecules, it would have required work to separate them when the gas increased in volume, this energy would have necessarily come from the molecules themselves, which would have thus lost kinetic energy; and therefore the temperature of the gas would have fallen. If there had been molecular forces of *repulsion* between the molecules, their potential energy would have been decreased by the increase in volume, their kinetic energy would then have increased an equal amount; and therefore the temperature of the gas would have risen. Either of these changes in the temperature of the gas would have affected the temperature of the surrounding water; and this would have been observed by the thermometer, if it were sufficiently delicate. Consequently Joule's experiments prove that to the degree of sensitiveness of his thermometer, there are no forces between the molecules of a gas; or, in other words, the internal energy of a gas is entirely kinetic.

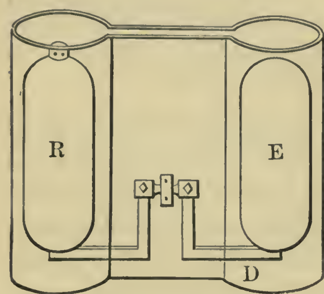


FIG. 118. — Joule's apparatus for studying the free expansion of a gas: two cylinders, *R* and *E*, are connected by a tube in which there is a valve; the whole is immersed in a tank containing water.

The later experiments of Joule and Thomson show, however, that gases do have measurable molecular forces, although they are extremely small, and that these in general are forces of attraction. (In hydrogen, at ordinary temperatures, the forces are repulsive; but at very low temperatures they are attractive.)

Expansion of a Gas when External Work is Done. — Joule modified his experiment by inclosing the two cylinders in separate tanks of water; and, when the expansion took place, he noted the temperatures of the water in the two tanks. He observed that the temperature of the one holding the cylinder in which there was the high pressure fell, while that of the other rose. The explanation is evident: the gas that stays behind in the high pressure cylinder after the expansion, has done work in giving kinetic energy to the escaping gas *as a whole*, *i.e.* in producing a *wind*, consequently its temperature falls. This blast of gas entering the other cylinder soon ceases, owing to viscosity, and the kinetic energy of the moving gas passes into the energy of the molecules, and is apparent by a rise of temperature. In other words, the gas in the first cylinder expands, doing work, and so its temperature falls; work is done upon the other gas in setting it flowing, and so its temperature rises after the motion is stopped by friction.

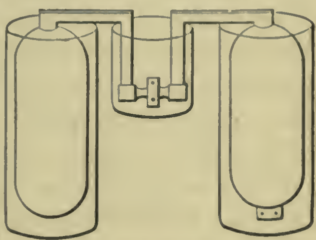


FIG. 119. — Joule's second experiment.

The fact that when a gas expands under such conditions as to do work its temperature falls is shown by allowing damp air inclosed in a vessel to expand suddenly. If there are nuclei in the air, drops of water will be condensed around them, thus forming a visible mist and showing that the air has been chilled. (See page 186.) This is the explanation in many cases of the formation of clouds in the air.

Expansion of a Gas in General. — When a compressed gas is allowed to expand out through a fine opening into a space where its pressure is less, changes in temperature take place owing to numerous causes. We shall consider two cases of practical importance. Let a cloth with fine meshes — a piece of cheese cloth or toweling — be folded over the nozzle from which the gas is escaping; it will expand with violence through these numerous openings into the air, thus forming a wind; the gas that does not have time to escape during any small interval thus does work on that portion which it blows out and so experiences a fall of temperature itself. This may be sufficient to liquefy or even solidify it; and the cloth will in this last case be found to contain the solidified gas. (This is the ordinary way of obtaining carbon dioxide in the solid form.) Again, let the opening or openings through which the compressed gas escapes be so fine and tortuous that the outcoming gas has no kinetic energy as a whole, *i.e.* there is no wind, it does not *flow;* in this case its temperature is lower than when in the compressed condition (except with hydrogen at ordinary temperatures), as shown by Thomson and Joule, owing to the fact that there are minute molecular forces of attraction, and as the potential energy of the expanded gas is increased, its kinetic energy, and therefore temperature, must be diminished. This fall in temperature varies directly as the difference in pressure of the gas in its two conditions, and so may be considerable. In any ordinary expansion of a gas from a small opening, both of these actions take place. The gas that is escaping at any instant has done work in pushing out the portion that was just before it, and so its temperature is lowered, quite apart from the influences of the molecular forces. (Some distance away from the opening, however, in the case of a jet or a blast, the temperature of the gas is increased, owing to the friction of the moving currents.) This method of securing a decrease in temperature by allowing a gas to expand through a fine nozzle is being used practically in recent machines for the liquefaction of gases. (See page 280.)

CHAPTER XII

CHANGES IN TEMPERATURE

Energy Relations when the Temperature is Raised. — It is by the change in temperature of bodies when exposed to some "source of heat" that our attention is directed to heat phenomena; and it is in terms of changes of temperature that heat energy is measured, as we have already seen in the definition of the "calorie." If, however, we wish to determine how much energy goes to producing the rise in temperature, it is necessary to ascertain how much is used in doing external work, and how much in overcoming the molecular forces. If there is a uniform pressure p over the body, and if its volume increases from v_1 to v_2, the external work done against this pressure is $p\ (v_2 - v_1)$, and so may be calculated. It is only, however, in the case of a gas that anything is known quantitatively in regard to the molecular forces. Therefore, when the temperature of a solid or a liquid is raised, although the external work may be calculated if the external pressure is known, it is impossible to separate into its parts the energy that is spent in internal work; and so we do not know how much goes to producing rise in temperature. But the case of a gas is different. We can regulate the amount of external work done as well as measure it, and we know further that the molecular forces are so small that they can be neglected. Consequently when the temperature of a gas is raised owing to the addition of heat energy, if we deduct the amount that is used in doing external work, the balance is all spent in raising the temperature, *i.e.* in increasing the kinetic energy of the molecules. The amount of energy that must be added to the gas, then, in order to raise its temperature a definite number of degrees, depends upon the amount of external work done by the gas; that is, upon the external conditions under which the gas is kept during the change in temperature. If this change is the same in different experiments, the energy that is spent in producing the increase in the kinetic energy of the molecules is the same whatever the external conditions; and so the difference in the amounts of heat energy added must equal the difference in the amounts of the external work done.

Special Case of a Gas. — For various reasons, partly practical and partly theoretical, the change in temperature of a gas is considered, as a rule, under two different conditions: (1), when the volume is kept constant, and so no external work is done; (2), when the pressure is maintained constant, so the external work equals the product of the values of this pressure and the change in volume.

If Q_1 is the heat energy added in one experiment when the volume is kept constant, Q_2 that in another when the pressure is kept constant at a value p while the volume is increased from v_1 to v_2, the change in temperature being the same in both,

$$Q_2 - Q_1 = p(v_2 - v_1),$$

where, of course, Q_2 and Q_1 are measured in mechanical units; *i.e.* in ergs or joules, since the product $p(v_2 - v_1)$ is so expressed.

Reverse Changes. — When a body cools, it gives out heat energy; but the external work done on it by the external forces is not necessarily the same as is done by the body against these forces when the body is heated and rises through the same range of temperature, because the external forces may have changed, or they may vary differently during the two processes. If, however, the pressure remains unchanged, or, when it is variable, if the series of changes is exactly reversed during the two processes, the external work done by the body during the rise in temperature equals that done on the body during the opposite change. If this is the case, the heat energy added in the former process must equal that given out during the latter; for at the end of the two changes the body is back in its original condition, and so there is no change in its internal energy, and there has also been no gain or loss of energy owing to external work. This fact serves as the basis of all methods for the measurement of quantities of heat energy.

Many simple experiments show that, to produce the same change in temperature in equal quantities of different bodies, requires different amounts of heat energy. One of these, due to Tyndall, is to raise to the same temperature in a bath of heated oil several spheres, made of different materials, but of the same weight and of the same size (naturally, some are hollow), and then to place them upon a cake of wax that melts at a low temperature. As the spheres cool down they melt different quantities of the wax, as is shown by their entering the wax to different distances; and at the end of the process they all have the same temperature. This proves that in cooling over the same range of temperature different bodies give out different quantities of heat energy. (This experiment does not prove this fact beyond question, because it is complicated by differences in the conductivity of the different bodies and by the consequent fact that some give out energy more quickly than others, and so more of the wax is melted before it can conduct the heat energy away. There are other actions, too; but in reality the main effect is that described above, and so the experiment is a proper one to illustrate the question referred to.)

Measurement of Heat Energy. — There are three methods in general use for the measurement of the heat energy required to raise the temperature of bodies, or of that given out when a body cools: (1) To put the body in a bath of water at a different temperature, and measure the change in temperature of the water when equilibrium is reached. If the mass of the water is m grams, and its change of temperature is $t°$ C., mt calories enter or leave the water, if we neglect the slight variations in the value of the calorie at different temperatures. (2) To put the body in a bath of ice and water, so that a certain amount of ice is melted. It will be shown later that it requires 80 calories, very approximately, to melt 1 g. of ice at 0° C.; and so, if m grams are melted, the ice must receive $80\,m$ calories. (3) To put the body in a bath of steam, so that a certain amount of this is condensed into water by contact with the cooler body. It will be shown later that 536 calories, approximately, are given out by 1 g. of steam when it condenses into water at 100° C. So, if m grams are condensed by the body at this temperature, it must receive $536\,m$ calories.

"Specific Heat." — The number of calories that must be added to a body whose mass is 1 g. in order to raise its temperature through one degree Centigrade, from $t°$ to $(t+1)°$, is called the "specific heat" of that body at $t°$ C.

It should be noted that the specific heat of a substance is a number that is independent, in a way, of the heat unit used or of the temperature scale adopted. We could adopt as a heat unit that quantity of heat energy required to raise the temperature of a unit mass (on any system) of water from $t°$ to $(t+1)°$ (on any scale); and define the specific heat of a substance at $t°$ as the number of these heat units that is required to raise the temperature of a unit mass of it from $t°$ to $(t+1)°$. Or, if we assume that the specific heat is the same at all temperatures, we may define the specific heat of a body as the ratio of the amount of heat energy required to raise the temperature of the body through any range of temperature to that required to raise the temperature of an equal mass of water through the same range of temperature.

With solids and liquids, as has been explained, we assume that the pressure is that of the atmosphere; but with gases we distinguish two special conditions, constant pressure or constant volume, and so have two corresponding specific heats. For most bodies the specific heat is practically constant for all temperatures that are far removed from the melting points; and so, if c is the specific heat of a body whose mass is m, and if the temperature is raised through $t_2 - t_1$ degrees, the number of calories added is $mc(t_2 - t_1)$.

In iron, boron, carbon, and a few other substances the specific heat varies to a marked degree at different temperatures; and with them c in the above expression is the "mean specific heat" for the range from t_1 to t_2.

Measurement of Specific Heat. — Corresponding to the three methods of measurement of heat energy referred to above there are, then, three methods for the measurement of the specific heat of a given substance. We assume in each case that there is no loss of heat by radiation, etc.

1. *Method of Mixtures.*

If
m_1 = mass of body,
m_2 = mass of water,
t_1 = original temperature of the body,
t_2 = original temperature of the water,
t_3 = final temperature of equilibrium,

$$m_1c(t_1 - t_3) = m_2(t_3 - t_2),$$

or

$$c = \frac{m_2(t_3 - t_2)}{m_1(t_1 - t_3)}.$$

The water is contained in some solid vessel, called a "calorimeter"; and its temperature is affected by the changes in that of the water. If m_3 is its mass and c' its specific heat, and if we assume that the changes in its temperature are the same as those of the water, the number of calories it receives is $m_3c'(t_3 - t_2)$; and, since this energy comes from the foreign body introduced into the water, it is seen that, in the above formula for c, m_2 must be increased by m_3c'. This quantity is called the "water equivalent" of the calorimeter.

This method was first used by Joseph Black about 1760; and it is the one most generally used at the present time. There are many objections to it, however. Chief among these are the difficulties of determining the water equivalent of the calorimeter and of avoiding losses by radiation, etc. Most of these are overcome in a modification of the apparatus due to Professor Waterman of Smith College.

2. *Method of Melting Ice.*

If m_1 = mass of body,

m_2 = mass of ice melted,

t = initial temperature of the body,

$$m_1ct = 80\,m_2.$$

This method was also first used by Joseph Black. The obvious difficulty in it is the measurement of the quantity of ice melted. This has been overcome most successfully in a form of apparatus due to Bunsen; but great skill is required to use it properly.

3. *Method of Condensation of Steam.*

If m_1 = mass of the body,

m_2 = mass of steam condensed,

t_1 = initial temperature of the body,

t_2 = final temperature of the body, which is never far from 100° C.,

$$m_1c(t_2 - t_1) = 536\,m_2.$$

With this method there is a correction for the water equivalent of the calorimeter; and one of the chief difficulties is to measure accurately the quantity of steam condensed. The apparatus used was invented by Professor Joly of Dublin.

Specific Heats of a Gas. — Any one of these methods can be used to measure the specific heat of a solid or a liquid; but with gases a difficulty enters owing to their small density — the correction for the calorimeter is the larger part of the heat energy. To measure the specific heat of a gas at constant volume, the third method may be used if the gas is compressed into a hollow sphere; but this is not very satisfactory. To measure the specific heat at a constant pressure the best method is to pass a large quantity slowly through a spiral tube which is immersed in a bath at a high temperature, then through another spiral surrounded by water which is cooler and whose temperature is thus raised, and finally out into the air or into some large reservoir. If the gas is forced through very slowly, its pressure remains practically constant.

There are two indirect methods by which the specific heat of a gas at constant volume may be determined: one depends upon a knowledge of the ratio of this to the specific heat at constant pressure; the other, upon a knowledge of the difference between these two quantities.

It may be proved by higher mathematics that the ratio of the specific heat at constant pressure to that at constant volume equals the ratio of the adiabatic coefficient of elasticity to the one at constant temperature (see page 194). But this last ratio is a constant for any one gas, which may be determined by measuring the velocity of compressional waves in this gas, as will be proved in the next section of this book. (See page 337.) It will be shown there that this velocity is given by the following formula,

$$V = \sqrt{\frac{cp}{d}},$$

where p is the pressure of the gas, d is its density, and c is the ratio of the two elasticities, and therefore of the two specific heats. It is not difficult, then, to determine c, since v, p, and d can all be measured; and, if C_p and C_v are the two specific heats, one at constant pressure the other at constant volume, $C_p = cC_v$; so, if c and C_p are known, C_v can be calculated.

The difference $C_p - C_v$ expressed in mechanical units, equals, from what has been said on page 248, the amount of external work done when 1 g. of the gas expands at constant pressure p, the temperature rising through 1° C. This work equals $p(v_2 - v_1)$, where v_2 and v_1 are the volumes of the gas at two temperatures differing by 1° C. We may calculate this by using the gas formula $\frac{pv}{T} = Rm$. In this case $m = 1$, hence $pv_1 = RT_1$ and $pv_2 = RT_2$, where $T_2 - T_1 = 1$. Therefore $p(v_2 - v_1) = R$. So that $C_p - C_v = R$. If C_p and C_v are measured in calories, and if J is the value of a calorie in ergs, $J(C_p - C_v) = R$. The value of R may be found by ordinary methods of gas measurement, as has been shown; and, therefore, if the value at C_p is known, that of C_v may be calculated.

Ratio of the Specific Heats. — We may obtain a better physical conception of c, the ratio of the two specific heats, by interpreting it in terms of the kinetic theory of gases; *i.e.* by considering the properties of a gas as identical with those of a set of elastic spheres.

The internal energy of a gas may be regarded as entirely kinetic; but this energy is not entirely energy of motion of the molecules themselves. The phenomena of radiation show that as the temperature of a gas is raised, the energy of motion of the parts of the molecules is increased. Let us assume, following Clausius, that the entire kinetic energy of a molecule — that is, the energy of its parts and its own energy as a whole — is proportional to its energy of translation. We can write it then $\frac{1}{2} mbV^2$, where m is the mass of the molecule, V its velocity of translation, and b a factor, which has the value 1 if there is no motion inside the molecule, and otherwise is greater than 1. If N is the number of molecules in a unit volume of the gas, the energy of the molecules in this volume is then $\frac{1}{2} mbV^2N$. But mN is the mass of the molecules in this volume; so the internal energy of a unit mass of the gas is $\frac{1}{2} bV^2$. This can be expressed in terms of the temperature; for the pressure of a gas is given by the formula $p = dRT$, if d is the density; and on the kinetic theory $p = \frac{1}{3} dV^2$. Therefore $V^2 = 3 RT$; and the internal energy per unit mass is $\frac{3}{2} bRT$. So, if the temperature is raised one degree, this energy is increased by an amount $\frac{3}{2} bR$. This quantity, then, is the specific heat of the gas at constant volume expressed in mechanical units; or, in symbols, $C_v = \frac{3}{2} bR$. But $C_p - C_v = R$; hence $C_p = C_v + R = R(\frac{3}{2} b + 1)$. Hence the ratio of the specific heats

$$c = \frac{C_p}{C_v} = \frac{R(\frac{3}{2} b + 1)}{\frac{3}{2} bR} = \frac{\frac{3}{2} b + 1}{\frac{3}{2} b} = 1 + \frac{2}{3b}.$$

The least value of b is 1; and therefore the greatest possible value of c is $1 + \frac{2}{3}$ or 1.67. For all other values of b, c is less than this.

It is a most striking fact that for certain gases, viz., mercury vapor, argon, helium, and a few others, the value of c found by experiments on the velocity of waves in them is 1.67, while for all other gases it is less than this, being about 1.41 for air, hydrogen, and oxygen, 1.26 for carbonic acid gas, etc. Those gases for which c equals 1.67 are called by chemists "monatomic"; and, whatever value may be attached to the above assumptions, it is certain that a large value of c indicates an extremely simple construction of the molecule or a molecule whose internal energy is small; while a small value of c indicates the contrary.

A few values of specific heats are given in the following table:

AVERAGE SPECIFIC HEATS

Alcohol	0°–40° C.	0.597	Mercury	20°–50° C.	0.0333
Aluminium	0°–100° C.	0.2185	Paraffin		0.693
Brass		0.09	Platinum	0°–100° C.	0.0323
Copper	0°–100° C.	0.0933	Silver	0°–100° C.	0.0559
Glass (crown)		0.161	Tin	0°–100° C.	0.0559
Iron	0°–100° C.	0.113	Turpentine		0.467
Lead	0°–100° C.	0.031	Water		1.00

GASES

	C_p	C_v	RATIO
Air	0.237	0.171	1.40+
Argon			1.66
Chlorine	0.121		1.32
Carbon dioxide	0.202	0.173	1.30

GASES — *Continued*

	C_p	C_v	RATIO
Helium			1.66
Hydrogen	3.40	2.40	1.41
Mercury (vapor)			1.67
Nitrogen	0.244		1.41
Oxygen	0.217		1.41

Law of Dulong and Petit. — When the values of the specific heats of a great many substances are compared, a connection becomes evident between them and the "atomic weights" of the substances. (For an explanation of this last quantity some book on chemistry should be consulted.) This was first noted by Dulong and Petit. It is found that the product of the value of the atomic weight of any solid substance and that of its specific heat is a quantity that is approximately the same for all substances, viz., 6.4, using the ordinary system of units; while the same constant for gases is 3.4. This means that the same amount of energy is required to raise the temperature of an atom, whatever solid substance it belongs to. This product is called the "atomic heat." Naturally, this law of Dulong and Petit is only approximately true; for, as has been said, the specific heat of a substance varies with the temperature, and it is impossible to know when different substances are at temperatures such that their conditions are comparable.

CHAPTER XIII

CHANGE OF STATE

Introductory. — The fact that heat energy enters or leaves a body when it changes its state is familiar to every one. In order to melt ice or boil water it must be exposed to some source of heat; if water evaporates from one's hand or from the surface of a porous jar, the latter is chilled, showing that heat energy has been taken from it; as steam condenses into water in steam coils or "radiators," they are heated, showing they have received energy; when an acid or salt is dissolved in water, its temperature is changed; when water freezes or dew is formed, the temperature of the surrounding air is raised slightly, etc. During these changes of state, not alone are there heat changes, but alteration in volume, and so external work is done; and we shall see that the external conditions are of fundamental importance. We shall consider in detail a few of the most important cases of change of state; viz., fusion, evaporation, sublimation, solution, and chemical changes.

Fusion

Freezing and Melting Point. — "Fusion" is the name given the process in which a solid body melts and becomes liquid; the reverse process is called "solidification." If a solid body that can form crystals, *e.g.* ice, is exposed to a source of heat, its temperature will rise until a point is reached when it begins to melt; and then, so long as there is any solid to melt, the temperature of the mixture of the solid and its liquid remains unchanged; but when the solid is entirely melted, the temperature again rises. Conversely, if the liquid thus formed is placed in such a condition as to lose heat energy, its temperature will fall until a point is reached at which some of the liquid solidifies; then the temperature remains unchanged until the liquid entirely changes into the solid form; and from then on the temperature falls again. This temperature which marks the transition from liquid to solid is the same as that which marks the reverse change. It is called the "melting point," or the "freezing point." If the solid and its liquid exist together in contact, and no energy is added or taken away from them, they will remain in equilibrium; so the melting point may be described also as that temperature at which the solid and its liquid can exist together in equilibrium.

Effect of Variations in the Pressure. — Experiments show that this temperature of equilibrium depends upon the external pressure, varying as it is changed. Thus, if ice and water are in equilibrium together, under ordinary atmospheric pressure and in a region where the temperature is therefore 0° C., an increase in pressure will cause some of the ice to melt, showing that the melting point is lowered and that heat energy flows into the ice from the surrounding region; a decrease in pressure will cause some of the water to freeze, showing that the melting point has been raised and heat energy flows out from the water into the surrounding region. The explanation of this variation of the melting point with the external pressure depends upon the fact that when a solid melts its volume changes, in some cases increasing, in others decreasing. If ice melts, its size decreases; that is, the volume of a certain mass of water in the solid form is greater than in the liquid. (Blocks of ice float on the surface of lakes.) Those metals which are used to form castings also expand when solidified; but in general bodies expand. Thus, gold and silver coins are stamped, not cast, because they expand when melted. When ice melts, the change in volume is of the same kind as would be produced by an increase in the pressure by mechanical means. In other words, increase of pressure helps on the process of melting; so that if ice is being melted by the addition of heat energy, the temperature does not need to be so high in order to secure melting; *i.e.* the melting point is lowered. If a body expands on melting, an increase in pressure will, for similar reasons, raise the melting point.

Regelation. — This pressure effect is small. In the case of ice and water, if the pressure is increased from one atmosphere to two, that is, by 76 cm. of mercury, the change in the melting point is only 0°.0072 C.; and, consequently, ordinary barometric variations have no measurable effect on the melting point of ice. The fact, though, that the melting point is lowered by an increase of pressure is shown by many familiar illustrations. If two pieces of ice with sharp points are squeezed together, the *pressure* may be enormous because the area of contact may be extremely small; so the melting point of the ice at the points where the pressure is great will be lowered and, if the surrounding bodies are at 0° C., some of the ice will melt, and the resulting water will be pressed out, so that it is at atmospheric pressure, and its freezing point is again 0°. But in order to melt this ice, heat energy must be taken from portions of the body near it, and their temperature is reduced below 0°; so the melted water, with a freezing point of 0°, is in contact with ice at a temperature below 0°, and it will therefore immediately freeze again. This is the explanation of the formation of snowballs; and this action also plays a most important part in the motion of glaciers. The phenomenon is called "regelation."

The gliding motion of a man skating is due to the fact that the ice melts under the pressure of the edge of the skate, and so he is actually moving on a thin layer of water. As the skate moves on, this water freezes again. Another illustration is the formation of so-called "ground ice," which is ice formed at the *bottom* of streams where there are eddies. The ice crystals are whirled round in the current and stick against the bottom, owing to regelation; then others stick to them, etc.

Non-crystalline Substances. — Bodies that are not crystalline, like waxes, plumbers' solder, etc., do not have a definite melting point; but as they are exposed to a source of heat, their temperature rises continuously until they are entirely melted. They pass through a "pasty" condition; and the temperature at which this begins is sometimes called the melting point. Similarly, as the liquid is cooled it begins to pass into the intermediate condition at a temperature called the freezing point. These two temperatures are not the same. These temperatures are affected by changes in pressure exactly in the same manner as those of crystalline bodies.

Undercooling. — The transformation from the liquid into the solid condition does not always take place as described above; for instance, if water is cooled gradually, its temperature will fall far below 0° and yet there is no ice formed. Its condition is, however, most unstable, because if it is shaken or if a minute piece of ice is thrown in, the liquid will solidify immediately and the temperature will rise to 0°. This phenomenon of a liquid existing below its freezing point, as above described, is called "undercooling"; it was discovered by Fahrenheit.

"Heat of Fusion." — The number of calories that must be added to a solid body at its melting point in order to make 1 g. of it melt, is called the "heat of fusion" at that temperature. (It should be noted that this number has the same value if we define the heat unit to be such a quantity of heat energy as will raise the temperature of a *unit mass* of water one degree Centigrade, and define heat of fusion as that number of these heat units which is required to melt a unit mass of the substance, quite regardless of the size of the unit mass.) This energy is spent in overcoming molecular forces and in doing external work if the body expands on melting; if it contracts, the external forces also do work in overcoming the molecular forces. The exact way in which this work is done cannot be determined; but, it is evident that, if the reverse process is carried out, the same amount of heat energy is given out by the body as is received during the direct one. Thus, the temperature of the air in a closed room may be kept from falling far below 0°, if a tub of water is placed in it; for, as the water freezes, a definite amount of heat energy is given off to the air. Again, by placing a pail of water under a fruit tree on a cold night, the fruit may be kept from being injured by the cold.

The heat of fusion of a substance may be determined by various methods, such as are used for the measurement of specific heats. Thus, a known number of grams of ice may be *immersed* (not allowed to float) in a vessel containing a known mass of water at a known temperature, and the fall in temperature may be noted. If

m_1 = mass of ice,

m_2 = mass of water, including the water equivalent of the calorimeter,

t_1 = initial temperature of water,

t_2 = final temperature of water,

L = heat of fusion of the ice;

$$m_2(t_1 - t_2) = m_1 L + m_1 t_2,$$

or

$$L = \frac{m_2(t_1 - t_2)}{m_1} - t_2.$$

Effect of Dissolved Substances; Freezing Mixtures. — The freezing point of a liquid is affected if there is a foreign substance dissolved in it. In every case the freezing point is lowered; and the change is, within certain limits, proportional to the amount of dissolved substance in a given quantity of the solvent, for certain substances. With others, the change is abnormally great; and it is to be noted that these substances are those which have an abnormal osmotic pressure. (See page 181.)

If the temperature of a solution is lowered to its freezing point, the solid formed is that of the *pure* solvent, in general; so that the solution becomes more concentrated. (In certain cases some of the dissolved substance is caught in the meshes of the solid solvent; but this is a mechanical process, not a thermal one.) Then, in order to freeze out more of the pure solvent, the temperature must be lowered still further; for, as said above, the freezing point of the solution falls as its concentration increases. A condition is finally reached with certain solutions such that the solution is saturated; if now heat energy is withdrawn, some of the solvent separates out in the solid form, and at the same time some of the dissolved substance is precipitated; the temperature remains unchanged, and as more and more heat energy is withdrawn, equivalent amounts of solid solvent and dissolved substance separate out. This complex solid mixture is called the "cryohydrate" of the two parts. It is in equilibrium with the solution of the same concentration, as we have just seen, at a definite temperature; so, if a cryohydrate is placed in a region at a higher temperature, it will melt. Thus, the cryohydrate of common salt and water has a composition of 23.8 parts by weight of salt to 76.2 parts of water, and its equilibrium temperature is – 22° C.; so, if salt and ice are mixed thoroughly and are at a temperature greater than this, the ice will melt and dissolve the salt. In this process the temperature of the mixture and of surrounding bodies falls, because heat energy must be supplied both to melt the ice and to dissolve the salt. (See page 283.) If the salt and ice are in exactly right proportions, this process will cease when the temperature – 22° C. is reached. Such a mixture of two bodies as this is called a "freezing mixture"; and the above description explains the use of salt and ice in lowering the temperature of surrounding bodies in "freezers," and also the effect observed when salt is thrown on ice or snow.

A freezing mixture of solid carbon dioxide and ordinary sulphuric ether, known as Thirlorier's mixture, allows one to secure a temperature as low as − 77° C.

The fusion constants of a few substances are given in the accompanying table:

	Fusion Point	Heat of Fusion
Copper	1100° C.	
Ice	0° C.	80
Iron	1400°–1600° C.	23–33
Lead	325° C.	5.86
Mercury	− 39° C.	2.82
Sulphur	115° C.	9.37
Zinc	415° C.	28.1

Evaporation

Boiling Point. — If a liquid stands in an open vessel exposed to the air, it is observed that the quantity of liquid continually diminishes; it is said to "evaporate"; the substance passes from the liquid to the gaseous condition. The gas rising from a liquid is called a "vapor," and an exact distinction between gases and vapors will be made later. (See page 278.) This process of evaporation requires that heat energy should be constantly added to the liquid, as may be proved by direct experiment, or as is seen by the fact that the hand is chilled when any liquid evaporates from it. Therefore the process may be hastened by applying some intense source of heat to the liquid. If this is done, its temperature rises until a point is reached when bubbles of the vapor form in the liquid, rise to the surface, and break. When this stage of "boiling," or "ebullition," is reached, the temperature ceases to rise, and remains constant until all the liquid is boiled away. This temperature is known as the "boiling point," and it is found to vary with the pressure of the air on the surface of the liquid. This is what we should expect, because in order that the bubbles may form and rise to the surface, the pressure of the vapor in them must be at least as great as the pressure of the air on the surface, and the pressure of the vapor as it rises from the surface must equal this; so, if this pressure on the surface is increased, the liquid must be raised to a higher temperature before it will boil. Similarly, if the external pressure is decreased, the boiling point is lowered. The process of boiling is one, then, of what may be called kinetic equilibrium, depending upon the equality of the pressure on the surface and that of the vapor as it rises from the surface.

Saturated Vapor. — If the liquid is contained in a closed vessel, it is observed that after a time the evaporation apparently ceases. The vapor above the liquid is now said to be "saturated." If the pressure and temperature of this vapor are noted, it is found that if the temperature is raised, the pressure increases, and more liquid is evaporated; while, if the temperature is lowered, the pressure decreases, and some of the vapor condenses to form more liquid. If the temperature is kept constant, however, the pressure remains the same, entirely independent of whether there is a small or a large amount of liquid present. This condition may be called one of statical equilibrium. On the kinetic theory of matter it is easily explained. The molecules of the liquid may attain sufficient velocity to break through the surface, thus requiring work to be done upon them. Similarly, the molecules of the vapor may strike against the surface and become entangled, thus losing kinetic energy. So, if both these processes go on together, there will be equilibrium when the number of liquid molecules which escape in any interval of time equals the number of vapor molecules which are retained by the liquid surface in the same time.

Spheroidal State. — A simple illustration of the kinetic nature of evaporation is afforded by what is called the "spheroidal state." If a small quantity of water is allowed to flow gently out of a tube or spoon on to a metal surface which is at a high temperature — far above 100° C. — and which is slightly hollowed out so that the water will not run off, it is seen to collect in a flattened drop, which does not rest on the surface, and it now rapidly evaporates. The explanation is evident; for, owing to the rapid evaporation on the lower side occasioned by the heat energy received from the hot metal, the molecules are leaving the drop on this side with such velocity and in such quantity that their mechanical reaction holds up the drop. There is thus a layer of vapor between the drop and the hot plate. This spheroidal state can be noticed when water is spilled on a hot stove, and in fact the hotness of a stove or a flatiron is often tested in this manner by seeing if it can produce this state in small drops of water. (This condition of a drop is sometimes called "Leidenfrost's Phenomenon," because the first observation of it is attributed to him, 1756.)

Vapor Pressure. — One of the simplest methods of observing the phenomena of saturated vapor is to introduce a small quantity of the liquid to be experimented on above the mercury column in a barometer which has a deep basin. Some of the liquid will evaporate, and equilibrium will be reached at a pressure depending upon the temperature. This may

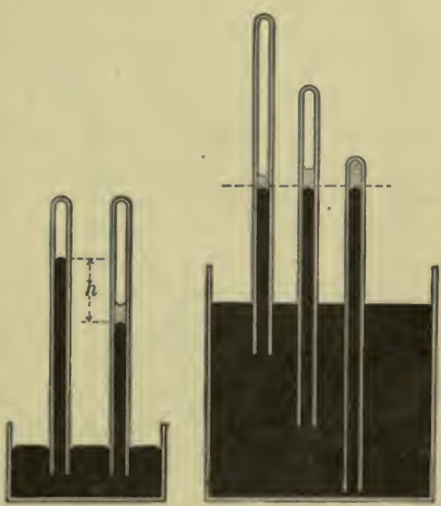


FIG. 120. — Experiments illustrating the laws of the pressure of saturated vapor: a small amount of the liquid, *e.g.* water, is introduced above the mercury in a barometer.

be varied at will by surrounding the tube with a bath of some liquid whose temperatures can be regulated. The pressure of this saturated vapor may be measured in terms of the atmospheric pressure by the ordinary law of hydrostatic pressure. (If h is the difference in height of the mercury column in this tube and of that in a barometer, the pressure of the vapor is less than the atmospheric pressure by dgh, where d is the density of mercury.) If the temperature is kept constant and the barometer tube slowly raised, some liquid will evaporate, but the *pressure remains constant;* similarly, if the tube is slowly pushed down, some of the vapor is condensed, but the pressure does not change. (If there were a gas above the mercury column instead of the vapor, its mass would not change, but its pressure would decrease and increase during the above changes, obeying Boyle's law: $pv =$ constant. But when there is a vapor in contact with its liquid, there is no change in pressure, but the mass of the vapor increases and decreases.) If, however, the temperature is increased, some liquid evaporates and the pressure increases; and, if it is lowered, some vapor condenses, and the pressure decreases. (If

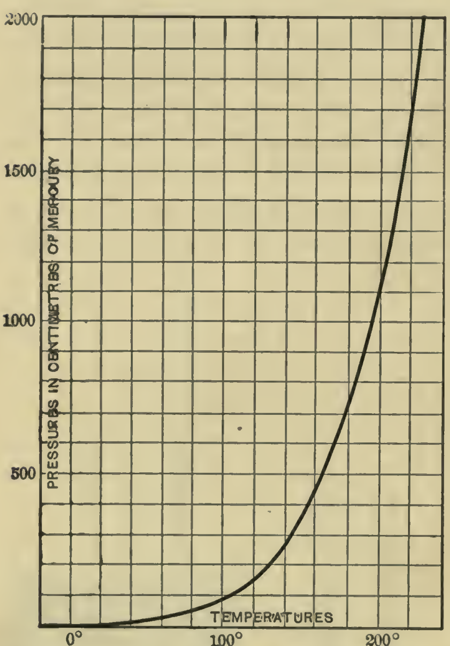


Fig. 121. — Curve showing connection between the temperature and the pressure of saturated water vapor.

there were a gas above the mercury column, its mass would remain constant during the above changes in temperature, and its pressure would change, but at a different rate from that of the saturated vapor.) There is thus seen to be a definite pressure of the saturated vapor which corresponds to a definite temperature when there is equilibrium, and conversely; the corresponding values may be found by either the statical method or the kinetic one, in which the boiling point at different pressures is determined. The results may be expressed by a curve drawn with axes of temperature and pressure. This curve for water vapor is given.

That the boiling point varies directly with the pressure is shown by the fact that the temperature of boiling water is much less than 100° C. on a mountain top ; by the high temperature in steam boilers where the pressure is great, etc.

From the description given above of the statical method, it is seen that there are two general methods available for condensing a vapor into a liquid : one is to lower the temperature ; the other is to decrease the volume. These will be discussed in full in a later section. If a vapor is not saturated, it obeys Boyle's law quite closely; and Dalton's law is also approximately exact for the pressure produced by a mixture of vapors.

A curve may be drawn that will express this law of a vapor in contact with its liquid when the temperature is constant. Lay off axes of pressure and volume; then, since a vapor keeps its pressure constant so long as the temperature is constant, the isothermal is a straight line parallel to the axis of volumes.

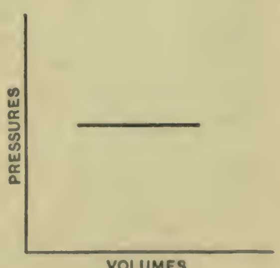


FIG. 122. — Isothermal of saturated vapor.

Formation of Dew, Clouds, etc. — We have seen that, as the temperature is lowered, the corresponding vapor pressure of the saturated vapor becomes less ; that is, if there is a certain amount of unsaturated vapor in a closed vessel, it will become saturated if the temperature is lowered sufficiently; and then, if it is lowered still more, some of the vapor will condense. This is illustrated by the formation of dew, of clouds, etc.

There is always a certain amount of moisture in the air, and the method of expressing it is as follows: we measure the temperature of the air and then by experiment find what temperature some solid body — like a metal can — must have in order to make moisture condense on it. This is called the "dew-point." The vapor pressure corresponding to these two temperatures is found from tables or from the curve given on page 266: one of these expresses the pressure that the water vapor in the air *might* have if it were saturated at the existing temperature; the other gives the pressure that the water vapor in the air actually has. The ratio of the latter to the former gives what is called the "humidity."

If we assume the truth of Dalton's law, we can easily calculate the mass of a unit volume of ordinary damp air. This equals the sum of the masses of the air itself and of the water vapor in the space. The quantity of water vapor in a unit volume corresponding to various dew-points is given in tables; *e.g.* if the dew-point is 10° C., the mass per cubic metre is 9.3 g. The pressure of this vapor corresponding to 10° C. is 0.914 cm. of mercury; thus, if the barometric pressure is 76 cm., the pressure due to the *air* is 75.086 cm.; and, if the temperature of the air is known, *e.g.* let it be 20° C., the mass of the air can be calculated, since the density of dry air at 0° C. and 76 cm. pressure is known to be 0.00129. Thus, using the gas law, the density at 20° C. and at 75.086 cm. pressure equals $\frac{75.086}{76}\,\frac{273}{293}\cdot 0.00129$, or 0.00119. Therefore the mass of the air in a cubic metre is 1190 g.; and the total mass of the cubic metre of damp air is $1190 + 9.3 = 1199.3$ g. If the air were perfectly dry and at 20° and 76 cm. pressure, its density would be $\frac{273}{293} \times 0.00129$, or 0.00120; and so the mass of a cubic metre of it would be 1200 g. It is seen, then, that the mass of dry air is greater than that of damp.

Boiling. — As already explained, the process of boiling consists in the formation of bubbles of the vapor in the interior of the liquid. Nuclei of some kind are required in order for these to form, such as sharp points or minute bubbles of some foreign gas, like air. As a liquid boils, the supply of such nuclei is used up, unless in some way it is renewed constantly, and it becomes more and more difficult for the liquid to boil. The temperature rises above the boiling point until the molecular forces are sufficient to form the bubble; there is a miniature explosion; and the temperature falls back to the boiling point. If the nuclei are removed from the liquid as completely as possible and if the walls of the containing vessel are smooth, the temperature of the liquid may be raised far above the boiling point; but this condition is of course unstable. If the liquid were entirely free from nuclei, it would never boil; but, if its temperature were gradually raised, it would finally explode.

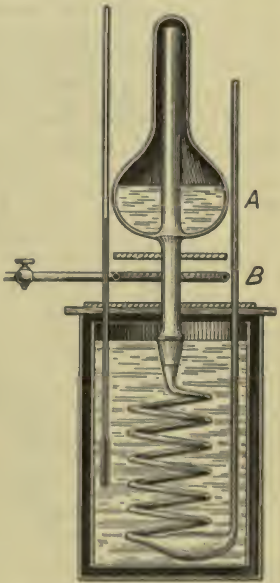


FIG. 123. — Apparatus for determination of the heat of vaporization of a liquid.

B is a ring-burner for heating the liquid contained in the vessel *A*; the vapor is condensed in the coils in the lower vessel.

Heat of Vaporization. — The number of calories that must be added to a liquid in order to make one gram of it evaporate, or boil at a definite temperature, is called the "Heat of Vaporization" at that temperature. (This number does not depend upon the size of the unit of mass if the heat unit is suitably defined. See page 260.) This energy is spent in securing internal changes corresponding to the transition from liquid to vapor and also in doing external work; this last can be calculated because the pressure must be kept constant, inasmuch as the temperature is assumed not to vary. The same number of calories is also given out when one gram of a vapor condenses at that temperature. This physical quantity is measured ordinarily by the method of mixtures; a known mass of vapor is condensed by making it enter a quantity of its liquid through a spiral tube, or in such a manner that it gives up the heat energy entirely to the liquid. A simple form of apparatus is shown in the cut.

If m_1 = the mass of vapor condensed,

m_2 = the mass of liquid in the calorimeter originally, including the correction for the calorimeter,

t_1 = initial temperature of the liquid,

t_2 = final temperature of the liquid,

t_3 = boiling point of the liquid at the given pressure,

c = specific heat of the liquid,

L = heat of vaporization,

then $m_1L + m_1c(t_3 - t_2) = m_2c(t_2 - t_1)$,

or
$$L = \frac{m_2c(t_2 - t_1)}{m_1} - c(t_3 - t_2).$$

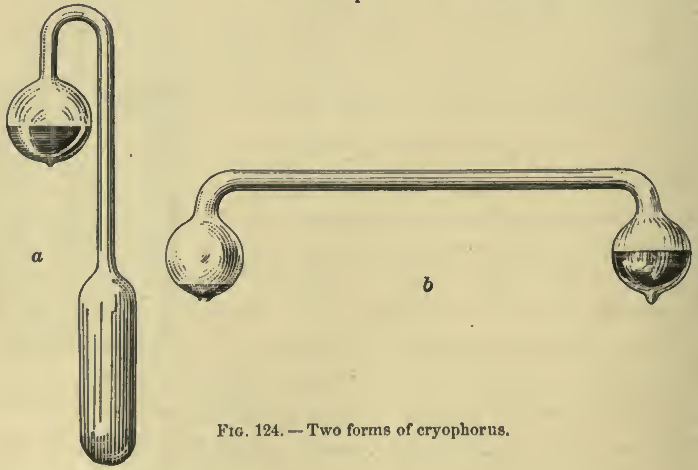
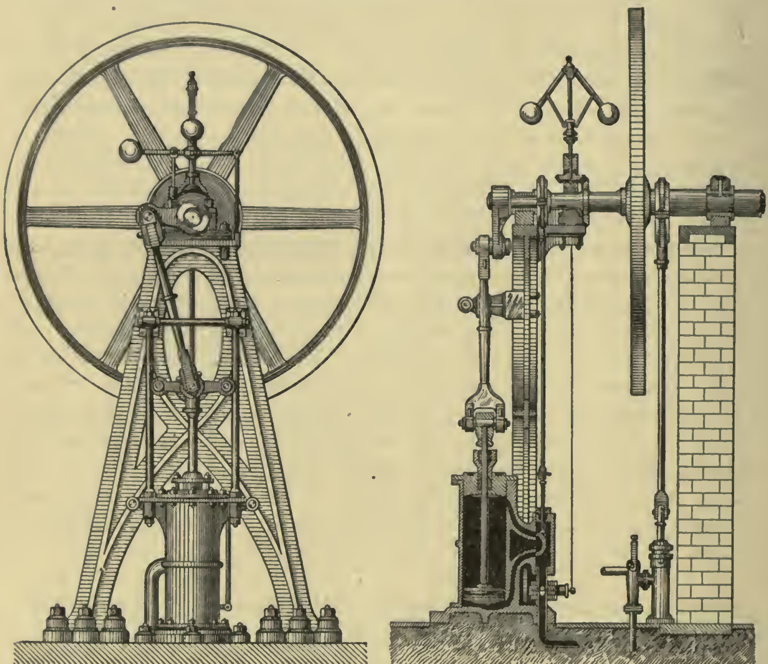


FIG. 124. — Two forms of cryophorus.

Several illustrations have been given already of the changes in heat energy when a liquid evaporates or a vapor condenses, but one or two more may be described. In an experiment due originally to Boyle, a small quantity of water is placed under a bell jar attached to an air pump, great care being exercised to guard against any possible conduction of heat to the water; as the air is exhausted, thus diminishing the pressure on the water and removing the vapor, the water evaporates so rapidly that the heat energy required is taken from the water left behind, and its temperature falls until it freezes. A somewhat similar experiment is one that involves the use of the "cryophorus," an instrument invented by Wollaston. This consists of two glass bulbs connected by a bent glass tube, as shown in the cut in two forms, *a* and *b*. There is sufficient water inside to half fill one of the bulbs. The experiment consists in placing the instrument *a* in a vertical position, with the curved portion uppermost, or the instrument *b* horizontal; the water is poured into one bulb, which is carefully shielded against heat loss or gain, and the other is surrounded by a freezing mixture of salt and ice. After a short time the water will be found to be frozen, owing to rapid evaporation at its surface, which is caused by the continuous condensation and freezing of the vapor in the other bulb. The action, then, is the same as if a substance "cold" were carried from one bulb to the other; hence the name "cryophorus," which means "carrier of cold." Some forms of apparatus for making artificial ice depend upon the same fact, that when a liquid evaporates, heat energy is required. In most cases the liquid which is evaporated is ammonia. This is placed in the space between the two walls of a double-walled vessel which contains water, and as the ammonia is evaporated, the water is frozen.

Steam Engine. — As a further illustration of the properties of a vapor, the steam engine may be mentioned — a diagram of a simple form of which is given in the cut. The principle of its action is as follows: Steam is produced in a "boiler" under high pressure, and therefore at high temperature; this steam is allowed at regular intervals to enter the "cylinder," in which there is a movable piston, at the instants when the piston reaches one end of the cylinder, and in such a manner as to exert a pressure on this piston, pushing it away from this end. Steam continues to enter, and the pressure is that of the steam in the boiler as long as the connection is maintained; but after the supply of steam is cut off, the steam, as it expands, decreases in pressure. In the meantime the pressure in the cylinder on the other side of the piston has been made as small as possible by one of three methods: (1) by opening it to the atmosphere—this makes a "non-condensing" engine; (2) by join-



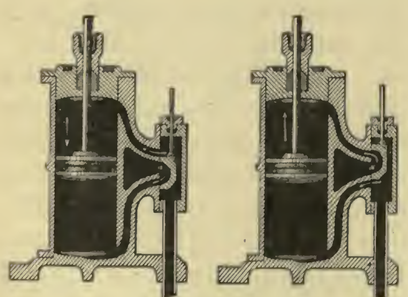


Fig. 125. — Steam Engine.

ing it to a large vessel, which is kept as nearly exhausted as possible by means of pumps and as cool as possible by means of coils or jets of water — this makes a "condensing" engine; (3) by joining it to another cylinder, exactly like the first one, only larger, and allowing the steam ejected from the first cylinder to work the piston in the second — this makes a "double, triple, etc., expansion" engine. When the piston reaches the other end of the cylinder, connection is made with the boiler, as above described, at this end, and the expanded steam is expelled, as the piston retraces its path, in one of the ways just mentioned. In the case of the non-condensing engine the steam escapes to the air and is lost; in that of the condensing engine it is condensed into water in the condenser and pumped back into the boiler; in the expansion or compound engine it is used over again, expanding more and more until, finally, it is condensed in a condenser and pumped back to the boiler. These changes are made automatically by certain valves; the "sliding valve" opens and closes the inlet pipes and also the exhaust pipe to the condenser. In all cases a certain quantity of saturated vapor is received at a definite pressure (and corresponding temperature); this expands, doing work on the piston, until a certain lower pressure is reached, then — in the condensing engine — it is condensed to water at the pressure corresponding to the temperature of the condenser; this water is pumped into the boiler, its temperature is raised until it boils, and the process begins again. There is thus a "cycle" of changes. While in the boiler, the water *receives* heat energy; and when water is being formed in the condenser, heat energy is *given out* by the steam. The steam does work in pushing on the piston, and work is done on it and on the water formed from it when the piston performs its reverse motion. If H is the heat energy received at the temperature of the boiler, and W is the *net* external work done, the ratio $\frac{W}{H}$ is called the "efficiency" of the process.

The amount of work done may be measured by a simple mechanical device. We have seen (page 161) that on a pressure-volume diagram, a closed curve describing the series of changes through which a fluid passes indicates by its area the total net work done. The curve giving the cycle of changes just described for the steam leaving the boiler, expanding, etc., must be somewhat as shown in the cut. A marks the instant when the *water* in the boiler begins to be changed into steam; B, when this process is finished, the pressure and temperature remaining constant; this steam can be imagined as formed directly back of the piston and exerting a pressure on it; so the instant marked by B is that of the "cut-off"; the steam now expands, and both temperature and pressure fall; at C, connection is made with the condenser, in which the pressure is that marked by the horizontal line $\overline{ED}$; so the pressure falls to D, and the steam condenses to water as marked by E; this water is forced by a pump into the boiler, and its temperature and pressure both increase until the point A is again reached. Looked at from another point of view, this curve indicates very approximately the changes in volume and pressure *of the space in the cylinder* which is open to the steam as the piston moves to and fro: when the piston is close to one end, the volume is small, and as steam is admitted, the pressure rises from E to A; then it remains constant as the piston is pushed out, thus increasing the volume, until the cut-off is reached at B; then, as the piston continues to move forward, the volume increases and the pressure falls; C marks the end of the motion of the piston; as it moves back, the volume decreases, but the pressure remains unchanged, being that of the condenser; etc. These pressure and volume changes may be recorded automatically by the engine in several ways: a piece of paper may be fixed to a drum and a pencil moved over it, motion in one direction being secured

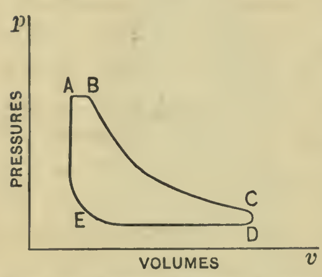


FIG. 126. — Diagram of a cycle of operations in a steam engine.

by attaching the pencil to a spring gage which measures the pressure in the cylinder and moves out and in as this increases and decreases, and motion at right angles to this being obtained by fastening the drum carrying the paper to a cord which is attached to the moving piston, and so lengthens or shortens as this moves, and thus makes the volume in the cylinder increase or decrease. These curves are called "indicator diagrams," and are of the greatest assistance in studying the actual working of an engine. A form of indicator mechanism often used is shown in the cut.

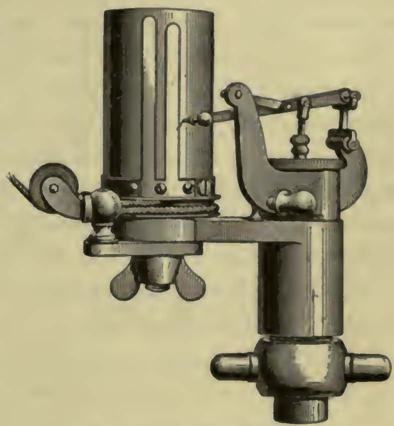


FIG. 127. — Indicator mechanism.

The first successful steam engine was made by Newcomen in 1705; but all the great improvements, as shown in modern engines, are with one exception due to James Watt. He invented the condenser in 1763; in succeeding years he conceived the idea of introducing the steam alternately on the two sides of the piston, and of cutting off the steam from the boiler so as to allow it to expand; he also invented the plan for surrounding the cylinder with a jacket into which steam could be admitted, so as to keep the cylinder always hot. Hornblower was the first to construct a compound engine, *i.e.* one with two cylinders, so as to have double expansion. This was done in 1781.

Effect of Dissolved Substances. — The vapor pressure of a liquid is affected if some substance is dissolved in it, being always lower at a given temperature than that over the pure liquid. Therefore the boiling point of a solution is higher than that of the pure liquid at a given pressure. These changes in the pressure or in the boiling point vary with the concentration of the solution. They are proportional to it in some cases, but in others the change is abnormally great. These last solutions are those for which there is an abnormal lowering of the freezing point and an abnormal osmotic pressure. When a solution evaporates or boils, the vapor formed is that of the pure solvent unless the dissolved substance is volatile; consequently, when the temperature of the boiling solution is to be measured, the thermometer must be immersed in the liquid, not in the vapor.

The vaporization constants of a few liquids are contained in the accompanying table:

	Boiling Point at Pressure of 76 Cm. of Mercury	Heat of Vaporization
Alcohol (ethyl) . . .	78° C.	209
Carbon dioxide . . .	– 80° C.	72 at – 25°
Chloroform	61°.20 C.	58.5
Cyanogen	– 20°.7 C.	103 at 0°
Ether (ethyl)	34°.6 C.	90
Hydrogen	– 238° C.	—
Mercury	357° C.	62
Oxygen	– 184° C.	—
Sulphur	444°.5 C.	—
Water	100° C.	535.9

Isothermal for a Change in State from Vapor to Liquid. — The various physical properties of liquids and vapors can best be expressed and studied by graphical means. One method is to draw isothermals on a diagram whose axes correspond to pressure and volume. Let us imagine the vapor inclosed in a cylinder that has a movable piston. We can then represent by curves the various conditions of the vapor, as the piston is forced in gradually, the temperature being kept constant. At first, as the volume is diminished, the pressure increases according to Boyle's law, since the vapor is unsaturated, as shown by the curve *AB*. But finally, a pressure is reached such that the vapor is now saturated for the given temperature; and, if the volume is still further decreased, the pressure remains unchanged, as shown by the curve *BC*. As this decrease in volume goes on, more and more of the vapor condenses and sinks to the bottom of the cylinder. When all the vapor is condensed, the cylinder is full of liquid, and to produce any further decrease in volume a great increase in pressure is

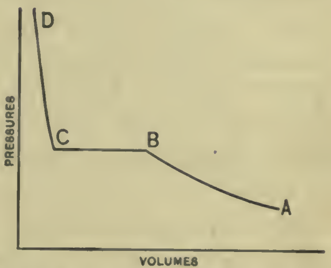


FIG. 128. — Isothermal for change from a vapor to a liquid.

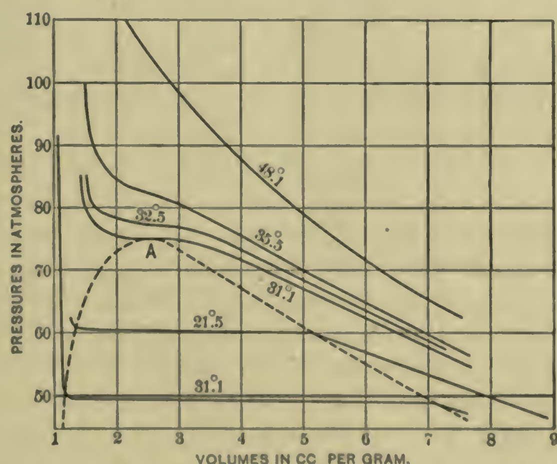


FIG. 129. — Isothermals of carbonic acid (CO_2).

required, as shown by the curve *CD*, which is nearly vertical. Curves of this kind were first determined and studied by Andrews, working with carbonic acid gas, in the years previous to 1869. The results of his investigation are given in the accompanying cut. It is seen that, as the isothermals at higher and higher temperature are drawn, they have the same general shape, but the horizontal portions become shorter, until a temperature is reached, the isothermal for which has no horizontal portion. This is called the "critical" temperature. For temperatures higher than it, the isothermals approximate more and more closely to those of a gas.

The Critical Temperature. — The only points on the diagram for which the matter is in the form of a liquid are along the horizontal portions of the isothermals, where there is a free surface separating the liquid and its vapor, and along the continuations of these lines to the left, where the liquid completely fills the vessel. Consequently, whenever we see a liquid partially filling a vessel, it is represented on the diagram by a point on a horizontal portion of an isothermal. In other words, if a vapor is to be liquefied, it must be at a temperature whose isothermal has a horizontal portion, that is, it must be at a temperature lower than the critical one. Bearing this fact in mind, all gases, with the possible exception of helium, have been liquefied. The name "vapor" may then be restricted to a body in the gaseous condition which is at a temperature below the critical one; while the name "gas" may be limited to temperatures above this. The exact point on the diagram, marked by A, where the critical isothermal has its point of inflection, that is, the point where an isothermal infinitely near the critical one, but below it, has a minute horizontal portion, is called the "critical point." If we have matter in this condition filling a vessel, and if the temperature is lowered, even the least amount, the liquid will separate out and sink to the bottom, filling approximately half the space and leaving the rest full of vapor. At the critical point, then, the surface of separation disappears, and the liquid and vapor dissolved in each other make a homogeneous form of matter. The pressure corresponding to the critical point is called the "critical pressure," and the volume of one gram of the substance in this critical condition is called the "critical volume."

In order to liquefy a gas, then, two steps are necessary; the temperature must be lowered below the critical temperature, and the volume must be decreased until the pressure is reached that corresponds to the state of saturation of the vapor for the temperature. After this, any further decrease in volume or temperature will cause the vapor to condense.

The values of the critical temperatures for various gases are given in the accompanying table:

CRITICAL TEMPERATURES

Alcohol	243°.6 C.	Hydrogen	– 242° C.
Ammonia	130° C.	Nitrogen	– 146° C.
Argon	– 120° C.	Oxygen	– 119° C.
Carbon dioxide	30°.9 C.	Sulphur dioxide	156° C.
Chloroform	260° C.	Water	365° C.

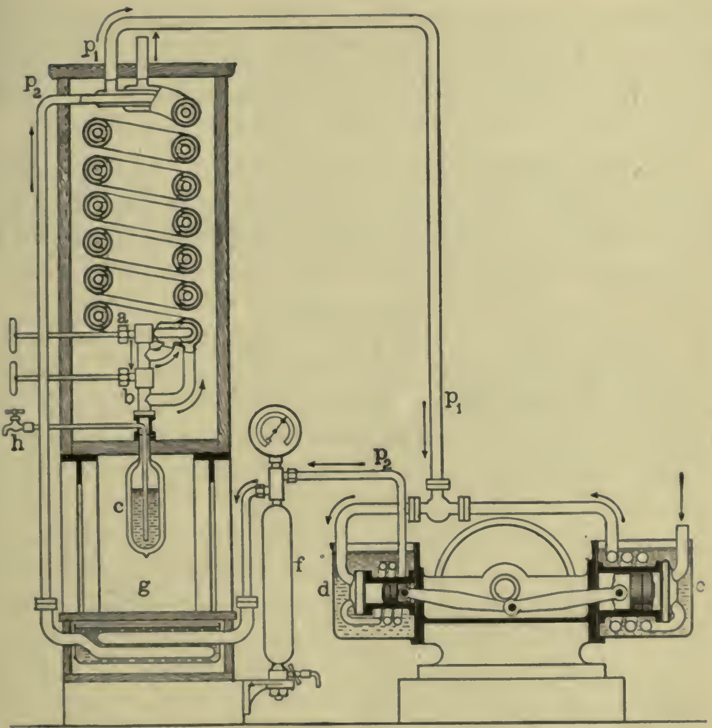
Liquefaction of Gases. — The critical temperatures of such gases as hydrogen, oxygen, nitrogen, etc., are seen to be extremely low; so that special means must be adopted in order to liquefy them. There are three methods in use for the production of low temperature; application of freezing mixtures, rapid evaporation of a liquid, expansion of a gas from high to low pressure. (See page 246.) The standard method for liquefying gases is a combination of these to a certain degree. The gas to be liquefied is compressed by pumps to a high pressure and is cooled by a freezing mixture or by evaporation of a liquid; it is then allowed to expand through a small opening, is again compressed; and the process is repeated. This expanded gas is colder than it was before expansion; and before being compressed again, as fast as it expands through the opening, it is drawn back over the vessel containing the unexpanded gas, thus chilling it. As the process continues, the temperature of the compressed gas gets lower and lower, until finally, on expansion, the critical temperature is passed and drops of the liquefied gas fall to the bottom of a Dewar bulb.

Two forms of apparatus are shown in the cut on page 281; one due to Linde, the other to Dewar. The most important parts of the former are the two-cylinder air compressor and the "counter-current interchanger." This last consists of a triple spiral of three copper tubes wound one inside the other. The cycle of operation is performed in such a manner that compressed air at the temperature of the coil *g* and at about 200 Atm. flows through the inmost tube of the spiral from top to bottom, and passes out at the lower end through a valve *a* under a pressure of some 16 Atm., it returns upwards through the annular space between the inner and middle pipes, is again raised to a pressure of 200 Atm. by the smaller cylinder *d* of the compressor, and then begins the same cycle over. The larger cylinder *e* of the compressor pumps a small amount of air from the atmosphere into the suction pipe of the small cylinder, that is to say at 16 Atm. A similar quantity of air must therefore leave the cycle at another point so that the pressures may remain constant. This escape of air takes place at the lower end of the counter-current interchanger immediately after the discharge from 200 to 16 Atms. so that a controllable amount of air issues from 16 Atms. to 1 Atm. through a second valve *b*. Part of this air, when the apparatus is cooled down to the temperature of liquefaction, becomes liquefied and collects in a "Dewar flask" *c*. That part of the air issuing from the second valve which is not liquefied leaves the apparatus, escaping through the space between the middle and outside pipes of the spiral into the atmosphere.

An iron pipe in the form of a coil *g*, which is cooled down to a few degrees below zero by a freezing mixture of ice and chloride of calcium freezes out the small quantities of water vapor contained in the highly compressed air until only traces remain, and also cools the air.

The properties of bodies at extremely low temperatures are entirely different from what they are at ordinary temperatures. The temperature of liquid air when boiling at atmospheric pressure is $-182°$ C.; under these conditions lead becomes elastic and iron and india rubber extremely brittle; egg shells, leather, etc., become phosphorescent; the electrical resistance of all pure metals decreases greatly; etc.

By suitable methods, air, oxygen, and even hydrogen may be solidified.



Apparatus of Linde for the liquefaction of air and other gases.

Apparatus of Dewar. *A* is a cylinder containing the gas under high pressure; *B* is a Dewar flask containing a freezing mixture such as solid carbonic acid and ether; *C* is a second flask containing some liquid which is made to evaporate rapidly, *e.g.*, ethylene or liquid air; *D* is a third flask. Coils bring the gas through *B*, *C*, and *D*; and it is allowed to escape through a fine opening, *G*, whose size is regulated by the rod *F*. As the gas escapes, it is drawn out through the opening *E*, so that as it rises it cools the coils in the tube *D*.

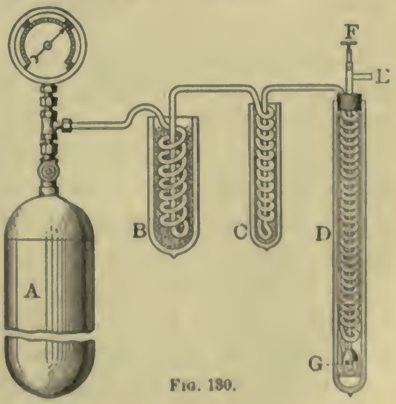


FIG. 130.

The melting point of hydrogen is estimated at − 257° C.; and the maximum density of liquid hydrogen is 0.086. By allowing a quantity of liquid air to evaporate slowly, and using the methods of fractional distillation, Ramsay discovered three new constituents of our atmosphere, which he called *Krypton*, *Xenon*, and *Neon*.

Continuity of Matter. — It is seen from the experiments of Andrews that it is possible to make a body pass from the state of vapor to that of liquid, and *vice versa*, by a series of *continuous* changes, because a path can be drawn from a point on the pressure-volume diagram where the body is in the form of a vapor to one where it is liquid, which does not pass through the region where the isothermals are horizontal, and where therefore the liquid and vapor exist separately, but in contact. The *isothermal* has two points of discontinuity, at the two ends of its horizontal portion, which correspond to molecular rearrangements; but a series of changes can be imagined, as above described, during which there will be no sudden molecular changes, but by which a vapor can be gradually and continuously changed into a liquid. This is ordinarily expressed by saying that matter is continuous from the liquid to the vapor state. Similarly, matter is continuous from the solid to the liquid state.

Sublimation

In many cases a solid body evaporates directly without passing through the liquid condition. This process is called "sublimation"; and it is illustrated by camphor, arsenic, iodine, carbon, many metals, snow or ice, etc. It is found by experiment that, if this process takes place in a closed vessel, there will be equilibrium between the solid and the vapor when a definite pressure is reached, which depends upon the temperature. If the latter is increased, the equilibrium pressure is higher, and conversely. The reverse of this process of sublimation is seen in the formation of frost.

It is found that all these substances which sublime under ordinary conditions may be obtained in the liquid condition if a suitable pressure and temperature are applied. (This is one step in Moissan's method of making artificial diamonds.)

The number of calories required to make one gram of a solid sublime at a definite temperature is called the "Heat of Sublimation." It equals the sum of the Heats of Fusion and of Evaporation at that temperature, in accordance with the principle of the Conservation of Energy.

Solution

Heat of Solution. — When a body — solid, liquid, or gas — is dissolved in a liquid, both being at the same temperature, there is a change in temperature, showing that heat changes are involved. If the temperature falls, it shows that work is required to make the substance dissolve, if we assume that there are no secondary molecular changes such as the formation of new molecules or dissociation. The heat energy that is gained or lost is measured by the product of the mass of the solution, its specific heat and the increase or decrease of temperature. If the temperature rises, heat energy is said to be "evolved"; if it falls, the energy is said to be "absorbed." If one gram of the substance is dissolved in a certain quantity of solvent, the heat energy thus involved will, in general, vary with the quantity of solvent; but, by continually increasing this, it is found that after a certain point the heat changes are independent of the quantity of solvent. The heat energy gained or lost, when one gram of a substance is dissolved in such a large quantity of the solvent as this, is called the "Heat of Solution" of the substance. Some values are given in the following table, the solvent being water. A plus sign indicates a rise in temperature, or "evolution of heat," and a minus sign, the opposite.

Ammonia gas	. +495.6 calories.	Caustic potash	. +223.3 calories.
Ethyl alcohol	. + 55.3 calories.	Sodium chloride	. − 18.2 calories.
Sulphuric acid	. +182.5 calories.	Silver chloride	. −110. calories.

Effect of Rise in Temperature upon Solution. — We can from these facts predict whether raising the temperature of a solution increases its solubility or not. Let us consider a saturated solution with an excess of dissolved substance precipitated; and let it be one in which heat energy is absorbed when solution takes place. If now heat energy is added, the solvent will dissolve more of the substance; and if heat energy is withdrawn, the solvent will precipitate some of its dissolved substance; because in this last case, for instance, if the effect were to make the solution more soluble, this act of solution would withdraw some heat energy, etc., and the condition would be unstable. Therefore, with a solution of this kind, an increase in temperature makes it capable of dissolving more; a decrease, less. Just the converse is true of those solutions in whose formation heat energy is evolved.

Dissociation; Ions. — The fact that certain solutions have an abnormally great osmotic pressure, and also have freezing and boiling points that differ abnormally from those of the pure solvent, can all be explained if it is assumed that in these solutions a certain proportion of the dissolved molecules are broken up into smaller parts. This has the immediate effect of increasing abnormally the number of moving particles due to the dissolved substance. If we assume, further, that these fragments of the molecules are electrically charged, we can explain the phenomena of electrolysis, as we shall see later; and all solutions of this kind do conduct electric currents. The whole science of Physical Chemistry is based upon these two assumptions, and they may be regarded as justified by experiments. When such a solution is formed, some of the molecules dissociate into their parts since thereby the potential energy is made less; or, as we may express it, there is a force producing dissociation. The process ceases — or equilibrium is reached — when this "solution pressure" is balanced by the electrical forces that are called into action. The equilibrium is not one in which there is no further dissociation, but one in which for each molecule dissociated there is one formed. The dissociated parts, called "ions," are moving to and fro in the solvent, and combinations and dissociations are taking place continually but at the same rate.

Chemical Reactions

Heat of Combination. — In all chemical reactions there are molecular changes, and consequent heat changes. If the reacting bodies are gases, these changes depend to a marked degree upon the external conditions that are maintained; for these determine the amount of external work. It is entirely immaterial, however, whether the change takes place in one or more stages. Thus, if one gram of carbon in the form of a diamond is converted into carbon monoxide (CO), 2140 calories are evolved; and if this is converted into carbon dioxide (CO_2), 5720 more calories are involved — 7860 in all. And, if the same amount of carbon is oxidized at once into carbon dioxide, the heat involved is the same. This is an illustration of the Conservation of Energy.

A few illustrations of these heat changes may be given. If 2 g. of hydrogen gas and 16 g. of oxygen, at 0° C. and 76 cm. pressure, combine to form 18 g. of water at 0°, the heat energy evolved is 68,834 calories. If 65 g. of zinc are dissolved in dilute sulphuric acid, 38,066 calories are evolved; while if 63 g. of copper are dissolved in dilute sulphuric acid, 12,500 calories are absorbed.

Dissociation. — One of the most interesting chemical changes is the dissociation of a gas into other gases. Thus some gases with complex molecules break down into others with simpler molecules when the temperature is raised to a high degree. In many cases it is observed that for a definite temperature equilibrium is reached at a definite pressure; and, if the temperature is increased, so is the corresponding pressure. This condition of equilibrium is one of continual recombination and dissociation.

CHAPTER XIV

CONVECTION, CONDUCTION, AND RADIATION

It has been shown that there are three general methods by which heat energy is added to or taken from a body: Convection, Conduction, and Radiation. Each will now be discussed in turn.

Convection

When a vessel containing a liquid is placed on a hot stove, the upper layers of the liquid receive heat energy from the lower ones by the process known as "convection"; the portions of the liquid in the lower layers have their temperature raised and therefore expand; and, since their density is thus diminished, they are forced upward by gravity, the cooler upper portions sinking. Thus the temperature of the whole liquid is made uniform. The mechanics of the phenomenon is not difficult to understand, because the molecules of the hot upward-moving portions of the fluid communicate by their impacts some of their energy to the other molecules; and thus the internal energy is distributed.

It is evident that convection processes can occur only in *fluids* under the action of *gravity*, and when the heat energy is applied to its lower portions. It should be noted that the energy is gained by the portions at low temperature and is lost by those at a higher temperature.

This method of distribution of heat energy is the one that forms the basis of the ordinary means of heating houses,—hot-air furnaces and stoves, hot-water systems, steam pipes, etc. Convection is of fundamental importance in the economy of nature, as is explained in Physical Geography, in connection with winds, ocean currents, etc. The metal bottom of a tea kettle (or of a steam boiler) does not become unduly hot so long as there is water in it, because, owing to convection, the cooler portions of the water are being continuously brought down to the bottom.

Conduction

General Description. — When one end of a metal rod, like a poker, is put into a fire, the temperature of this end rises, and in a short time that of the other portions not too far from the fire rises also. The temperature of the end in the fire is the highest; and that of the other points of the rod decreases gradually as one passes from the fire, until a point in the rod is reached that is at the temperature of the surrounding air. If a thin transverse portion or slab of the rod is considered after it has come to a steady state, it is evident that its side near the flame is at a higher temperature than the one away from it; the molecules at the former section have more energy of motion than those at the latter. As a consequence, the former molecules give up some of their energy to the molecules of the slab; and its temperature would rise were it not for the fact that the slab is losing heat energy by convection in the surrounding air (and by radiation, also, a process to be described presently), and that the molecules in the cooler end of the slab are themselves handing on energy to the other portions of the rod. This process by which molecules give up some of their energy to contiguous molecules, there being no actual displacement as in convection, is called "conduction." Thus, considering the slab across the rod, we say that it gains heat energy at the hot face and loses it at the cooler one by conduction; and the difference between the quantities gained and lost must equal that lost at the surface of the rod by convection (and radiation), since the rod is in a steady state. It is important to note that the heat energy is conducted from the hotter portions of matter to the colder ones. When the rod is not in a steady state, *e.g.* immediately after one end is put in the fire, part of the energy that enters the slab goes to raising its temperature, to doing external work, etc.

"Conductivity" for Heat. — If the rod is in a vacuum, there is very little energy lost, in general, from the surface, because there is now no convection; and, when the bar is in a steady state the energy conducted in at one section of a slab equals that conducted out at the other. If t_1 is the temperature of one section and t_2 that of the other, if A is the area of each section and a the thickness of the slab, experiments show that the quantity of heat energy conducted through from the former section to the latter, if t_1 is greater than t_2, is proportional to $\frac{(t_1 - t_2)A}{a}$, but is different for rods of different material. This fact may be expressed by the following equation, in which Q is the quantity of heat energy conducted by the slab, $Q = k\frac{t_1 - t_2}{a}A$, where k is a factor of proportionality, which is different for different bodies. It is called the "conductivity" for heat. If the conductivity of one body is greater than that of another, it is said to "conduct better." Thus silver conducts better than copper; copper, better than iron; all metals, better than wood and other non-metals; etc. The conductivity of any one body varies slightly with its temperature.

If the conductivity of a fluid is to be determined, the *upper* surface must be made the hotter, so as to avoid convection. All liquids conduct poorly, with the exception of fused metals; and all gases conduct still worse. Thus loss of energy from a body by the processes of conduction and convection may be avoided by inclosing it in a quantity of eider down, feathers, or loose wool or felt; because these solids are poor conductors and motion of the air inclosed by them is prevented, as it is contained in small cavities. The best method of all, however, for avoiding these losses is to have the body inclosed by another and to have the space between completely exhausted of air. By using a Dewar flask (see page 228), liquid air and hydrogen may be kept for hours in a room at ordinary temperatures.

Illustrations. — The fact that metals conduct well is shown by countless experiments. Thus, if a piece of wire gauze with fine meshes is lowered over a flame, *e.g.* one from a Bunsen burner, the latter burns below the gauze only; because the molecules of the gas as they pass through the wire meshes lose so much of their heat energy by conduction to the outer portions of the gauze beyond the flame that the temperature of the gas as it rises through the gauze is lower than that at which it burns. However, the temperature of the gauze gradually rises, owing to the flame, and as soon as the temperature of combustion is reached, the flame will burn on both sides of the gauze. Or, if the gas is turned on through the burner, but is not lighted, and the gauze is held close to the burner, the gas rising through the gauze may be ignited by a match, but the flame will not strike back below it. (This is the principle of the miner's safety lamp invented by Sir Humphry Davy.) Again, a bright luminous flame may be made smoky by bringing a large piece of metal close to it, so as to conduct off the heat energy and thus lower the temperature. The cracking of a tea cup or tumbler when hot water is poured into it is due to the sudden expansion of the inner surface before the outer one has time to be affected; this may often be prevented by putting a silver spoon (not a plated one) in the cup or tumbler and pouring in the water along it; the silver is such a good conductor that it prevents the temperature from rising too high at once.

Conduction in a Gas. — The process of conduction in a gas is evidently simply the redistribution of the kinetic energy of the particles. When the temperature is high at one point in the gas, the kinetic energy there is great; and so, owing to the increased velocity of these particles, this energy is communicated to the neighboring ones. On the assumption that a gas behaves like a set of elastic spheres, we can deduce a value for the conductivity in terms of the mean free path, etc. (See page 202.)

The conductivities of a few bodies at 0° C. are given in the following table, in which the heat unit is a calorie and the C. G. S. system is used:

Silver	1.096
Copper	0.82
Aluminium	0.34
Zinc	0.307
Iron	0.16
Mercury	0.0148
Water	0.0012

Radiation

Radiation as a Wave Motion. — When one's hand is exposed to sunlight, a sensation of hotness is perceived; similarly, if a body is brought near a flame, — even when not above it, — its temperature rises, or if brought near a block of ice, its temperature falls. There is neither convection nor conduction involved in these changes of temperature, yet heat energy is being gained or lost. The process is called "radiation." Boyle noted as early as the seventeenth century that it went on through a vacuum, and this fact is proved also by the heating action of the sun which we observe here on the earth. When we discuss, in Chapters XVI *et seq.*, the phenomena of waves and show how wave motions may be detected, it will be proved that this process of radiation consists in the motion of waves in the ether, *i.e.* in the medium which occupies space when ordinary matter is removed and which permeates ordinary matter as water does a sponge or as air does the stream of motes revealed by a beam of sunlight entering a darkened room. Without going into details in regard to waves, several facts may be mentioned which are familiar to every one from observations of waves on lakes or the ocean, or of waves along a rope. One is that in wave-motion we do not have the advance of matter, but the propagation of a certain disturbance or condition; each particle of matter makes oscillations about its centre of equilibrium, but does not move away from this as the wave itself advances; and therefore by the "velocity of waves" is meant the distance this disturbance advances in a unit of time. (Consider the waves produced in a long, stretched rope when one end is shaken sidewise.) Thus, in order to produce waves, there must be some centre of disturbance or vibration, and this centre is giving out energy, for it is evident that a medium through which waves are passing has both kinetic and potential energy. Thus, as waves advance into a medium, energy is carried forward, owing to the action of the particles of the medium on each other. We say, then, that "waves carry energy," although of course this energy is associated with the material particles. If wave motion ceases gradually as the waves enter a different medium (*e.g.* if a stretched rope is so arranged as to pass through some viscous liquid, waves sent along it will cease when they enter the liquid), this medium is said to "absorb" the waves; it gains the energy which the waves carry. Further, waves are of different lengths, depending upon the nature of the disturbing vibration; to produce short waves, that is, waves in which the distance from crest to crest is short, requires very rapid vibrations; while long waves are due to slow vibrations. We know, too, that waves suffer reflection, as is seen when water waves strike a large pier with a solid wall. If several waves are passing through the same medium at the same time, the resulting motion is the geometrical sum of the individual waves.

Production of Radiation. — When ether waves are discussed, it will be shown that they have a velocity of 3×10^{10} cm. per second, or about 187,000 mi. per second, and that they are known to have lengths varying from $\frac{1}{1000000}$ cm. up to many kilometres, depending upon the frequency of the vibrating centre where they are produced. It will be shown presently that all portions of matter, whatever their temperature, are producing spontaneously waves in the ether, whose lengths are so small as to be comparable with the size of molecules. The exact mechanism of this is not known; but it is clear that there must be some *mechanical* connection between the ether and the minute particles of matter, and that these last must be making exceedingly rapid vibrations. If the wave length is called l and the velocity of the waves v, the number of vibrations in a unit of time is $\frac{v}{l}$, because during each vibration the waves advance a distance l; and so, if there are n vibrations in a unit of time, the waves advance a distance nl in that time, or $v = nl$. If, then, there are waves in the ether whose length is $\frac{1}{10000}$ cm., the number of vibrations per second is $3 \times 10^{10} \times 10^4$ or 3×10^{14}, *i.e.* 300 trillions. Consequently these vibrating particles are thought to be the infinitesimal parts of a molecule. Our conception, then, of the structure of ordinary matter is as follows: it consists of molecules which are moving to and fro, vibrating about centres of equilibrium in solids, or traveling from point to point in a fluid, and at the same time the parts of the molecule are vibrating and producing the waves in the ether — this is similar to the case of a man moving a ringing bell, for the bell moves as a whole and its parts are making vibrations. Ether waves are also produced during certain electrical changes, as will be shown later; and they are short if the body experiencing this change is small, but long if the body is large.

Measurement of Radiation. — In order to study the nature of these waves in the ether and the connection between them and the material bodies which produce them, it is necessary to have some instrument which will detect their presence and measure the energy they carry. To do this some body must be found which absorbs the waves, for then some change which can be observed will be produced in it, depending upon its own properties and the length of the waves. Thus, if the waves are short, they may produce vibrations in the particles of the molecules in accordance with a simple mechanical principle known as that of "resonance." This is illustrated by a boy setting in motion another who sits in a swing; this has a natural period of vibration, and may receive a large amplitude if a series of pushes are given each time the swing passes through its lowest point, going in the same direction; but if the pushes are given at irregular intervals, one may neutralize another; so the force applied must have the same period as the natural period of the swing. Similarly, if the waves of a certain period enter a material body, the particles of whose molecules have a natural period the same as this, they will be set in vibration by the waves, and will therefore absorb them, gaining their energy. (If the waves have a much longer period, they may produce electrical changes in the body.) If the energy of the waves is absorbed by the particles of the molecules, further changes will occur, determined by the nature of the molecules and its parts. Thus certain ether waves falling upon the eye produce changes which result in vision; again, if certain ether waves are absorbed by photographic films, molecular changes go on which may be detected later; but in general the energy gained by the particles of the molecules is diffused among the molecules themselves, and is manifested by the appearance of heat effects. To measure the energy of the radiation, the absorbing instrument should absorb all the energy, and the corresponding change produced should be one which is in direct proportion to the energy added to the body. It is found (see below) that polished bodies like bright metallic surfaces absorb very little energy, while rough, blackened ones absorb nearly all the incident radiation, if the waves are not too long. So ordinary radiation produced in the ether by material bodies is detected and measured by instruments which are sensitive to heat changes, and whose surfaces are covered with a layer of lampblack, or copper oxide or platinum black (finely divided platinum).

It is worth while to describe a few of these instruments briefly. An ordinary mercury thermometer with its bulb blackened has been used; but a more sensitive instrument is a platinum resistance thermometer, with its strip of platinum blackened; this is called a "bolometer," and was invented by Professor Langley of the Smithsonian Institution. Another instrument is the thermocouple; in some cases the electric current produced, as the temperature of one junction is raised, is measured by a galvanometer, while in other instruments, called "radio-micrometers," and invented by Professor Boys, the wire in which there is the junction, and which carries the current, is suspended between the poles of a magnet in such a manner that it rotates when there is a current in the wire. The most sensitive of all instruments, probably, is the "radiometer," which is a modification of the instrument invented by Sir William Crookes and described on page 204; these alterations are in the main due to Professor E. F. Nichols, now of Columbia University. In its present form this instrument consists essentially of an exhausted glass bulb closed at one point by a window of fluorite, — which is particularly transparent to ether radiations of all lengths, while glass is not, — and containing a fine vertical quartz fibre which carries a horizontal arm; to each end of this is attached a thin piece of mica, polished on one side and blackened on the other. The blackened face of one mica disk comes opposite and parallel to the fluorite window; so, if radiation enters this, it falls upon the blackened face of the mica, whose temperature therefore rises and which then moves backward. This motion twists the quartz fibre; and when the torsional moment of reaction of the fibre equals the moment due to the "repulsion" of the blackened disk occasioned by its rise in temperature, everything comes to rest. The angle of deflection of the horizontal arm measures, then, the intensity of the radiation.

Another class of instrument must also be used in order to describe radiation from any source; this is one which analyzes it, and so distributes it that waves having different wave lengths proceed in different directions and may be studied separately. This process is called "dispersion" and is illustrated by the action of a glass prism on the light from a lamp.

Radiation Spectra; Energy Curves. — Using these instruments, certain facts have been established. All material bodies in the universe, so far as we know, are producing waves in the ether. Solid and liquid bodies emit waves of all wave lengths between certain limits, whereas gases emit trains of waves of definite wave lengths. The emission of a solid or liquid depends largely upon the condition of its surface, other things being the same. A polished metallic surface emits very little radiation, *i.e.* the energy of the radiation is small; whereas, rough or blackened surfaces emit a great deal. (This is the reason why stoves, steam pipes, etc. are blackened.) Again, the amount of the radiation from a body depends largely upon its temperature. As this is raised, the energy carried by each train of waves of a definite wave length increases; but this increase is greater for the short than for the long waves. This fact can be represented by a graphical method. Let two axes be drawn, distances

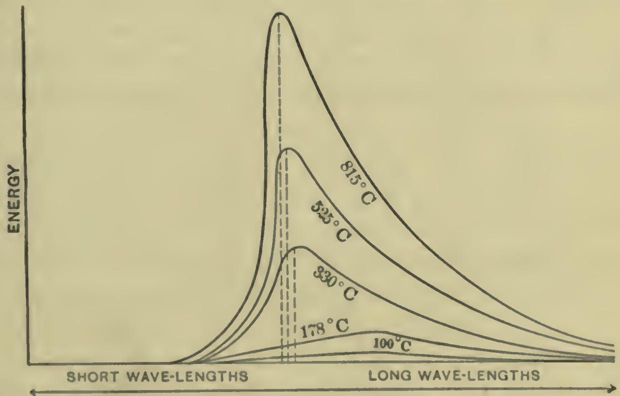


FIG. 131. — Radiation from blackened copper.

along the horizontal one to represent the wave lengths of the component waves, vertical distances to represent quantities of energy carried by the individual trains of waves. Several curves are given of the radiation from blackened copper at different temperatures. It is seen that these curves are in accord with the statements made above. If we consider any individual wave length at the extreme ends of the curves, it marks evidently the limiting power of the measuring instrument used; and therefore waves whose energy at one temperature of the body is so small that they cannot be detected may be so intense at a higher temperature as to permit of observation. For instance, the longest waves which affect the human eye in such a manner as to produce the sensation of light are those that cause the sensation we call *red;* therefore, if the temperature of a solid body, *e.g.*, a piece of iron, is raised, the experiment being performed in a darkened room, the solid is invisible (except for stray reflected light) until the temperature becomes so high that the energy of the waves whose length corresponds to "red light" is sufficiently intense to affect our eyes. (Actually, the fact must be taken into account that the human eye is not sensitive to all colors alike, and that if the light of any color is feeble, the eye perceives "gray.") The body is now "red hot"; and as the temperature rises still higher, its color changes continually, and finally it appears white and is said to be "white hot."

Laws of Radiation. — Careful observations upon the radiation of a blackened body have shown a most intimate connection between the total quantity of energy emitted and the temperature. If Q is the energy emitted at a given temperature, and T the absolute temperature, *i.e.* $T = t°\text{C.} + 273$, $Q = cT^4$, in which c is a factor of proportionality.

This statement is called "Stefan's law," having been first proposed by him. Again, there is a connection observed between the temperature of a blackened body and the wave length of the train of waves which carries more energy than any other train, *i.e.* the wave length which corresponds to the highest point of an energy curve for that temperature, as shown on page 295. This relation is due to Weber; and calling T the absolute temperature of the body and L_m the wave length just defined, $TL_m = a$, where a is a constant. These two laws, which have been verified over wide ranges of temperature by most painstaking investigations, offer convenient means of obtaining the temperature of bodies when they are so hot as to render it inconvenient to use ordinary means. If we *assume* that the laws are true for temperatures which are higher than those for which they have been verified, we may assign them numbers. (In this way, making the above *assumption*, the temperature of the sun is observed to be about 5700° C.)

Prévost's Law of Exchanges. — It is thus seen that the radiation of any body is independent of other neighboring bodies, because it depends upon the vibrations of its own minute particles. So, if two bodies are associated in such a manner that one receives the radiation of the other, each radiates independently; and the temperature of either one will fall if it radiates more energy than it absorbs. This principle of the complete independence of the radiation of two bodies was first stated by Prévost (1792) in his "Law of Exchanges," which is equivalent to the above.

Newton's Law of Cooling. — If a body is surrounded by an inclosure at a lower temperature, it loses more heat energy than it receives; and, if the radiation from the two bodies is that which is characteristic of a blackened body, this *net* loss may be expressed at once. Let T_1 and T_2 be the absolute temperatures of the body and the inclosure; then the heat energy lost by the former diminished by that received from the inclosure is $c(T_1^4 - T_2^4)$. This equals

$$c[(T_1 - T_2)(T_1^3 + T_1^2T_2 + T_1T_2^2 + T_2^3)];$$

and, if T_1 is only slightly greater than T_2, it may be written $4\,cT_1^3(T_1 - T_2)$. So the net loss in heat energy varies as the difference in temperature between the body and the inclosure. This is called "Newton's law of cooling," and it is true of other bodies than "black" ones for small differences in temperature.

Absorption

Reflection and Absorption. — When radiation falls upon a body, some is absorbed, some is reflected, and some is transmitted. A body which allows waves of a certain wave length to pass through it is said to be "transparent" to them; but no body is perfectly transparent to any waves; if it is sufficiently thick, it will absorb them. In a thin layer, however, a body may absorb certain waves completely and may transmit others comparatively freely. Thus, ordinary glass permits those waves to pass which affect our sense of sight, but either absorbs or reflects other waves which are shorter or longer.

This is the explanation of the action of the glass roof of a greenhouse. The "visible waves" from the sun are transmitted through the glass and are then absorbed by the black earth or the green leaves. The temperature of these is raised,—but not sufficiently to make them self-luminous,—and they radiate waves which are so long that they are reflected by the glass. Thus the energy which enters through the glass is trapped and stays inside; consequently the temperature is raised.

A body which transmits comparatively freely those waves which carry the greatest amount of energy is called "diathermous"; but the word is not often used, because of its indefinite character.

If we wish to compare the properties of reflection and absorption, it is best to consider a body which is so thick as to transmit no waves. It is then at once evident that if a body absorbs well it must be a poor reflector, and conversely. Thus, a blackened surface absorbs well and reflects poorly; while a polished metal reflects well and absorbs hardly at all. In order to secure what is called "regular" reflection, as from a mirror, not alone must the body be itself large in comparison with the length of the waves, but its surface must be smooth to such an extent as to have no irregularities so large; otherwise the different portions of the surface reflect the waves in different directions and so scatter them. Under the above conditions a body reflects at its surface waves of all lengths to a greater or less extent; but in every case certain waves enter the body, although their intensity may be very small.

Absorption. — Let us consider the process of absorption more closely. When ether waves fall upon a body, certain particles in the molecules are set in more violent vibration by resonance, and thus the waves lose energy. In some bodies these vibrating particles emit waves immediately, without the temperature sensibly rising. This is the case with pieces of fluor spar, thin layers of kerosene oil, and with a few other bodies, as will be shown later under Fluorescence and Phosphoréscence. In other bodies the absorbed energy is distributed among the molecules and becomes apparent in heat effects. This absorption, where the energy goes into heat effects, is called "body absorption." Many bodies absorb only waves of definite wave-length, and transmit others. They are said to have "selective" absorption.

Metals and substances that have strong selective absorption *reflect* certain waves more intensely than others; they are said to have "selective" reflection. Thus, the reason why gold appears yellow to our eyes is because, when viewed in ordinary white light, besides the waves that are reflected at the surface and that would make the gold appear white, there are certain waves, of such a wave length as to produce the sensation of yellow, that are reflected more intensely than the others. Those waves which enter the gold are absorbed in the surface layer of molecules and produce heat effects.

It is evident, then, that if white light is reflected again and again from a series of gold surfaces, in the end the only waves which will leave the last surface will be those which produce in our eyes the sensation of yellow. The waves which leave the last surface after a great number of reflections from the same material are called the "residual" ones.

If one looks at a bundle of needles in white light, the points being turned toward the eye, they appear black, because the waves are reflected to and fro from needle to needle, but are continually getting weaker and weaker and being deflected down the needle; thus no waves come back to the eye, and the points appear black. Finely divided silver and platinum appear black for the same reason.

Connection between Radiating and Absorbing Powers. — Since absorption is due to resonance, it is simply a restatement of this to say that a body absorbs to a marked extent waves of the same period as those which it has the power of emitting. But we can say more, if we consider the intensities of the waves absorbed and emitted, and if we assume that there are no chemical or other molecular changes in the body. This excludes fluorescence, phosphorescence, etc. If several bodies at different temperatures could be inclosed inside a vessel which absolutely prevents any heat energy from entering or leaving, and which keeps a constant volume (so that no external work is done), there is every reason for believing that equilibrium would finally be reached, but not until the temperature of all the bodies inside was the same as that of the walls of the vessel. When this is the case, each body must be absorbing and turning into heat effects as much energy as it emits, *provided there are no chemical or other energy changes*, otherwise its temperature would change. That is, the absorbing power of a certain body at a definite temperature exactly equals its emissive power at that same temperature; where by absorbing power we refer to body absorption. (In general language, a body which absorbs well, in the sense of transforming radiant energy into heat energy, radiates well; *e.g.* a blackened surface.) We can imagine, moreover, a body in the vessel described above, which is entirely inclosed by some envelope which allows to pass through it waves of only one wave length; therefore, when equilibrium is reached, the body inside must radiate as much energy in the form of waves of this wave length as it absorbs. Consequently, the amount of energy of a definite wave length which a body emits at a given temperature equals exactly the amount of energy in the form of waves of this same wave length which it absorbs at that temperature. In other words, the absorptive power of a body at a certain temperature equals both in *quantity* and *quality* its emissive power at that same temperature. If at ordinary temperatures a body appears black when viewed in white light, it is owing to the fact that it absorbs those waves which affect our sense of sight; and, if raised to such a temperature as will enable a body to emit such short waves, it will emit them and so shine brightly in a darkened room. Similarly, if a body appears red at ordinary temperatures when viewed in white light, it is because it absorbs all waves except those which produce in our eyes the sensation of red; these are either transmitted or are reflected out from the interior by some small foreign particles. (Thus, a colored liquid appears perfectly black except by transmitted light, if it is entirely free from small solid particles; but, if a minute quantity of dust is stirred in it, it appears colored when viewed from any direction.) Then, if such a red body is heated until its temperature is sufficiently high, it will emit all the waves except those which correspond to the sensation of red, and so, if viewed in a dark room, will appear bluish green.

This law connecting radiation and absorption was first stated by Balfour Stewart, but was discovered independently by Kirchhoff. The latter, however, in expressing it, did so in a more mathematical form. He took as a standard of absorption a hypothetical body, which is called a "perfectly black body" and which is defined to be such a body as will absorb and turn completely into heat effects all radiations which fall upon it. (Any non-reflecting body, if sufficiently thick, is such a body.) We can approximate to such bodies experimentally by using lampblack or platinum black as the surface layer. (It is seen, too, that the radiation inside any hollow inclosure, provided it is rough, is that which is characteristic of a perfectly black body at that temperature, because after a sufficient number of irregular reflections any train of waves will be totally absorbed, so the walls of the inclosure finally produce the effect of a black body.) It will be shown later that radiation may be produced and varied by other means than by raising or altering the temperature of the radiating body; but this law of Stewart and

Kirchhoff refers only to radiation that is due to the same cause which conditions the temperature of bodies, and to absorption that results in heat effects.

Atmospheric Absorption. — One important case of absorption of radiation is that of the solar rays incident upon the earth. As these pass through the earth's atmosphere, a certain percentage of their energy is reflected by the floating particles and drops in the air, and also by the molecules of the air themselves; and so does not reach the earth. There is also true body absorption as the waves traverse the atmosphere. The energy that is not reflected or absorbed reaches the earth and is there almost completely absorbed and spent in producing heat effects. The earth itself is therefore also radiating energy, and this again passes through the atmosphere and is partially absorbed. If there are clouds, this radiation from the earth itself is absorbed by them, and they radiate a certain proportion back toward the earth. Our appreciation of the temperature of the air in which we live depends largely upon the quantity of heat energy absorbed by the air, not so much upon the radiation received by us directly from the sun. Thus we see the reason why the temperature is lower upon mountain tops where the air is rare and so absorbs poorly than at sea level where the air is denser and absorbs more.

CHAPTER XV

THERMODYNAMICS

Nature of Heat Effects. — Throughout the previous chapters we have assumed that heat effects are due to work done against the molecular forces of a body and that for a definite amount of energy received the same effect is produced regardless of the source of the energy. These assumptions are justified by countless experiments, some direct and some indirect. Thus Joule, in a series of investigations beginning about 1843 and lasting over forty years, caused heat effects to be produced in many different ways; compression of gases, friction of various kinds, conduction of an electric current through a wire, chemical reactions, etc. The only possible explanation of the results of these experiments is the assumption that heat effects are due to energy being given the molecules of the body and are proportional to the amount received. Thus, if a certain amount of heat energy received from friction is used to boil some water or to melt some ice or to raise the temperature of water, we can, by allowing this water to cool through this temperature range, obtain an amount of energy which will melt an equal amount of ice, etc. — it being remembered that heat energy passes from high to low temperature. Again, when the energy received from any two sources of heat — for instance, a candle and the sun — is compared, if under any condition they raise the temperature of the same amount of water through the same range of temperature, they will melt the same amount of ice, or boil the same amount of water, etc. Therefore the production of heat effects depends upon the energy received, not upon the temperature or condition of the source of the energy.

Mechanical Equivalent of Heat. — As stated before, the practical unit of heat energy is that required to raise the temperature of one gram of water from 15° to 16° C. The value of this in ergs, or the "Mechanical Equivalent of Heat," has been determined by different observers. For many years it was thought that the specific heat of water was the same at all temperatures; and in the early work no distinction was made between the amounts of energy required to raise the temperature of water between different degrees.

Robert Mayer calculated the mechanical equivalent from experimental data secured by other observers on the heat energy required to raise the temperature of a gas at constant volume and at constant pressure. We have shown that, on the assumption of no internal forces, if J is the "mechanical equivalent," $J(C_p - C_v) = R$ for a gas; and hence knowing C_p, C_v, and R for a gas, J may be calculated.

Joule, and afterwards Rowland, used the method of turning a paddle rapidly in water, and measuring the mechanical work, the quantity of water, and the rise in temperature. The mechanical work was measured by a simple dynamometer method (see page 121); and in Rowland's work most accurate results were obtained. In fact, it was he who first measured the variations in the specific heat of water. This method was later modified by Reynolds and Moorby, who used revolving paddles to raise the temperature of a known quantity of water from 0° to 100° C.; and so their results are independent of thermometers.

The most recent work, and the best, has been done by Griffiths, Schuster and Gannon, and Callendar and Barnes, using electrical methods. It will be shown later that when an electric current passes through a conductor, heat-energy is produced; the amount of this depends upon the electrical quantities involved, all of which can be measured with great exactness, much more so than mechanical power. If i is the strength of the current; R, the resistance of the conductor; t, the time the current flows; the heat energy produced is i^2Rt expressed in ergs, if the electrical units are properly chosen. Therefore, calling the measured heat energy H,

$$JH = i^2Rt.$$

In practice the current is passed through a wire which is in contact with water; so H is measured directly. Thus J may be determined. (The weak point in the method is the uncertainty as to the value of the electrical units in terms of mechanical ones.) As a result of all of these experiments we know that to a high degree of approximation the work required to raise the temperature of one gram of water from 15° to 16° C. is 4.187×10^7 ergs.

The First and Second Principles of Thermodynamics. — The statement, made so frequently, that the conservation of energy includes heat effects is sometimes called the "first principle of thermodynamics," — the science which applies principles of mechanics to heat phenomena. The additional statement that heat energy of itself always flows from bodies at high to those at low temperatures is called the "second principle of thermodynamics." The founder of the science of thermodynamics was Sadi Carnot (1796–1832). He was interested in the practical problem of increasing the efficiency of the steam engine. The action of this engine is complicated. The steam — or "working substance" — receives heat energy in the boiler and loses it while expanding, owing to conduction of the cylinder walls, and also while being condensed. It does work during part of the process, and has work done on it during the rest. The efficiency of the process has been defined to be the ratio of the *net* external work done in any one stroke, W, to the heat energy received in that time from the boiler, H. Carnot conceived a simpler type of engine, which has been called

"Carnot's engine." In it the processes are considered as taking place as follows: the working substance expands slowly from A to B, the temperature remaining constant; it expands from B to D in such a manner that no heat energy enters or leaves, and so its temperature falls; it is compressed from D to C at a constant temperature; and finally it is compressed from C to A under conditions such that no heat energy enters or leaves, and so its temperature rises to its initial value. The curves $\overline{AB}$ and $\overline{CD}$ are then "isothermals"; $\overline{AC}$ and $\overline{BD}$ are "adiabatics." In this cycle the working substance in passing from A to B is supposed to receive heat energy from some large reservoir at the temperature indicated by the isothermal, and to lose heat energy in passing from D to C to another large reservoir at the temperature of this lower isothermal. The external work done equals the area of the curve $ABDC$. If H_1 is the heat energy received, and H_2, that lost, and W the external work done, $W = H_1 - H_2$ by the conservation of energy (using the same units for heat energy and external work), and the efficiency is $\frac{H_1 - H_2}{H_1}$. In discussing the action of this engine Carnot was led to several most important conclusions which Clausius and Lord Kelvin have shown to be rigid consequences of the two principles of thermodynamics. One of these was that the efficiency of his ideal engine was independent of the nature of the substance used to work it: steam, water, alcohol, etc. Another was that the efficiency of his engine varied directly as the difference in temperature between the two reservoirs referred to above.

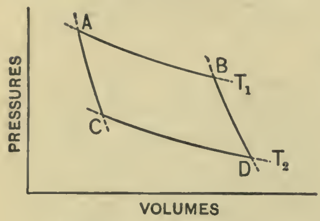


FIG. 132. — A Carnot cycle.

Absolute Thermometry. — This last fact led Lord Kelvin, then William Thomson, in the year 1848 to propose a system of thermometry depending upon the use of Carnot's engine; this has the great merit of being independent of the substance used in the thermometer (or engine). Thomson's system of "absolute" thermometry, as it is therefore called, is equivalent to defining the ratio of the temperatures of two bodies as equal to the ratio of the quantities of heat energy received and given out by a Carnot engine working between these two temperatures. Thus, if in the above description of an ideal Carnot's engine the ratio of the temperatures on Thomson's scale of the hot and cold reservoirs is written $\frac{T_1}{T_2}$,

$$\frac{T_1}{T_2} = \frac{H_1}{H_2}.$$

Hence the efficiency is $\frac{T_1 - T_2}{T_1}$. Since it is impossible to have an efficiency greater than unity, there is a minimum value of temperature, that for which $T_2 = 0$; for, if T_2 had a negative value, the efficiency would be greater than unity. This minimum temperature is called "absolute zero."

Thomson's definition of absolute temperature does not specify the "size" of a degree, but simply the *ratio* of two temperatures: we can choose the degree to suit ourselves. Let us agree to use the Centigrade scale, so that if T is the temperature of melting ice, $T + 100$ is that of boiling water. By a series of most ingenious experiments Thomson showed that this system of temperature agrees most approximately with that which we have been using, namely, that of a constant-pressure hydrogen thermometer on the Centigrade scale, if we add to each temperature reading the reciprocal of the coefficient of expansion of the gas, *i.e.* 273 approximately. (This is what we have called on page 240 "absolute gas temperature.") Thus, if large quantities of boiling water and melting ice could be used as the reservoirs between which a Carnot engine worked, the quantities of heat energy received and given out, H_1 and H_2, would have such values that their ratio $\frac{H_1}{H_2}$ almost exactly equaled $\frac{373}{273}$. It can be proved that, if the gas had no internal forces and obeyed Boyle's law exactly, the agreement between Thomson's absolute system of thermometry and that of a gas thermometer as above described would be perfect. Therefore, in the formula for a gas, $pv = RmT$, T is very nearly equal to Thomson's absolute temperature. The slight differences between this system and that we have been using were determined by Thomson and Joule; and in all exact work in Heat this correction is made to the gas temperatures.

There are several ways of determining the absolute temperature of a body and of calculating the value of absolute zero on the system of thermometry ordinarily used. The accepted value is $-273°.10$ C.

Historical Sketch of Heat Phenomena

It was believed by the Greek philosophers that all the phenomena of a material body which we associate with the word "heat," such as expansion, change in temperature, boiling, etc., were due to the addition to the body of a *substance;* but to Newton and his immediate predecessors and associates it seemed clear that in some way they were due to motion of the parts of the body. Thus Boyle gave a correct explanation of the heat effects observed when a hammer strikes a nail. The materialistic theory of heat, however, was again proposed, and prevailed for nearly two hundred years. Its great defenders were Gassendi (1592–1655), Euler (1707–1783), and Black (1728–1799). Even as late as 1856, in the eighth edition of the *Encyclopædia Britannica*, this theory is offered as the accepted explanation of heat phenomena. One great reason why this theory was so universally accepted was because it was so analogous to the accepted explanation of combustion. Stahl (1660–1734), who was professor at Halle, advanced the theory that during the process of combustion a material substance, called "phlogiston," was given off; and this idea persisted until the work of Lavoisier, about 1800, and even later. Fantastic theories in regard to the properties of phlogiston and of the substance heat (or "caloric") were of necessity brought forward in order to account for the observed facts.

Joseph Black showed that when two bodies at different temperatures were brought together he could speak of "quantities of heat" leaving or entering the bodies; and we owe to him our ideas and methods in regard to specific heats. Black considered two kinds of effects when "heat" was added to a body: if the temperature was raised, the heat was called "sensible," and it was supposed to be free in the body; but if the temperature did not change, the heat was said to be "latent," and it was supposed to form some kind of a compound with the molecules of the body changing their state. But the experiments of Rumford, in 1798, and of Davy, in 1799, convinced nearly every one that heat effects in a body were due primarily to the transmission of motion to its minute parts. Rumford showed that in such a process as that of boring out a brass cannon, "quantities of heat" could be produced, limited only by the amount of work done. Similarly, Davy arranged an apparatus which caused one block of ice to rub violently against another, and showed that the quantity of ice melted varied directly with the work done.

The first one, however, to express clearly the belief that heat effects were due entirely to the addition of energy to the small parts of a body was Robert Mayer, in 1842. He was followed by Joule, in 1843, and later by Helmholtz, in 1847. By the epoch-making researches of Joule, the principle of the conservation of energy — a phrase of Rankine's — was soon extended so as to cover all heat phenomena.

The fact that "radiation" is a phenomenon due to wave motion in the ether, of exactly the same nature as that which produces the sensation of light, has been established by a long line of investigators. William Herschel showed, in 1800, that there were rays in the solar spectrum invisible to the eye and yet having the power of affecting a thermometer. Herschel speaks of these rays as subject to the laws of reflection and refraction; and this fact was fully established by Melloni some thirty years later. It was proved in the following years that these rays could be diffracted, be made to interfere, be polarized, etc.; and that, in short, they were due to waves in the ether.

BOOKS OF REFERENCE

EDSER. Heat for Advanced Students. London. 1901.

PRESTON. The Theory of Heat. London. 1894.
An advanced text-book, containing extended descriptions of the experiments which form the basis of our knowledge of heat quantities.

TYNDALL. Heat a Mode of Motion. London. 1887.
In this the author describes a series of experiments illustrating the fact that heat effects are due to energy changes.

BRACE. The Law of Radiation and Absorption. New York. 1901.
This contains the original papers of Prévost, of Balfour Stewart, and of Kirchhoff.

AMES. The Free Expansion of Gases. New York. 1898.
This contains the papers of Gay-Lussac and Joule on the expansion of gases through small openings.

RANDALL. The Expansion of Gases by Heat. New York. 1902.

TAIT. Heat. London. 1884.

MAXWELL. Theory of Heat. London. 1892.

YOUMANS. The Correlation and Conservation of Forces. New York. 1876.
This contains reprints of papers by Helmholtz, Mayer, and others on the Conservation of Energy.

STEWART. The Conservation of Energy. New York. 1874.

GRIFFITHS. Thermal Measurement of Energy. Cambridge. 1901.
This gives a most interesting description of various methods for the measurement of heat energy.

VIBRATIONS AND WAVES

CHAPTER XVI

WAVE MOTION

General Description. — One of the most important phenomena in nature, and one that is of most frequent occurrence, is the transmission of energy from one point to another by what is called "wave motion." This motion is illustrated in many ways: if a stone is dropped in a pond, waves are produced on its surface, which do work on any movable object which they meet; a vibrating bell produces waves in the air, which do work in a similar way, or by bending a suitably stretched membrane such as the drum of the human ear or the diaphragm of a telephone instrument; if one end of a long, stretched rope is fastened to a movable object and if the other is given a sudden sidewise or lengthwise motion, a disturbance will pass along the rope and will do work on the object; etc. In all these cases it is evident that there is a vibrating centre which produces motions in those portions of the surrounding medium immediately in contact with it; these portions affect those next them, etc. The vibrations at the centre of disturbance must be of such a nature as to produce in the medium a displacement or change which can be propagated by it. Thus, if the hand is moved through air or water, or if a pendulum vibrates in air or water, waves are not produced, because the fluid flows around the obstacle as it moves through it, and is not compressed; in order to produce waves, the vibration must be so rapid or the motion so sudden that the fluid does not have time to flow, and is therefore compressed on one side and expanded on the other. Again, any *sidewise* motion in a fluid of an object like a thin board, however sudden, would not produce waves, because in a fluid there is no elastic force of restitution when one layer is moved over another. In order that a medium should carry waves, there must be forces of restitution called into action when its parts are displaced; for these forces are due to the action of the neighboring parts on each other; and owing to the reaction of the displaced parts on those in contact with them the latter are displaced also, and so waves are produced and propagated. If these conditions as to the medium and the centre of disturbance are satisfied, and if the motion of this centre is a vibration, or a series of vibrations, the various portions of the surrounding medium will in turn be set in vibration. If the motion of the vibrating centre ceases, that in the medium will persist until forces of friction (or other causes) bring it to rest. In all these cases it is clear that the particles of the medium vibrate, but do not advance with the waves; a certain "condition" moves out from the centre.

"Water Waves" and Elastic Waves. — There are several kinds of forces of restitution that enable a medium to propagate waves. If the surface of a pond is disturbed at one point by the motion up and down of a stick dipping in the water, waves spread out, owing to the fact that the force of gravity tends to maintain a level surface. Again, if a vertical cord, like a fishing line, is moved sidewise through the surface of a pond (or, if a quietly flowing river flows past a stationary vertical cord), short waves called "ripples" may be observed on the side toward which the cord is moving (or up stream in the other case). These waves are due to the force of restitution of surface tension.

The waves just described occur on the *surface* of a liquid; but any portion of matter that is elastic can also carry waves through its *interior*. Thus, fluids can propagate *compressional* waves, that is, waves produced by having a centre of disturbance where the fluid is compressed or expanded; as is illustrated by a bell vibrating in air or under the surface of a lake. Fluids cannot, however, propagate a distortional disturbance. An elastic solid body can carry both compressional and distortional waves, as is illustrated by a long, stretched wire. If this is disturbed at some point by a transverse vibration, or if it is twisted back and forward, there will be distortional waves; if it is pulled to and fro longitudinally, in the direction of its own length, at some point, there will be compressional waves. (This fact is also illustrated by the disturbances that we call earthquakes; for these are due to some great disturbance in the interior of the earth that produces both kinds of waves in the body of the earth. As we shall soon see, these two types of waves in a solid travel with different velocities; and this fact is observed in all earthquakes.) Compressional waves are often called "longitudinal," and distortional ones, "transverse," for obvious reasons.

Polarization. — One distinction between longitudinal and transverse waves, other than the one which gives rise to the names, is worth noting. A longitudinal wave, in which the particles of the medium move to and fro along the line of advance of the waves, appears the same to the eye from whatever side this line is viewed. But, since in a transverse wave the particles are vibrating in planes that are at right angles to the direction of propagation, it will as a rule appear different when viewed from different directions. Thus, if a transverse train of waves is produced in a long, stretched rope by moving one end up and down in a vertical line, the rope will at any instant have a sinuous form when viewed from the side, but will appear straight when viewed from above. In this case all the particles are vibrating in straight lines through which a plane can be drawn; and such a train of waves is said to be "linearly" or "plane polarized." Similarly, if the end of the rope is moved rapidly in a circle or in an ellipse, all the particles of the rope will in turn move in circles or ellipses; and the waves are said to be "circularly" or "elliptically polarized." Thus only transverse waves can be polarized; and, conversely, if in any wave motion we can detect polarization phenomena, we know that the waves must be transverse.

Intensity. — In all classes of waves it is at once evident that we are dealing with the propagation of energy. For wherever there is wave motion the moving parts have kinetic energy, and the existence of the waves presupposes forces of restitution, so when the parts are displaced there is potential energy. Therefore, as the waves spread out from a centre of disturbance, the medium into which they advance gains energy. If the cause of the waves is a temporary disturbance, any portion of the medium gains energy when the waves reach it and loses it again when the motion ceases, owing to the waves passing on. Thus a portion of the medium simply *transmits* the energy. It gains none permanently unless there are forces of friction when the particles of the medium move relatively to each other as the wave passes. (Of course if there are foreign bodies immersed in the medium carrying the waves, they may be set in vibration and may continue to vibrate after the wave passes; in which case this portion of space — not the medium itself — has an increased amount of energy in it afterward.) If at any point in a medium a plane having a unit area be imagined described at right angles to the direction in which the waves are propagated, the amount of energy transmitted through it in a unit time is called the "intensity" of the waves at that point. (If the waves are varying, the exact definition is as follows: if E is the energy transmitted in time t through an area A, the intensity is the limiting value of the ratio $\frac{E}{tA}$, as t and A are taken smaller and smaller.)

Detection of Waves. — The effect of the waves is perceived in two ways. If the medium is limited in one direction by some object, with which it is connected in such a manner that the vibration of this object would produce waves in the medium, this will be set in motion by the waves unless it is restrained by mechanical means. Thus, waves in the air set in motion the drum of the ear or the diaphragm of a telephone receiver. If the object is surrounded by the medium, but cannot move bodily with the waves, it may be set in vibration by them, owing to "resonance," a process which will be described in detail later. Thus, waves produced in the air by one tuning fork may set in vibration another one, if the periods of vibration of the two are identical; waves in the ether may set in vibration particles of matter, as described in the chapter on Radiation.

Other Kinds of Waves. — So far we have spoken of mechanical waves only, that is, waves in material media or in the ether, in which the disturbances are displacements of particles. But we can have many other kinds of waves, depending upon the property of the medium that is varied. Thus, if the temperature of one end of a metal rod is first raised gradually, then lowered, raised again, etc., we have a vibration of temperature; and, if we observe the temperature of any point in the rod not too far away from this end, we shall find that its temperature also rises and falls. Since it takes time for the conduction of heat, the temperatures at different points of the rod will not have their maximum values at the same time, so there is a wave of temperature in the rod. This is illustrated in the daily heating and cooling of the earth's surface as it is turned toward and away from the sun, also by the seasonal heating and cooling owing to the revolution of the earth around the sun. There are temperature waves, then, going down into the earth for a short distance. The daily wave is appreciable for a depth of about 2 or 3 ft.; the annual one for about 50 ft.

The *mean* temperature of the earth's crust increases gradually but continuously with the depth; the rate of increase varies greatly with the geological conditions, but is on the average about 1° C. for a depth of 28 m. This condition requires that heat energy should be continually flowing from the interior of the earth to the surface. From considerations based on this fact it is possible to make an estimate of the "age of the earth"; that is, the interval of time since the earth was in a liquid condition. This is probably about 50,000,000 years.

Similarly, if one end of a long electrical conductor, *e.g.* an ocean cable or a telephone wire, has its end suddenly joined to an electric battery, the effect is gradually felt along the conductor; and, if the electric battery at the end is varied, there is an "electric wave" in the conductor.

Again, if a metallic body is charged electrically, there are electric forces, so called, at points in space near by; so, if this electric charge is varied, these forces will vary; and, as they change, variations are produced at neighboring points. Therefore, when the electric charge on a body varies, that is, when there are "electric oscillations," waves of electric force are produced in the surrounding medium. The medium which carries these waves has been proved to be the same as that which carries the waves that affect our sense of sight; namely, the "luminiferous ether." It has been proved, too, that, wherever there are variations in the electric force, there are also variations in the magnetic force; so these waves are called "electro-magnetic." (It should be borne in mind that if we look upon matter as the fundamental concept in nature, then as soon as we are able to explain electric and magnetic forces as in some way due to the motion of matter, we shall be able to describe electro-magnetic waves in the ether as displacements of material portions of the ether. But, if an electric charge is the fundamental concept, as soon as we can explain the properties of matter as due to the motion of charges, we shall be able to describe all waves in matter in terms of electric forces.) These electro-magnetic waves may be detected by suitable means, as is shown by the various systems of "wireless telegraphy."

CHAPTER XVII

HARMONIC AND COMPLEX VIBRATIONS

We shall now proceed to discuss in detail the two fundamental features of waves: first, the properties of the centre of disturbance; second, those of the medium through which the waves pass. The effects produced when waves in the air or in the ether are perceived by our senses of hearing or of vision will be considered later in the sections devoted to Sound and to Light.

The Kinematics of Vibrations

Simple Harmonic Vibration. — A disturbance may be periodic or not; that is, it may after a definite period of time called the "period" be repeated identically, and again at the end of another period, etc.; or it may be irregular. Thus, the vibrations of a pendulum, of a tuning fork, of a violin string, etc., are periodic; while the motion of a piece of tin as it is "crackled," of two stones when struck together, or of one's hand as it is moved to and fro at random, are not periodic. The simplest case of periodic motion is that which is called "simple harmonic," and which is discussed on page 48. The period of this has been defined above; and the number of vibrations in a unit of time is called the "frequency"; this is, of course, the reciprocal of the period. The "amplitude" has been defined as one half the length of the swing, or as the value of the maximum displacement. Two harmonic motions having the same period and amplitude may yet differ in "phase"; that is, the instants at which they pass through their origins may be different.

Composition of Harmonic Motions. — 1. *In the Same Direction.* If a point is subjected to two harmonic motions, the resulting motion may be found by compounding the displacements geometrically at consecutive instants of time, or by simple algebraic processes. Thus, if the two harmonic motions have the same period and are in the same direction, they may be represented by $x_1 = A_1 \cos(nt - a_1)$ and $x_2 = A_2 \cos(nt - a_2)$ (see page 51); and the resultant motion is

$$x = x_1 + x_2 = A_1 \cos(nt - a_1) + A_2 \cos(nt - a_2).$$

By ordinary trigonometrical formulæ this takes the form

$$x = A \cos(nt - a),$$

where $$A^2 = A_1^2 + A_2^2 + 2 A_1 A_1 \cos(a_1 - a_2),$$

and $$\tan a = \frac{A_1 \sin a_1 + A_2 \sin a_2}{A_1 \cos a_1 + A_2 \cos a_2}.$$

This shows that the resulting motion has the same period as that of its two components, but a different amplitude and phase.

Graphical Methods. — If the two component harmonic motions are in the same direction but have different periods, the algebraic formulæ are much more complicated, but their resultant may always be found by a simple graphical method.

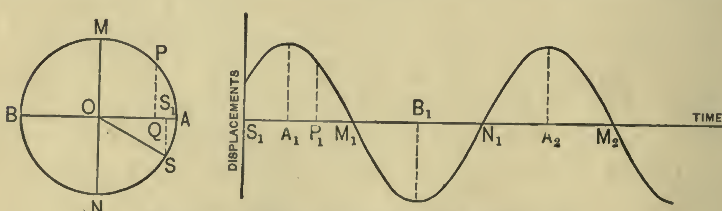


FIG. 133. — Graphical representation of harmonic motion.

Any harmonic motion may be represented by a curve drawn on a diagram whose axes are *intervals of time* and *displacements*. Thus, the motion $x = A \cos(nt - a)$ will be given by a curve, a portion of which is shown in the cut —

which is called the "sine curve." Thus, let, as on page 48, the harmonic motion be that of the point Q, the projection on the diameter of a point P, which is moving in a circle with constant speed. If, when we begin to count time, the point P is at S, the "initial" value of the displacement is the projection of $\overline{OS}$ on the diameter. From this time on the displacement assumes different values; $\overline{OA}$ is the greatest, $\overline{OB}$ is the least; when P is at M, the displacement is zero, etc. Therefore, if we erect at each point of the "axis of time" a line whose length equals the *displacement* of Q at that instant, and if we remember that displacements in one direction are positive, but in the other negative, we obtain a curve like that shown in the cut, which is known as a "sine curve." S_1, A_1, P_1, M_1, B_1, etc., indicate points corresponding to points S, A, P, M, B, etc., in the circular diagram. As the motion is periodic, the curve repeats itself.

This curve can be obtained practically by fastening a wire to the bottom of a heavy pendulum, and drawing under it, at right angles to

FIG. 134. — Sine curve drawn on smoked glass.

the plane of vibration, a piece of glass which has been blackened with camphor smoke, in such a manner that the point of the wire just scrapes off the black soot, thus leaving a trace on the glass.

The amplitude is the maximum value of x, *i.e.* the length of the line $\overline{OA}$.

The line $\overline{OS_1}$ gives the value of the displacement at the epoch of time when we begin counting, *i.e.* at $t = 0$. Hence $\overline{OS_1}$ equals $A \cos a$. (As the curve is drawn, $\overline{OS_1}$ is positive, hence a is an angle lying in the first or fourth quadrants.) So the length of this line varies with the phase of the vibration, other things being unchanged.

At the points M_1, N_1, M_2, etc., the displacement is zero; and if the corresponding instants of time are called t_1, t_2, t_3, etc.,

$$A \cos (nt_1 - a) = 0; \quad \therefore \cos (nt_1 - a) = 0.$$

$$A \cos (nt_2 - a) = 0; \quad \therefore \cos (nt_2 - a) = 0.$$

$$A \cos (nt_3 - a) = 0; \quad \therefore \cos (nt_3 - a) = 0.$$

$$\text{etc.} \qquad\qquad \text{etc.}$$

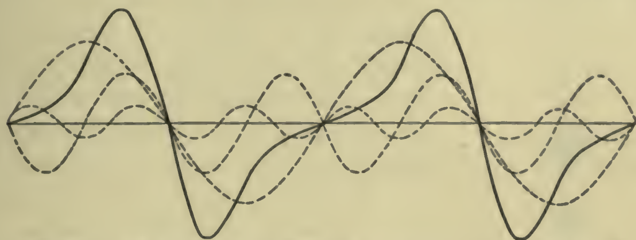
So if $nt_1 - a = \frac{\pi}{2}$, it follows that $nt_2 - a = \frac{3\pi}{2}$, $nt_3 - a = \frac{5\pi}{2}$, etc.; and $t_3 - t_1 = \frac{2\pi}{n}$. The interval of time taken for the displacement to traverse the curve from M_1 to M_2 is the period; and so we have a proof of the fact that $\frac{2\pi}{n}$ is the value of the period.

This is the length of the line M_1M_2; and therefore, if different harmonic motions have different periods, *i.e.* different values of n, they will have curves which cut the axis of time at points whose distances apart are different.

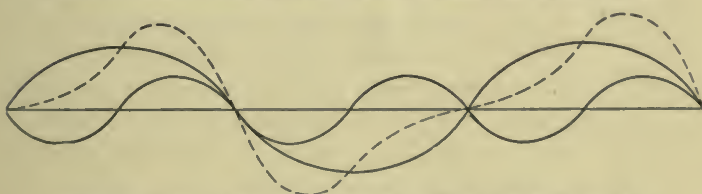
The method, then, of compounding harmonic motions which are in the same direction is to draw on one diagram the curves for each one and to superimpose them. This is done by adding the values of x which correspond to each value of t, remembering that, when the curve is below the axis, the values of x are negative. Illustrations of this process are given in the accompanying cuts.

This addition of harmonic motions in the same direction may be illustrated physically by suspending a series of simple pendulums one from another, as shown in the cut on page 322, and setting them in vibration through small amplitudes. The motion of the lowest pendulum is the sum of the harmonic motions which each of the pendulums would have by itself. It is at once seen how this varies if the component pendulums have different phases, periods, or amplitudes.

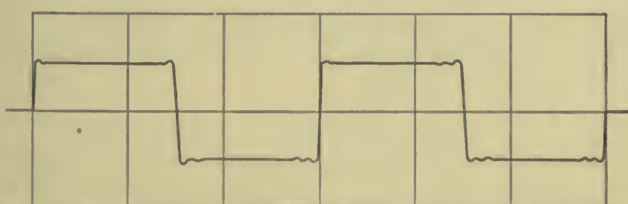
Complex Vibrations; Fourier's Theorem. — It is evident that if the component vibrations are numerous, the resulting vibration is, as a rule, exceedingly complicated. The study of the converse problem, that of separating a complex vibration into simpler parts, led Fourier to a most important and striking mathematical theorem: any complex motion can be



Composition of three harmonic vibrations whose periods are in the ratio 1 : 2 : 3.



Composition of two harmonic vibrations whose periods are in the ratio 1 : 2.



Composition of 100 harmonic vibrations whose periods are proportional to 1, $\frac{1}{3}$, $\frac{1}{5}$, etc., and whose amplitudes are proportional to 1, $\frac{1}{3}$, $\frac{1}{5}$, $\frac{1}{7}$, $\frac{1}{9}$, etc.

FIG. 135.

analyzed into simple harmonic motions of proper amplitudes whose periods are in the ratios of 1 : 2 : 3 : etc. That is, if the longest period is T, the others are $\frac{T}{2}$, $\frac{T}{3}$, etc. These component vibrations are called "harmonics."

There are other modes, however, of analyzing a complex vibration than by Fourier's theorem; and, if the component vibrations do not have frequencies in the above simple ratios, the one of least frequency is called the "fundamental," and the others "upper partials."

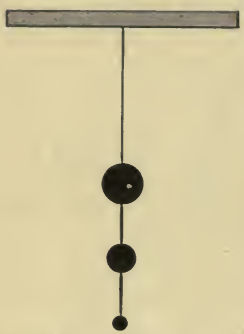


FIG. 136.—A complex pendulum.

The exact mode of vibration of any body may be ascertained by the smoked glass method, as described on page 319. The curve obtained is in general complex; but this can be analyzed by certain mechanical methods into its simple parts. Instead of using a wire attached to the body and the smoked glass, various optical methods may be used, depending upon photography. By the same methods the frequency of a vibration may be determined with great exactness, a comparison being made between it and a known frequency or a known velocity. The student may consult for further details Poynting and Thomson, *Sound*, page 71.

Composition of Harmonic Motions. — 2. *At Right Angles to Each Other.* When the two component simple harmonic motions have the same period, but are at *right angles to each other*, they may be written

$$x = A_1 \cos nt,$$

$$y = A_2 \cos (nt - a).$$

The resulting motion may be represented graphically by choosing two axes, one for x, the other for y, and laying off points for different values of t. This may be done as follows: Let the line of vibration of one motion be $\overline{BA}$, and let the amplitude of the vibration be $\overline{OA}$; then with O as a centre and $\overline{OA}$ as radius describe a circle; beginning at A, divide the

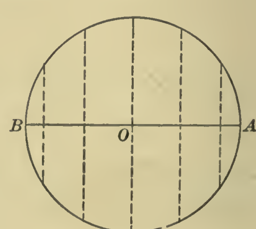


FIG. 137. — Diagram showing how to divide a line BOA into a number of lengths which correspond to equal intervals of time for a point making harmonic motion with the amplitude OA.

semicircumference into a number of equal lengths so that there is a whole number included in each quadrant; from the ends of these equal lengths drop perpendiculars upon the diameter $\overline{BA}$; the points thus marked off on the diameter correspond to equal intervals of time for the point making harmonic motion, as is evident from the definition of this motion as given on page 48.

If the amplitudes of the two vibrations are equal, two lines $\overline{AB}$ and $\overline{CD}$ may be drawn at right angles to each other at

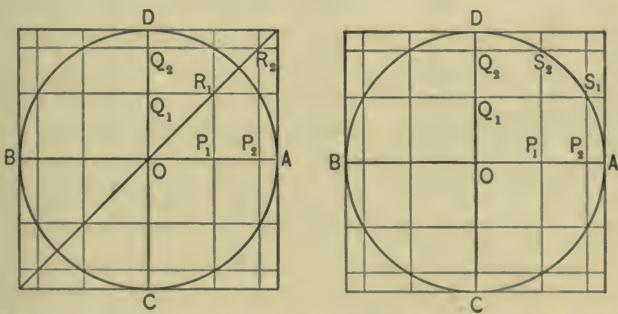


FIG. 188. — Composition of two harmonic vibrations at right angles to each other, whose periods and amplitudes are equal and whose difference in phase is (1) zero, (2) quarter of a period, or $\frac{\pi}{2}$.

their middle points, as shown in the cut, and the points corresponding to equal intervals of time will be as indicated. Draw through these points lines parallel to $\overline{AB}$ and $\overline{CD}$.

If the vibrations are in the same phase, both will be represented by the point O at the same instant; then, when one vibration has reached P_1, the other has reached Q_1; and the geometrical sum is given by R_1. When the former vibration reaches P_2, the latter reaches Q_2; and the geometrical sum is given by R_2; etc. It is evident that the resulting motion is a straight line.

If the vibrations differ in phase by a quarter of a period, *i.e.* by $\frac{\pi}{2}$ in angular measure, one vibration will be at the end of its path, A, when the other is at 0. The geometrical sum is given by A. When the former vibration reaches P_2, the latter is at Q_1; and the geometrical sum is given by S_1; etc. The resulting vibration is evidently a circle.

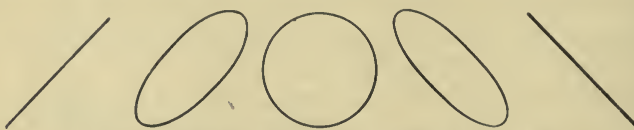


Fig. 139. — Periods equal; difference in phase $O, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \pi$.

If the difference of phase is one eighth of a period, *i.e.* $\frac{\pi}{4}$, the resulting vibration is an ellipse. The curves are shown for a number of different differences in phase.

If the amplitudes of the two vibrations are not the same, the geometrical methods are exactly similar. Lay off two lines, $\overline{AB}$ and $\overline{CD}$, perpendicular to each other at their middle points; divide them into lengths that correspond to equal intervals of time; at these points draw lines parallel to $\overline{AB}$ and $\overline{CD}$. If the phase of vibration is the same, the resulting motion is in a straight line. If the difference in phase is one eighth of a period, the vibration is in an ellipse, as shown, etc.

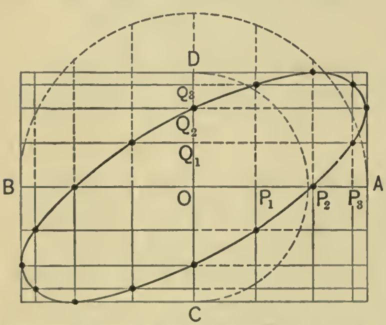


Fig. 140. — Periods equal; amplitudes unequal; difference in phase one eighth of a period.

The simplest way, however, of compounding the vibrations is to eliminate t from the equations, and plot the resulting equation. Different curves will be obtained by giving a different values.

The curves of Fig. 139 may be obtained by several physical processes. One is to use, as described on page 319, a pendulum that can trace a path on a piece of glass; but in this case the glass is kept stationary and the pendulum is set swinging, not in a plane through its origin, but in a cone, as a result of a sidewise push given it when it is held out at the end of its swing. Another method, due to Lissajous, is to use two large tuning forks whose frequencies are the same, and to place them, with their vibration planes perpendicular to each other, in such a position that a pencil of light from a small source incident on the end of a prong of one fork is reflected to the end of a prong of the other and thence to a screen, or into the eye of the observer. This arrangement is shown in the cut. If only one fork is vibrating, a straight line is seen; but if this fork is quiet and the other is vibrating, another straight line at right angles to the first is seen. These lines are caused by the rapid harmonic motions of the two forks. If now both forks are set vibrating, the path of light seen is an ellipse. If the forks are started again after having come to rest, the shape of the ellipse will be different, in general, owing to the fact that their difference in phase is not the same as before. This is on the assumption

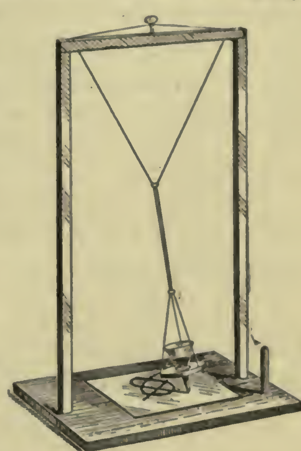


FIG. 141. — Apparatus for compounding vibrations of different periods. See page 327.

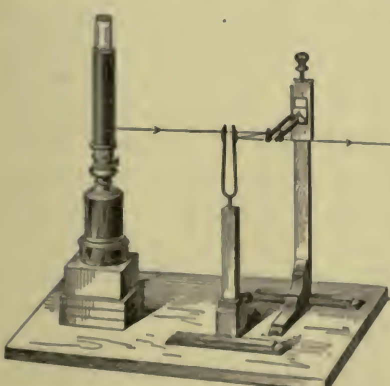


FIG. 142. — Lissajous' arrangement of two tuning forks.

that the frequencies of the two forks are *exactly* equal; if they are not, the shape of the ellipse will change as one looks at it, showing that the difference of phase between the vibrations has changed. The reason for this is seen at once if one considers the two equations for the forks. If their periods are not quite the same, these may be written

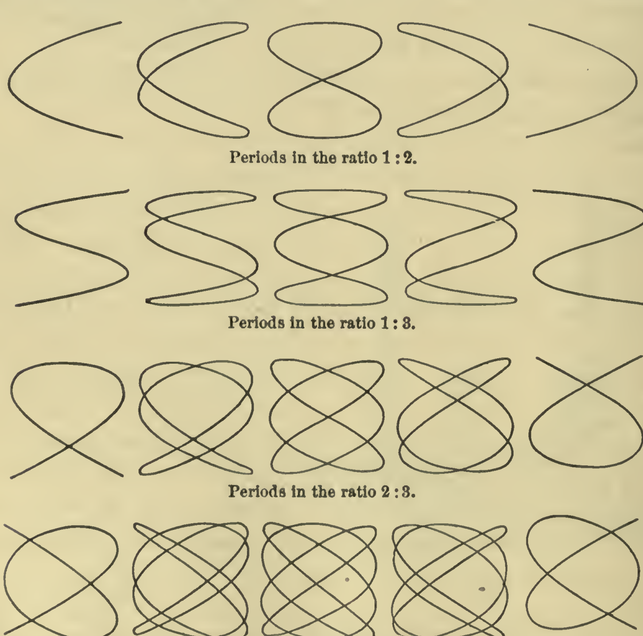
$$x = A_1 \cos nt,$$

$$y = A_2 \cos (n_1 t - a),$$

where $n_1 = n + b$, and b is a small quantity. Therefore substituting for n_1 its value,

$$y = A_2 \cos [nt - (a - bt)].$$

Comparing this with the equation for x, the difference in phase is seen to be $a - bt$; and this is different for different



Periods in the ratio 1 : 2.

Periods in the ratio 1 : 3.

Periods in the ratio 2 : 3.

Periods in the ratio 3 : 4.

FIG. 143. — Lissajous' figures.

values of the time. In words, if one vibration has a period slightly less than that of the other, it will gain on the latter, thus passing through various differences in phase; it will finally gain a whole vibration, when the cycle of changes will repeat itself. If the periods agree exactly, the particular kind of ellipse seen will be a matter of chance, depending upon the manner of starting the forks, but it will persist until the motions of the forks die down owing to friction and to loss of energy caused by the production of waves.

Lissajous' Figures. — Similar statements may be made in regard to compounding two harmonic vibrations of different periods, whose directions are at right angles to each other. The curves obtained when the frequencies bear certain simple relations to each other are given in the accompanying cuts. The different curves in any one cut correspond to different differences in phase; and if, using Lissajous' method, the frequencies are not exactly in the proper ratio, one curve will gradually change into another. The number of cycles of changes in a unit of time depends, of course, upon the divergence of the frequencies of the forks from exact adjustment; and, if this number is counted, the general shape of the curve noted, and the frequency of one fork known, that of the other may be deduced at once by a simple formula.

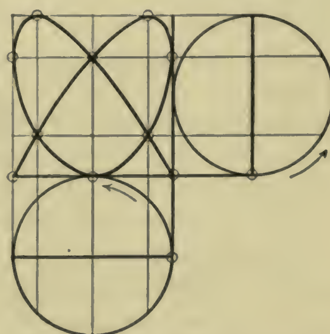


Fig. 144. — Geometrical method of compounding two harmonic vibrations whose periods are in the ratio 8 : 4, whose amplitudes are unequal, and whose phases are different.

Note. — Although these curves are generally called Lissajous' figures, they were first drawn and described by Nathaniel Bowditch of Salem, Mass., in 1815, forty years before Lissajous did the same.

Energy of Vibrations

Damped Vibrations; Dynamics of Vibrations. — If any actual harmonic motion is observed, it is seen that its amplitude slowly decreases; it is said to be "damped." If this decrease is very gradual, there is no change in the period; but if it is rapid, as for instance if a pendulum has a piece of paper fastened to it, the period is sensibly increased. This decrease in amplitude is due to loss of energy by the vibrating body, generally by friction as it moves through the air or at the pivot.

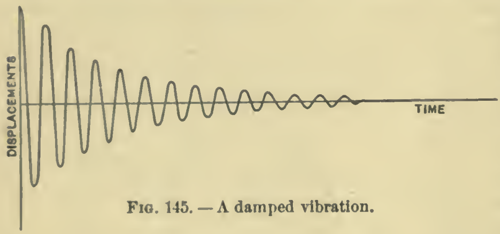


FIG. 145. — A damped vibration.

This is evident if we calculate the energy of the vibrating particle. If its motion is given by $x = A \cos(nt - a)$, its speed at any instant is $s = An \sin(nt - a)$ (see page 51); and so if its mass is m, its kinetic energy at this instant is $\frac{1}{2} mA^2n^2 \sin^2(nt - a)$. This varies at different instants of the vibration, but is always proportional to A^2. (The mean value over one period may be proved by the infinitesimal calculus to be $\frac{1}{2} mA^2n^2$.) During the motion, as fast as kinetic energy is gained, potential energy is lost; and so the mean total energy during a vibration is twice the mean kinetic energy; and this, from what has just been said, varies directly as the square of the amplitude. Therefore as the amplitude of vibration decreases, the particle loses energy.

Forced Vibrations; Resonance. — Even though a body which can make vibrations like a simple pendulum has a definite period of its own, it may be given a different period, if it is attached to some other vibrating system; thus, if the point of support of a pendulum is moved to and fro by a hand making harmonic motion, the motion of the pendulum is the resultant of two, one due to the force applied by the hand, the other its own natural motion. If this last is greatly damped by attaching a sheet of paper to the pendulum, it soon dies down; and the final motion of the pendulum is that due to the harmonic force of the hand. *This motion has the same period as that of the harmonic force;* and is called a "forced vibration," to distinguish it from the natural free vibration. The amplitude of this forced vibration depends to a great extent upon how closely the period of the force agrees with that of the natural vibration; if they are exactly equal, the amplitude is very large. This condition is called "resonance," and is illustrated in many ways. The case of a child in a swing being set in motion by a series of pushes given at intervals agreeing exactly with the natural period of the swing has been mentioned already. In a similar manner a heavy church bell may be set swinging. If a tuning fork is vibrating near another one of the same frequency, the latter will be set in vibration. Many other simple mechanical cases of resonance are given in Rowland's *Physical Papers*, page 28.

If the period of the force varies slightly from the natural period of the vibrating body, the amplitude is not so great as when there is resonance; and in most cases one can tell with considerable accuracy when the resonance is exact.

The fact that a harmonic force produces in a system whose own vibrations are greatly damped a vibration whose period is the same as its own is of great importance. If the force is periodic, but complex, each of the component harmonic forces produces a corresponding harmonic vibration having its period; but the phases and amplitudes of these component vibrations bear relations to each other that are not the same as for the component forces. Therefore, the resultant complex vibration is different from the complex force in "form." It is only a harmonic force that can "reproduce" itself, of course with variations in the amplitude.

CHAPTER XVIII

VELOCITY OF WAVES OF DIFFERENT TYPES

Wave Front. — When waves spread out from a centre of disturbance, a surface can be described that marks at any instant the points which the disturbance has reached. This is called the "wave front." Thus, if a stone is dropped in a pond, or if a raindrop falls on a pool of water, the wave front is marked by an ever-expanding circle. If waves are produced in air by a vibrating tuning fork, the wave front at some distance from the fork is very approximately a sphere. These are called "spherical" waves. The waves in the ether that reach us from a distant star, or from any distant terrestrial source of light which is small, have a spherical wave front; but this sphere has such a large radius that the portion of the wave front that affects us is practically a plane; so we call these "plane waves."

Intensity of Spherical Waves. — If we consider a point source, which therefore produces spherical waves, we can easily calculate the relative intensities (see page 314) at different distances from the source. Let us describe two spherical surfaces of radii r_1 and r_2 around the source; their areas are $4\pi r_1^2$ and $4\pi r_2^2$. So if the source emits in a unit of time an amount of energy equal to E, and if there is no absorption by the medium, the intensity at any point of the first surface is $\frac{E}{4\pi r_1^2}$, and that at any point of the other surface is $\frac{E}{4\pi r_2^2}$. If the former intensity is called I_1 and the latter I_2, it is seen that

$$I_1 : I_2 = \frac{1}{r_1^2} : \frac{1}{r_2^2}.$$

Or, in words, the intensity varies inversely as the square of the distance from the source. (If the source is of such a kind as to produce harmonic motions at all points in the surrounding medium, we see at once that the amplitude of the vibration at a point in the first spherical surface bears a ratio to that at a point in the second surface given by $A_1 : A_2 = \frac{1}{r_1} : \frac{1}{r_2}$; for we have shown on page 328 that the energy of a harmonic vibration is proportional to the square of the amplitude, *i.e.* $I_1 : I_2 = A_1^2 : A_2^2$; and therefore by the formula just deduced for the intensity, the one for the amplitude follows at once.)

Velocity of Waves. — The rate at which the wave front advances is called the "velocity" of the waves. In the next article we shall deduce its value in certain simple cases in terms of the physical properties of the medium; but from general considerations it is evident that in an elastic medium the velocity will be increased if the elastic force of restitution of the medium is increased, and will be decreased if the inertia of the medium is increased; and conversely. In fact, we can prove without difficulty that the velocity of compressional waves in a homogeneous fluid whose density is d and whose coefficient of elasticity is E, is given by the formula

$$V = \sqrt{\frac{E}{d}}.$$

This formula is due to Newton, and is deduced in the *Principia*.

The wave front of waves in the air is affected naturally by winds. Thus, if plane waves are advancing in a direction opposite to the wind, the upper portions of the wave front will be more retarded than the lower, because, owing to friction, the wind near the earth has a less velocity than higher up. So the wave front will lean backward and will proceed up in the air away from the earth. Similarly, if the wind is blowing in the direction of the advance of the waves, the wave front will lean forward.

If a "point source" that is producing spherical waves has motion of translation apart from its vibration, the wave front may change. Thus, if the source is moving in a straight line with a constant velocity, it is evident that, so long as this velocity is less than that of the waves, the wave front remains spherical. If, however, the velocity of the source is greater than that of the waves, things are different.

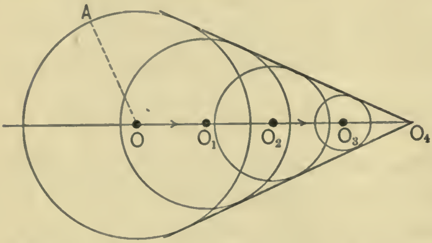


FIG. 146. — Waves produced by a point source which is moving with a velocity greater than that of the waves in the medium. While the source moves from O to O_4, the waves from O reach a distance $\overline{OA}$.

In the cut let O, O_1, O_2, O_3, O_4 be the position of the source at instants zero, t, $2t$, $3t$, $4t$; draw a sphere round O with a radius $4tv$, where v is the velocity of the waves; one round O_1 with a radius $3tv$; etc. These spheres mark the distances the disturbances have spread out at the time $4t$; so, when the source is at O_4, the wave front is a *cone*, whose section is shown by the cut. This fact is illustrated by the waves seen at the bow of a rapidly moving boat.

A convenient model for the study of wave motion in an elastic medium is made by suspending a great number of lead balls by long strings at equal intervals in a straight horizontal line, and joining them by spiral springs. If the end ball is pushed in quickly toward the next one, a compression is propagated down the line; if it is pulled away rapidly from the next

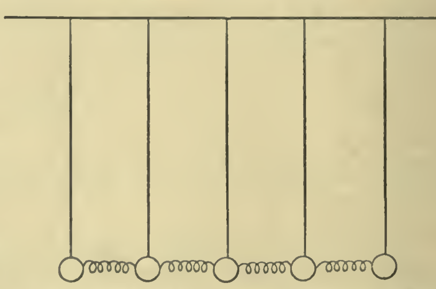


FIG. 147. — A ball and spring model for illustrating wave motion.

ball, an expansion is propagated; if it is given any periodic motion, a corresponding periodic train of waves is produced. The velocity of these disturbances is evidently increased by using stiffer springs or lighter balls, and diminished by using heavier balls or weaker springs.

Reflection of Waves. — When waves in any medium meet an obstacle, there is, as a rule, reflection, as is illustrated by water waves being thrown back by a pier wall, by the phenomenon of echoes, by the use of a mirror in light, etc. We can deduce the exact conditions of reflection by considering two ball-and-spring models like the above, which have different velocities for compressional waves, and which are connected so as to form a continuous "medium."

Let the two sets of balls be arranged in the same straight line, the ones which carry waves with the less velocity on the left, as shown in the cut. We can call this set the "slow" one, the other the "quick" one. When a compression is propagated from left to right, it travels with constant velocity until the second set is reached. At the boundary the compression is carried on more rapidly by the quick set of balls, and therefore the spring joining the two sets is not compressed as much as it would be if the velocity were not increased; the last ball of the slow set is thus pulled slightly toward the right, and so produces an extension of the spring back of it; etc. Therefore a compression in the slow set produces a compression in the quick set, and an expansion is reflected back along the slow set from the

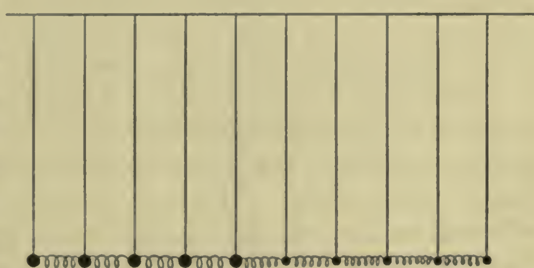


FIG. 148. — A model illustrating two media; in the one on the left the velocity of waves is less than in the one on the right.

boundary. If an expansion is propagated from left to right, there will be an expansion produced in the fast series of balls; but at the boundary the first ball of this set yields more easily to the force of the spring pulling it toward the left than does the last ball of the slow set. Therefore the spring at the boundary is not extended as much as the others, and this will then produce a compression in the springs of the slow set. An expansion, then, propagated along the slow set produces an expansion in the fast one, but a compression is reflected back along the slow one. If a series of waves consisting of alternate compressions and expansions, such as would be produced in this model by giving harmonic motion to the first ball of the slow set, meets a set of "faster" balls and springs, the reflected waves are of the same nature; but whenever a compression reaches the boundary, an expansion is reflected, and conversely. In an exactly similar manner it may be shown that, if such a series of waves in a set of fast balls and springs is incident upon a boundary beyond which there is a slow set, a similar series of waves is reflected; but in this case a compression produces a compression, and an expansion an expansion at the boundary. There is a difference, then, in the reflection at the boundary between the fast and slow sets, depending upon the direction from which the incident waves come; and this difference is equivalent to a substitution of a compression for an expansion, or *vice versa*.

It should be particularly noted that if the velocities of the waves in the two media are the same, there is no reflection. (It is assumed that the waves are not damped, that is, that there is no absorption.) Therefore, in order to have reflection of waves at a boundary separating two media, these must be such that the velocity of the waves is different in the two. (The "rolling" of thunder is an obvious illustration of the reflection of air waves owing to the presence in the air of foreign bodies, namely clouds, or of regions in which the velocity is different.) Another obvious condition for securing reflection from an obstacle is that its area should be large compared with the length of the waves; otherwise they will pass around it.

Velocity of Transverse Waves along a Flexible Stretched Cord. — It is not difficult to calculate the velocities of certain classes of waves. This is true of the propagation of transverse disturbances along a stretched but perfectly flexible cord in which the tension is constant. Imagine a tube, which is straight except for a circular portion near its middle, slipped over this cord and moved rapidly along it with a constant velocity v. Let us consider the motion of the particles of the cord as the curved portion of the tube reaches it, and the forces which the tube exerts on the cord. As this curved

FIG. 149. — A cord, over which has been slipped a bent tube, is stretched between P and Q.

portion of the tube reaches any particle of the cord, it gives the particle a motion which may be resolved into two components: uniform motion in a circle with constant speed v, and uniform motion parallel to that of the tube along the cord with constant speed v. The former motion will be considered presently. As the particle enters the curved portion, it is given, therefore, a momentum along the line of the cord, which it keeps until it leaves the other end of the curved portion, when it is given an equal momentum in the opposite direction and brought to rest. The two forces that produce these changes in momentum are due to the tube; but one balances the other exactly; so there is no resultant action or reaction due to them. The only other acceleration is that occasioned by the particle being made to move in a circle with constant speed. If l is the length of any minute portion of the cord, d its mass per unit length, and r

the radius of the circular portion of the tube, the force which must be exerted on this portion of the cord to make it move in the circle is $dl\frac{v^2}{r}$. There are three forces acting on it, the tension in the cord acting at its two ends and the reaction of the side of the tube against which it presses. Let us calculate the former. In the cut let A, B, C represent three consecutive points in the cord, drawn on an immense scale; let the length of the arc $\overline{ABC}$ be l, the portion of the cord which we are considering; let O be the centre of the circle drawn through these three points; and T be the tension in the cord. Owing to this tension there are two forces acting on this portion of the cord, as shown in the cut. Calling the angle between $\overline{AO}$ and $\overline{OB}$ (and also between $\overline{OB}$ and $\overline{OC}$) N, the component of each of these forces along $\overline{BO}$ is $T \sin N$, and their components perpendicular to $\overline{BO}$ balance each other. The resultant, then, of these two tensions is $2\,T \sin N$, and its direction is along $\overline{BO}$. N is an extremely small angle, and so $\sin N$ may be replaced by N; but $2\,N$ is the angle (AOC), and this equals the length of the arc ABC, *i.e.* l, divided by r, the radius. So the resultant force due to the tension acting on the elementary portion of the cord considered is $\frac{Tl}{r}$. If the tube is moving with such a velocity that this equals the force required to make the portion l move in a circle, *there is no reaction of the tube on the moving cord;* that is, the tension in the cord is itself sufficient to maintain the cord in the given curvature; and the tube now exercises no force on the string whatever. In other words, the tube could now be removed entirely, and the "hump" in the cord would be propagated of itself

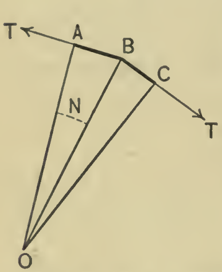


FIG. 150.—A, B, C are three consecutive points of the curved portion of the cord.

along the cord with the velocity of the tube before it was removed. The mathematical condition for this velocity, as given above, is that

$$\frac{Tl}{r} = dl\frac{v^2}{r}, \text{ or } v = \sqrt{\frac{T}{d}}.$$

This shows that the velocity of any hump, whatever its radius, is a constant quantity for a given cord under a definite tension; therefore, any kind of a disturbance, whatever its nature, will be propagated with this same velocity. The fundamental conditions are that there is no force in the string except its tension and that this is constant.

Velocity of Compressional Waves in a Fluid. — The velocity of compressional waves in a fluid may also be calculated; but the process is not quite so simple as in the previous case. The student will find it given in Maxwell, *Heat*, 4th edition, page 223, and in many other text-books. The result is to prove that, if E is the coefficient of elasticity of the fluid, and d its density, the velocity of the waves is given by the formula $V = \sqrt{\frac{E}{d}}$. Nothing need be said in explanation of d; but we have seen that the "elasticity" of a gas is an indefinite quantity, unless the condition is defined under which the change in volume occurs (see page 194). In the present case, the number of compressions and expansions in one second is so great that the change in volume of the fluid takes place adiabatically; for there is not time for heat transfer to occur. Consequently in the above formula E is the adiabatic coefficient of elasticity.

Its value for any *gas* has been given already (page 253). If c is the ratio of the two specific heats, and p the pressure of the gas, E equals the product cp. Therefore, for a gas, $V = \sqrt{\frac{cp}{d}}$; and this can be simplified by using the gas law $p = RdT$. Thus, $V = \sqrt{cRT}$. Therefore, the velocity varies directly as the square root of the absolute temperature; and further, if V, R, and T are known for a gas, c may be calculated.

Laplace was the first to see (1816) that the coefficient of elasticity in this formula was the adiabatic one. Newton, who was the first to derive and apply the formula, used the isothermal coefficient, whose value equals p, and thus made an error.

Assuming the value of c for air to be 1.40, and substituting proper values of p and d, the velocity of air waves may be calculated at any temperature. The value of this velocity at 0° C. is thus equal to 33,170 cm. per second, if the C. G. S. system is used.

The fact that the velocity varies with the temperature is illustrated by the observations of arctic travelers who have noticed that the so-called "velocity of sound" is less at low temperatures. The velocity of waves in air is also seriously affected by the presence of moisture, because the density of the air is changed.

The velocity is seen to be independent of the nature of the disturbance propagated, and also of the pressure. When there is a violent explosion in the air, there are slight variations in the velocity near the centre of disturbance, owing to the fact that the value of the elasticity of the gas, as given above, is true for small variations in the pressure only. At some distance away from the centre, however, the velocity becomes normal.

Similarly, the velocity in other gases may be calculated. The velocity of waves in liquids may also be deduced from the original formula $V=\sqrt{\frac{E}{d}}$. For water at 8° C. it is found, as stated before (page 172), that an increase in pressure of one atmosphere, *i.e.* of 76 cm. of mercury, decreases a unit volume by 0.000047 of its value. If the thermal effects may be neglected,

$$E=\frac{76\times 981\times 13.59}{0.000047}, \text{ and } d=1:$$

therefore, $V=145{,}000$ cm. per second.

There are also experimental methods for the determination of these velocities in gases and liquids; the details of which are described in larger text-books such as Poynting and Thomson, *Sound*. These methods may be divided into two classes: direct and indirect. In the former, a disturbance is produced at some point and the time taken for the waves to reach a point at a measured distance away is accurately measured. The disturbance may be the ringing of a bell, a mild explosion, etc.; and the instant of arrival of the waves may be determined by the mechanical motion of a diaphragm or by the perception of the sound. In the indirect methods, the gas or the liquid is set in vibration by some periodic disturbance whose frequency is known: since this is due to resonance, the frequency of the vibration of the gas or liquid is known; and, as will be shown presently, the velocity of waves in it may be at once calculated.

Velocity of Waves in a Solid. — In solids a purely compressional wave cannot exist, because when there is a compression produced by two opposite forces there is at the same time a distortion. The velocity of longitudinal waves in a solid that extends in all directions is given by the formula

$$V = \sqrt{\frac{k + \frac{4n}{3}}{d}},$$

where k is the coefficient of elasticity for a change in volume, and n the one for a change in shape. A train of waves can, however, be produced where there is only distortion, as when one end of a long wire is rapidly twisted to and fro. In this case, if n is the coefficient of rigidity for the solid, $V = \sqrt{\frac{n}{d}}$. If longitudinal waves are produced in a wire or a rod by stroking it lengthwise with a resined or damp cloth, the velocity of the waves is given by $V = \sqrt{\frac{E}{d}}$, where E is Young's modulus. (See page 154.)

Water Waves. — The velocity of waves upon the *surface* of a liquid depends upon many quantities, and we can do no more here than state certain facts in regard to liquids whose viscosity may be neglected. These statements involve the quantity known as the "wave length," which in the case of waves on liquids may be defined to be the distance from crest to crest or trough to trough. If this quantity has the value l, we have the following formulæ for the velocity:

If the liquid is deep, $V=\sqrt{\frac{gl}{2\pi}}$, where g has its usual meaning. If the liquid is shallow, $V=\sqrt{hg}$, where h is the depth of the liquid. For ripples, $V=\sqrt{\frac{2\pi T}{ld}}$, where T is the surface tension and d the density.

The general formula for waves on liquids which are deep is

$$V^2=\frac{2\pi T}{ld}+\frac{gl}{2\pi};$$

and it is clear that, if the waves are long, the first term is negligible, while, if they are short, the second one is. It is seen by calculation that in the case of water if $l > 10$ cm., the first term may be neglected, and if $l < 0.3$ cm., the second. For intermediate values of l, the full expression must be used.

There are many most interesting applications of these formulæ. The fact that, if one sailing boat has a longer water-line than another, the latter is given a "time allowance" in a race, is due to an attempt to equalize the advantage of the longer boat; for a boat moving through the water produces waves that are comparable in length with its own; and as the boat is helped on by these waves, the longer boat is helped the more because the velocity of the waves it produces is greater. Again, as waves approach a shelving shore, if they are oblique to the shore line, they will gradually turn so as to approach parallel to it, owing to the fact that in shallow water the waves are faster in the deeper portions than in the ones less so.

The motion of the individual particle of a liquid as a wave passes over its surface is in general an elliptical path; and the effect of the waves is felt only a short distance down from the surface, as the amplitude of the vibration decreases rapidly with the depth.

This is not the place to discuss the velocity of temperature waves or of electric waves along wires. But it may be stated that in both these cases part of the energy of the waves is dissipated in heat effects throughout the media and in other ways, and as a consequence of this the waves die down and are not propagated as far as they otherwise would be. The waves are said to become "attenuated." It may be proved also that long waves persist for a greater distance than short ones; and this fact is of fundamental importance in telephone service, as will be shown later.

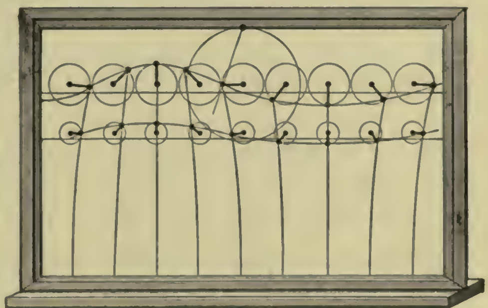


FIG. 151. — A drawing of Lyman's wave model for water waves, showing the form of the wave, the motions of the individual particles, etc.

CHAPTER XIX

HARMONIC AND COMPLEX WAVES—"STATIONARY WAVES"

Trains of Waves and Pulses. — In discussing many of the properties of wave motion it is essential to distinguish two types of waves: one is produced by a sudden irregular disturbance, and may be called a "pulse"; the other is produced by a periodic disturbance, and is called a "train of waves."

Harmonic Waves: Wave Length, Wave Number, Amplitude. — The simplest type of a train of waves is one produced by a centre of disturbance whose motion is harmonic. This is called a "train of harmonic waves," or a "harmonic train." As a consequence of this disturbance, each particle of the medium will be set in harmonic motion, but the *phase* of the vibration varies from point to point at any one instant. (This may be illustrated on the ball-and-spring model.) The simplest mode of representing waves graphically is to choose two axes, one giving the distance the wave front advances in any direction, the other the displacement at any instant at points along a line drawn in this direction.

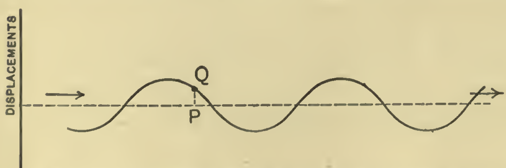


FIG. 152. — Distances along a line in the direction of propagation. A harmonic train of waves.

A curve on this diagram gives the displacements at any instant of all the particles in the medium along a line in the direction of advance of the waves.

If the waves are due to a harmonic disturbance, the curve is as shown, where the line $\overline{PQ}$ indicates the displacement of the particle whose position in the undisturbed medium is P. (If the waves are transverse, this line may be the actual displacement; if they are longitudinal, it equals, or is proportional to, the displacement of P in the direction of the axis. In this latter case a displacement of P toward the right may be indicated by $\overline{PQ}$ being drawn vertically upward, as shown; and a displacement toward the left by the line $\overline{PQ}$ being drawn downward.) As time goes on, the displacements all change, the waves advance; and successive conditions in the medium are illustrated in Fig. 154, where three different curves are drawn for three different instants of time; and it should be noted that in drawing these it is assumed that there is no decrease in the amplitude. If the displacement of any particle P in the medium is observed, it is seen that it is given in succession by $\overline{PQ_1}$, $\overline{PQ_2}$, and

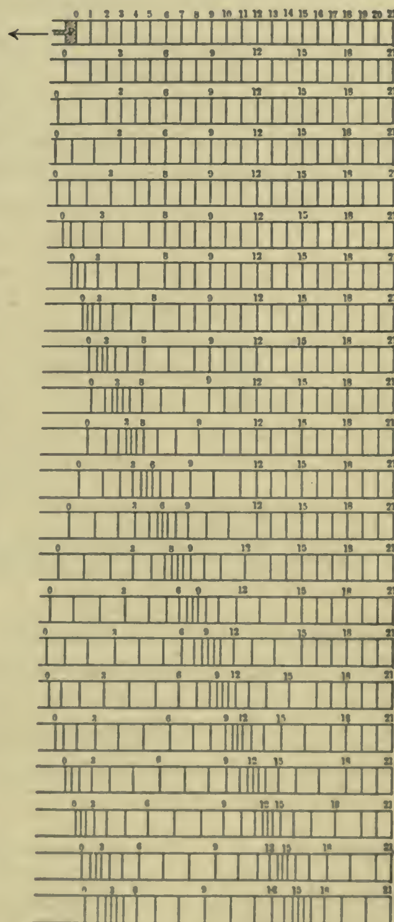


FIG. 153. — A compressional train of waves advancing toward the right. The left-hand particle — marked O — can be considered as kept in vibration by a piston moving to and fro.

$\overline{PQ_3}$; so it is making vibrations as previously described. The amplitude and the period of the vibration of any particle are called the amplitude and the period of the waves, and the frequency of vibration of the particle is called the "wave number." The distance along the line of propagation from any point to the next one where at the same instant the motions are identical is called the "wave length"; *e.g.* from P to R. (This may also be defined as the distance between two consecutive points in the direction

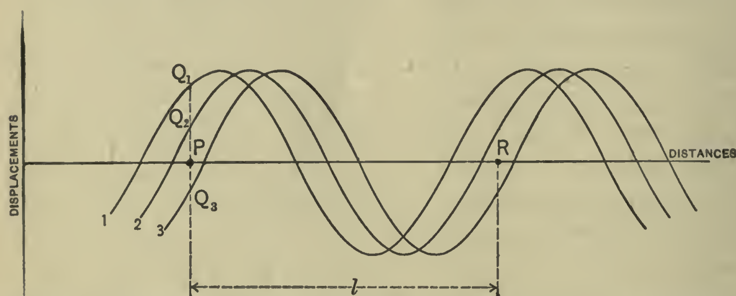


Fig. 154. — Diagram showing three successive positions of the train of waves; the vibration of an individual particle of the medium P; and the wave length $\overline{PR}$.

of propagation at which the *phase* of vibration is the same.) In the time of one complete vibration of a particle the waves advance a distance equal to the wave length, and so in a unit of time the waves advance a distance equal to the product of the wave number and the wave length. Therefore, if V is the velocity of the waves, l the wave length, and N the wave number, $V = Nl$. Or, if T is the period of the waves, $V = \frac{l}{T}$. (The velocity of an individual particle depends upon the instant we consider it, for it is making harmonic vibrations; and so the velocity varies from zero to a maximum value, then decreases to zero, etc. It is clear that there is not the faintest connection between this varying velocity and the constant velocity of the waves.)

Doppler's Principle. — If we speak of that portion of a train of waves which is a wave length long as a "wave," we may say that, if the wave number is N, the source emits N waves in a unit of time; and, in general, N waves pass any point in the medium in a unit of time. This is true if the vibrating source, the point in the medium, and the medium as a whole are not moving. It is interesting, however, to consider the two cases, when the source is moving and when the point in the medium where the waves are counted is moving, the medium not being in motion in either.

If the vibrating source is at rest, and the point in the medium is moving toward it in a straight line, let N be the frequency of the source, V the velocity of the waves in the stationary medium, l their wave length, and v the velocity of the moving point. Owing to the motion of this point it would pass $\frac{v}{l}$ waves in a time if the waves were stationary in the medium; but since they are moving toward the point at such a rate that N waves would pass a fixed point in a unit of time, the total number that passes the moving point in that time is the sum $N + \frac{v}{l}$. $\left(\text{But } N = \frac{V}{l}\text{; and so this number may be written } \frac{V+v}{l}.\right)$ For similar reasons, if the point is moving *away* from the fixed source with a velocity v, the number of waves which pass it in a unit of time is $N - \frac{v}{l}$ $\left(\text{or } \frac{V-v}{l}\right)$.

If the point in the medium is fixed, and the vibrating source is approaching it in a straight line with the velocity v, the case is entirely different. If the source were not moving, the length of a wave would be l; but, when it is moving, the wave front advances a distance V from the source in a unit time, and in this same time the source advances a distance v; so the N waves that have been emitted in this time are crowded together in the interval of space $V-v$, and the length of a wave is now $\frac{V-v}{N}$. The new wave number, or the number of these waves that pass a fixed point in a unit time, is the velocity V divided by this wave length, or $\frac{V}{V-v}N$. Similarly, if the vibrating source is receding from the fixed point with the velocity v, the new wave number is $\frac{V}{V+v}N$.

If v is small compared with V, these last two expressions may be written $\left(1+\frac{v}{V}\right)N$ and $\left(1-\frac{v}{V}\right)N$, or $N+\frac{v}{l}$ and $N-\frac{v}{l}$; so the formulæ for this and the preceding case agree under this condition. It is thus seen that when the source of waves and the point under consideration are approaching each other, the wave number is apparently increased; while, if they are receding from each other, it is apparently decreased.

These formulæ express what is called Doppler's Principle. It is illustrated in the case of a star approaching or receding from the earth, in a whistling locomotive approaching or receding from a station, etc.

Attenuation of Waves. — The energy of a harmonic vibration varies as the square of the amplitude, and hence the intensity of harmonic waves varies as the square of their amplitude. This amplitude decreases as the waves advance, owing to various causes. One case has been considered already on page 331, where it was proved that in spherical waves the amplitude varies inversely as the distance from the source. Further, there may be friction involved in the relative displacements of the particles of the medium, as is the case to a greater or less degree with all waves in all forms of matter; or, motion may be given particles of foreign matter immersed in the medium (see page 311); in both of which cases the amplitude of the waves decreases and they lose energy. Similarly, if waves in one medium are incident upon a boundary separating it from another in which the velocity is different (see page 333), waves are transmitted into the latter medium, and waves are also reflected back into the former. The combined intensity of these two trains of waves must equal that of the incident train; and so the amplitude of the transmitted waves and that of the reflected waves must both be less than that of the incident waves. (An obvious law connects them.)

Superposition of Waves; Complex Waves. — There can be two harmonic trains of waves of the same wave length and amplitude, but differing in phase at any instant, depending upon when or how their motion was begun. Similarly, we may have in any medium waves of different wave length, different amplitude, etc.; and their combined action may be found by compounding them as was done for the vibrations of a particle, for it may be proved that this is allowable, provided the individual displacements of the particles of the medium are small in comparison with the wave length. The best method of considering the superposition is a graphical one. Thus, two plane polarized transverse waves (see page 313) which are harmonic, and whose directions of vibrations are at right angles, may be compounded as shown in Lissajous' figures. In particular, two such waves having the same period would combine to produce an elliptically polarized train of waves, or a circularly polarized train if their amplitudes are equal and their difference in phase $\frac{\pi}{2}$. Conversely, an elliptically or circularly polarized train of waves may be resolved into two plane polarized trains of waves which are harmonic and whose vibrations are at right angles to each other.

Two plane polarized transverse waves which are harmonic and whose directions of vibration are the same, or two longitudinal harmonic trains of waves, may be compounded in the same manner as were two vibrations in the same direction; and any of the illustrations given on page 321 may be applied to the case of waves. Conversely, by Fourier's theorem, any complex train of waves may be resolved into trains of harmonic waves whose wave lengths are in the ratio of 1 : 2 : 3 : etc. There are other modes of resolution, also, which often are more convenient.

We see, then, that a harmonic vibration produces a harmonic train of waves; a complex vibration, a complex train. A special case of this last is a non-periodic or a confused vibration; it will produce a corresponding wave. If the disturbance is intense, but lasts only a short time, it produces in the medium what we have called a "pulse." Its effect when it reaches a portion of the medium containing foreign matter is naturally different from that of a long train of harmonic waves; because owing to these last there are periodic forces brought into action on the foreign particles, and resonance may follow.

Distortion of Complex Waves. — As we have seen in speaking of the attenuation of waves owing to their decrease in amplitude, there are cases in which long waves are less affected than short ones. (See page 341.) In such cases, if a complex vibration is producing a train of complex waves, its harmonic components of long wave length will persist longer than those of short wave length; and so the character of the complex vibration at different points in the medium will vary. This phenomenon of the change in the "form" of the wave is called "distortion." The further one is from the vibrating source, the more nearly does the vibration approach that of being simple harmonic. This fact is illustrated by ocean cables and by long-distance telephone wires. However complex the electrical disturbance at one end of a cable, that at the other is nearly, if not quite, harmonic. In using a telephone over a long distance the quality of the sound is entirely changed, only the graver notes being heard. The great merit of Pupin's new system of constructing cables and telephone lines is that it not alone decreases the attenuation, but also diminishes greatly the distortion by making the attenuation of all the waves, long and short, the same.

Nodes and Loops

"Stationary Waves." — A most important effect of wave motion is illustrated by a simple experiment which may be performed with a long flexible cord; *e.g.* a long spiral spring or a long rubber tube, one of whose ends is fastened to a fixed support and the other is held in the hand. If the cord is stretched fairly tight and the free end moved sidewise with a rapid harmonic motion, waves will be produced in the cord which will be propagated up to the fixed end, and will there suffer reflection and be propagated back to the hand, etc. Consequently, at any instant there are in the cord two trains of waves traveling in opposite directions. If the frequency of the motion of the hand is exactly right, it will be observed that the cord ceases to have the appearance of being traversed by waves, and vibrates transversely in one or two or more portions or "segments." That is, there are certain points in the cord where there is very little, almost no, motion, which are called "nodes"; and the cord in between these vibrates just like a short cord whose two ends are fastened. The points halfway between the nodes, where therefore the motion is greatest, are called "loops." This type of vibration is sometimes, but improperly, called "stationary" or "standing" waves. In reality, the *waves* have disappeared in the production of a *vibration*. Such vibrations as this are extremely common; and may occur with waves of all kinds.

The explanation is evident. Consider a medium through which are passing in opposite directions two harmonic trains of waves which are not suffering attenuation and whose wave lengths and amplitudes are equal. At any instant they may be represented by curves as in the cut. Since the actual motion in the medium is found by superimposing the two wave motions, there will be certain points, one of which is P_2 in the cut, at which the displacement is zero, owing to one wave neutralizing the effect of the other. As the waves advance — in opposite directions — they continue to neutralize

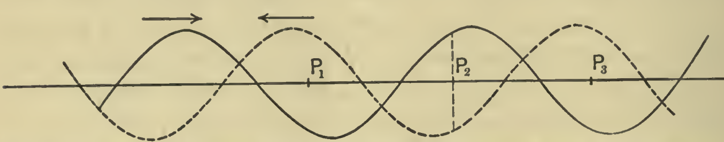


FIG. 155. — Formation of nodes and loops by two trains of waves advancing in opposite directions. P_1, P_2, P_3, etc., are nodes.

each other at this point, as is seen from the cut; therefore this is a node. The importance of the conditions that the waves should not be damped and that the amplitudes of the two trains should be the same as well as their wave lengths is evident. Further, since in a train of waves at a distance of *half a wave length* from any point the displacement is exactly reversed, if two trains of waves neutralize each other at a point P, they will also do so at points distant from P by half a wave length, a whole wave length, etc. *Therefore, the distance apart of two nodes equals one half the wave length of either of the component trains of waves.*

In the case of the cord which was first considered, the fixed end is obviously a node; and the one held in the hand is approximately one, as is evident if one observes the motion. If the long flexible cord is held suspended vertically from a balcony so that the lower end hangs free, a vibration of the same kind can be produced; only in this case, since reflection takes place at a "free end," this point is a loop.

As the frequency of vibration of the hand is increased gradually, it is found that there are certain definite frequencies for which the cord separates into vibrating segments; and the number of these segments increases, that is, the distance apart of the nodes decreases, as these critical frequencies increase. Similarly, if the tension of the cord is increased, the critical frequency must be increased also. The explanation of these facts is not difficult. Let us consider the first case, that of the cord with its end fixed; and let the length of the cord be L, the wave length of the component waves be l, the distance apart of two nodes be d, the velocity of the waves in the cord be V, the frequency of the vibration be N, and the number of segments be n (necessarily a *whole* number). Then, we have the relations: $d = \frac{l}{2}$, $V = Nl$, $L = nd$; and therefore, on substitution, $N = \frac{nV}{2L}$. So if the tension is kept constant, *i.e.* if V is constant, and if the length L of the cord is not varied, n varies directly as N; that is, the frequency of the vibration must have a definite value, since n must be an integer, and if the number of segments is increased from 1 to 2 to 3, etc., the frequency must be increased in an equal ratio. Again, if the tension in the cord is increased, the velocity V is increased; and therefore, if n is the same, N, the frequency, must be increased.

Vibration of a Stretched Cord. — These vibrations in a stretched flexible cord are not always produced by the method described. If the cord is stretched between two fixed points, it may be set vibrating by using a violin bow, by plucking it with a finger, by striking it a blow, etc. The vibrations are exactly like those just described. By lightly touching the middle point of the cord, so as to hold it nearly at rest, and bowing the cord or plucking it at a point distant from the end by quarter of the length of the cord, the string vibrates in two equal segments; if the point touched is at a distance from one end equal to one third the length of the cord, it may be made to vibrate in three segments; etc. The positions of the nodes and loops of any vibration of the cord may be determined experimentally by putting astride the vibrating cord, which is supposed to be horizontal, small paper saddles. They will be thrown off at the loops, but not at the nodes. (This method was used as early as 1677 by two Oxford scholars, William Noble and Thomas Pigott.) If the cord is bowed or plucked at random, the vibration is a complex one, which can be resolved into the simple components just described. (Naturally, the point of the cord which is plucked or bowed must be a loop; and so only those component vibrations are present in the complex one which have a loop at this point.)

We shall consider each of these simple modes of vibration more in detail. The general formula is $N=\frac{nV}{2L}$; and so, if there is one segment, $N=\frac{V}{2L}$; if there are two segments, $N=\frac{2V}{2L}$; if there are three segments, $N=\frac{3V}{2L}$; etc. The frequencies are therefore in the ratio 1:2:3: etc. These vibrations are, as we have said in speaking of Fourier's theorem (page 321), called "harmonics." The first one, in which $n=1$, is also called the "fundamental"; and the others, "upper partials." (It is seen that the frequency when there are two segments is the same as if the cord were of half the length and had but one segment.) A few cuts are given of these transverse vibrations of stretched cords.

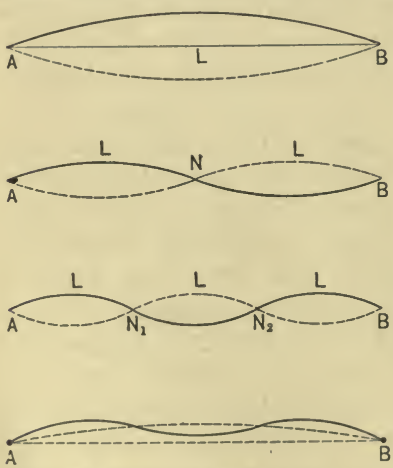


FIG. 156. — Vibrations of stretched cords: (1) one vibrating segment; (2) two vibrating segments; (3) three vibrating segments; (4) superposition of 1 and 3.

The velocity of transverse waves in such a stretched flexible cord has been proved to be given by the formula, $V=\sqrt{\frac{T}{d}}$, where T is the tension in the cord and d is the mass per unit length. Substituting this in the general formula, we have $N=\frac{n}{2L}\sqrt{\frac{T}{d}}$; which shows that if the tension of cord is increased, the frequency is increased; if the length of the cord is increased, the frequency is decreased; etc. All of these facts are illustrated in various musical instruments, as will be noted later.

Vibrations of a Column of Gas. — The same type of vibration can be produced in a column of gas, such as we have in the case of an organ pipe, a flute, a horn, etc. There are many ways in which the vibrations may be produced: by blowing a blast of air over the sharp edge of an orifice opening into the column; by making some solid body in contact with the gas at one end vibrate harmonically with a suitable frequency; etc. (The first is illustrated when one blows an ordinary whistle or when one blows over the end of a hollow key; the latter when one blows a horn by means of the vibrations of the lips or holds a tuning fork over the mouth of a bottle in which water is poured until there is resonance.) If a particle in a column of gas is at a node, that is, if its motion is a minimum, it must be in contact with the end wall or it must be held stationary by symmetrical conditions on its two sides, up and down the column. Thus, as the particles in the gas vibrate, there are the greatest fluctuations in pressure and density at the nodes. Similarly, at a loop there is the greatest motion, but the least change in pressure and density. The vibration of a column of gas is illustrated in the accompanying cut, in which the transverse lines indicate the positions of transverse layers of gas.

If the column of gas is closed by a solid partition, this point is a node; while if the end is open to the air, so that the pressure there cannot change greatly, this point is approximately a loop. (Actually, the loop is a short distance beyond the open end in the air outside. If the tube contain-

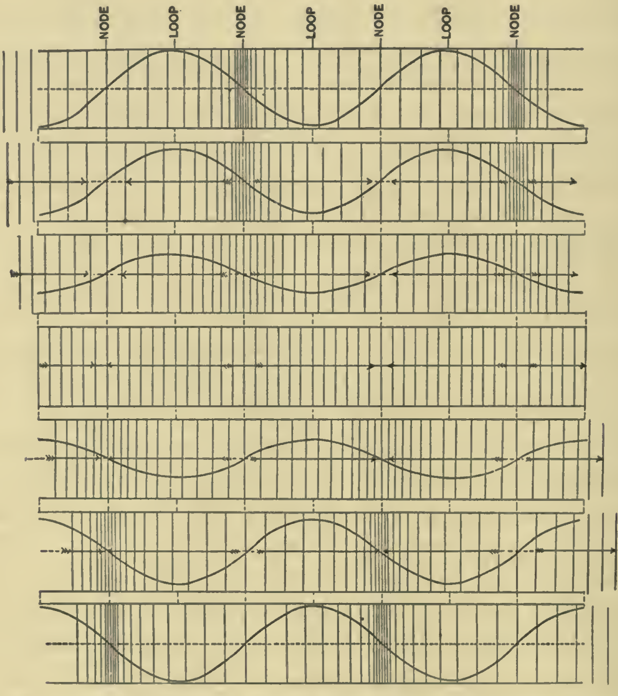


Fig. 157. — Stationary vibration in a column of gas. Vertical lines represent positions of layers of gas. *Curves* represent by their *vertical* displacements the *horizontal* displacements of the layers of gas from their positions of equilibrium. Arrows represent the directions of motion of the layers of gas.

ing the gas is a circular cylinder with a radius R, the loop is at a point beyond the end at a distance given approximately by $0.57R$.)

As in the case of a stretched cord, the column of air can vibrate in different ways, depending upon the number of segments. We shall consider several special cases:

1. *A column of gas closed at both ends.* — There is therefore a node at each end. The simplest mode of vibration is when there is only one loop, which must then be at the middle point; in this case the distance from node to node is the length of the column L. This must equal half the wave length of the component waves, $\frac{l_1}{2}$; and, if N_1 is the frequency of the vibration and V the velocity of the compressional waves in the gas, $l_1 = \frac{V}{N_1}$. Hence, $L = \frac{l_1}{2} = \frac{V}{2N_1}$, or $N_1 = \frac{V}{2L}$. The next simplest mode of vibration is when the column of gas is divided into two segments by a node at its middle point. In this case the distance between two nodes is half the length of the column, $\frac{L}{2}$. Hence,

$$\frac{L}{2} = \frac{l_2}{2} = \frac{V}{2N_2}, \text{ or } N_2 = \frac{2V}{2L} = 2N_1, \text{ etc.}$$

The analogy with the transverse vibrations of a cord stretched between two fixed points is complete. The possible vibrations make a complete harmonic series.

2. *A column of gas open at both ends.* — There is then a loop at each end. The simplest mode of vibration is when there is only one node, which must be at the middle point. We saw in our general discussion that a loop came halfway between two nodes; so the distance from loop to loop equals that from node to node. In this case, the

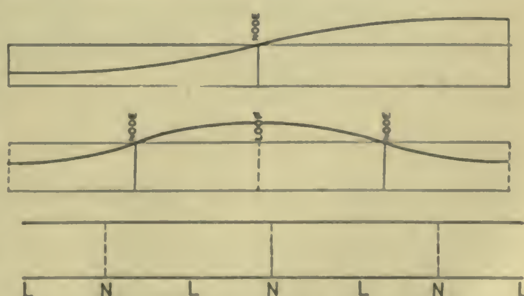


Fig. 158. — Vibrations of a column of gas open at both ends: (1) fundamental; (2) first partial; (3) second partial.

distance from loop to loop is the length of the column, neglecting the slight correction for the ends; and so $L = \frac{l_1}{2} = \frac{V}{2N_1}$, or $N_1 = \frac{V}{2L}$.

The next simplest case is when there is a loop at the middle point and a node at each of the points halfway from it to the ends. The distance from loop to loop is one half the length of the column. Hence,

$$\frac{L}{2} = \frac{l_2}{2} = \frac{V}{2N_2}, \text{ or } N_2 = \frac{2V}{2L} = 2N_1, \text{ etc.}$$

So this case is the same as the previous one; the vibrations form a complete harmonic series.

This kind of a column of gas is called an "open" one; and, as in previous cases, if it is set vibrating by some random or indefinite means, the resulting motion is a complex vibration equivalent to the addition of these simple ones.

3. *A column of gas open at one end and closed at the other.* — There is thus a loop at one end and a node at the other. The simplest mode of vibration is when there are no other nodes or loops. The distance from node to loop is one half the distance from node to node; so in this case $2L = \frac{l_1}{2} = \frac{V}{2N_1}$, or $N_1 = \frac{V}{4L}$.

The next simplest mode of vibration is when there is a node at a distance $\frac{L}{3}$ from the open end, and a loop at a distance $\frac{2L}{3}$. The distance from node to node is then $\frac{2L}{3}$; and we have the relation $\frac{2L}{3} = \frac{l_2}{2} = \frac{V}{2N_2}$, or $N_2 = \frac{3V}{4L}$. Similarly, the next mode of vibration will give a frequency $N_3 = \frac{5V}{4L}$; etc. So in this case the vibrations have frequencies in the ratios 1 : 3 : 5 : etc.; a series of the odd harmonics only. It should be noted that the fundamental in this column of gas has a frequency one half that of the fundamental in the other two cases, when their lengths are the same. This kind of a column is called a "stopped" one. If it is set vibrating at random, its complex vibration is compounded of these simpler ones just described.

Manometric Flame. — A simple method of studying the vibrations of the column of gas in a tube is to pierce openings at intervals along the tube and close them with flexible rubber diaphragms; these are covered on the outside with small hemispherical caps into which there are two openings: one permits the introduction of illuminating gas, the other is joined

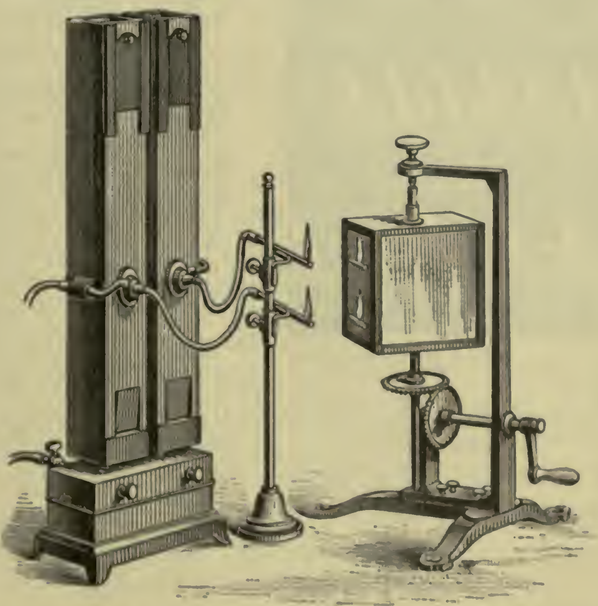


FIG. 159. — Apparatus for the comparison of the vibrating movements of two sonorous tubes.

to a gas burner with a fine opening. If the gas is ignited at this opening, it will flicker if the diaphragm comes at a node in the vibrating column of gas, because the fluctuations of pressure there will cause the diaphragm to vibrate and so affect the pressure on the illuminating gas. These fluctuations have the frequency of the vibration of the column of gas, and are therefore, as a rule, very rapid. So the flickering of the flame must be observed in a revolving mirror, which separates the images when one looks at the reflection of the flame in the mirror. This apparatus is called a *manometric flame.* In its present form it is due to Rudolph König.

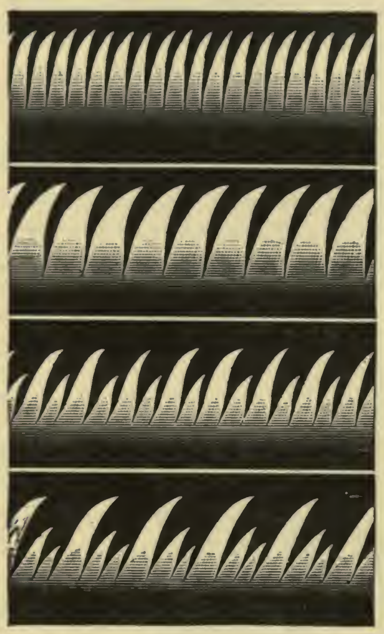


Fig. 159 *a.* — Manometric flames.

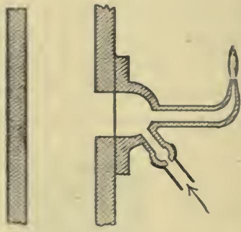


Fig. 160. — Attachment to organ pipe in order to obtain manometric flames.

Longitudinal Vibrations of a Wire or Rod. — Similarly, a wire or a rod can be set in *longitudinal* vibrations by stroking or rubbing it lengthwise. If the wire is stretched between two fixed points at a distance L apart, there are nodes at these points; and if V is the velocity of *compressional* waves in the wire, we have the following relation for the frequency of the fundamental vibration,

$$L = \frac{l}{2} = \frac{V}{2N}, \text{ or } N = \frac{V}{2L}.$$

If a rod is clamped in the middle, this point is a node and the ends are loops. If L is the length of the rod and V the velocity of compressional waves in it, the frequency of the compressional vibrations is given as before by the formula $N = \frac{V}{2L}$.

It may be proved by methods of the infinitesimal calculus, as has been already stated, that the velocity of these compressional waves in a wire or rod is given by the formula $V=\sqrt{\frac{E}{d}}$, where E is Young's modulus of the wire and d is its density. (This modulus E is found, however, by experiment not to have exactly the same value as the one found for the same material by the statical method as described on page 154.)

Other longitudinal vibrations than the fundamentals can be produced in wires and rods; but their frequencies do not have simple ratios with that of the fundamental; so they are not harmonics, but upper partials.

Transverse Vibrations of a Rod. — A rod can also be set in *transverse* vibration. If it is not clamped at any point, but is supported in such a manner as to be perfectly free to vibrate, its fundamental mode is shown in the cut; and its next simplest mode of vibration is as shown. The ratio of the frequency of this vibration to that of the fundamental is 2.92; and in a similar manner the other possible vibrations do not have frequencies which bear simple relations to that of the fundamental or to each other; so these are upper partials, but not harmonics.

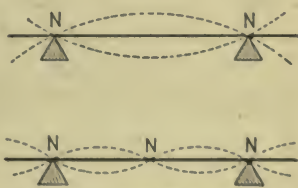


FIG. 161. — Transverse vibrations of a rod: (1) fundamental; (2) first upper partial.

If the rod, instead of being straight, is bent into the shape of a long, narrow U, and has a projection attached at its middle point, as shown in the cut, it makes a "tuning fork." In all actual forks the rod is not uniform, but is thickened at its base; so that the two nodes, which in the uniform rod are near its end when vibrating in its fundamental mode, are brought closer together. The point where the projecting stem is attached is still a loop, as may be shown by setting the fork vibrating and then pressing the stem against a board; the latter will be set in vibration. A tuning fork is, in fact, generally attached to a hollow box, open at one end, and of such a length that the column of air inside has the same frequency as the fundamental of the fork. This is called a "resonance box." Another effect of giving the fork its peculiar shape is so to enfeeble the vibrations other than the fundamental that the latter is the only one present when the fork is bowed or struck so as to be set in vibration.

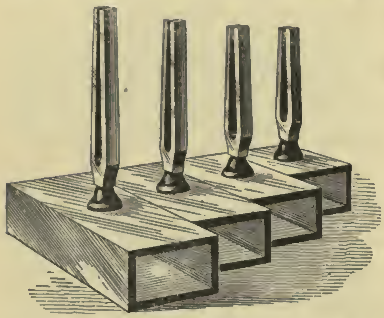


FIG. 162. — Set of tuning forks on resonance boxes.

The frequency of a fork may be measured with great accuracy by comparing its period with that of a standard clock; but for details of the methods of comparison, reference should be made to some larger text-book, such as Poynting and Thomson, *Sound*, Chapter III, or to some laboratory manual for advanced students. (The frequencies of two forks may be compared by Lissajous' figures, or by "beats," if the frequencies are close together. See pages 327 and 412.)

Vibrations of Metal Plates. — Thin metal plates cut in squares or circles can be made to vibrate transversely by clamping them at some point—generally the centre—and drawing a violin bow across their edges. When this is done, it is easy to show that there are certain lines along the plates where there is comparatively no motion; these are called "nodal lines." The simplest method of proving their existence is to scatter fine dry sand over the plates before they are set vibrating; when the vibrations begin, the sand collects along the nodal lines, being thrown there by the other vibrating parts of the plates. There are thus formed most beautiful, regular geometrical figures, which are called "Chladni's figures," after the physicist who first systematically investigated them. Their shape and complications depend upon the point of support of the plate, the point where it is touched with the finger so as to make a node, the point of bowing, and the manner of bowing, which, to a certain degree, determines the number of vibrations, besides the fundamental, which are present.

Similar nodal lines may be observed on stretched membranes, such as drumheads, when they are set in vibration in a proper manner.

If some extremely light powder is used instead of the sand, with which to cover the plate, it is observed that it gathers, not at the nodal lines, but over those portions of the plate which are moving with the greatest amplitude. The reason for this was discovered by Faraday, who showed that small whirls were formed in the air near the vibrating portions of the plate, and that these carry the light powder off the nodal lines on to those portions.

Vibrations of Bells. — Metal bells or bell-shaped objects, like glass tumblers or bowls, may be set in either transverse or longitudinal vibrations. The simplest mode of transverse vibration is shown in the cut. The position of the loops is fixed by the point where the clapper strikes. Many other vibrations than the fundamental are always present, but there is no simple relation between their frequencies. If there is an irregularity in the thickness of the rim at some point, this will produce comparatively little effect if it comes at a node; but if it is at a loop, it will result in a change of the frequency. So a bell like this can vibrate in two ways, giving frequencies that are not very different; and if the bell is struck a random blow, both these vibrations will occur. This is the

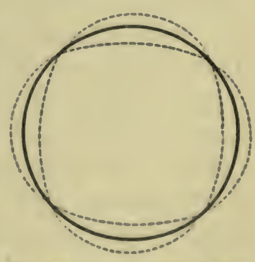


FIG. 163. — One mode of vibration of a bell.

cause of the "beating" of large church bells, as will be explained in a later chapter. (See page 412.)

Other Illustrations of Stationary Waves. — The illustrations so far given of these vibrations due to the superposition of two trains of waves in opposite directions in a medium have been purely mechanical ones, the medium being either a solid or a fluid. But waves in the ether can produce these vibrations also, as is shown by the experiments of Wiener. A node in this case corresponds to a point where there is no motion in the ether (if we retain our mechanical concept of the ether); and a loop, to a point where there is the greatest motion. Rapid vibrations in the ether produce in the surrounding matter various effects, such as chemical changes which may be shown by photographic processes, fluorescence, etc.; and by all of these the existence of nodes and loops in the ether has been proved, when waves in it fall upon a mirror and suffer reflection. The distance apart of the nodes equals half the wave length of the incident waves.

Electrical waves along wires and electromagnetic waves, produced by ordinary electric oscillations, cannot produce these vibrations with nodes and loops, because these waves are all "damped"; and owing to this fact the amplitudes of the reflected and incident waves are not equal.

Measurement of the Velocity of Waves. — These facts in regard to the vibrations of stretched cords, of wires and rods, and of columns of gas (or, in fact, liquids also), lead at once to obvious methods of comparing the velocities of waves in different media. Thus, suppose two stretched cords have the same frequency of vibration when vibrating transversely in their fundamental modes; then if L_1 and L_2 are their lengths, and V_1 and V_2 are the velocities of *transverse* waves along them,

$$\frac{V_1}{2L_2} = \frac{V_2}{2L_2}, \quad \text{or} \quad V_1 : V_2 = L_1 : L_2.$$

Similarly, if the fundamental longitudinal vibrations of two wires or rods of different material are the same, and if L_1 and L_2 are their lengths, we have the same relation, $V_1 : V_2 = L_1 : L_2$, for the velocities of *compressional* waves. Or, if two organ pipes contain different gases, and if their fundamental vibrations have the same frequency, the same formula applies for the velocities of waves in these gases.

There are two general methods for determining when two vibrations have the same frequency if the medium is a material one. As we shall see later, when two vibrating bodies that are not moving bodily are producing sounds, they have the same frequency if the "pitches" of the two sounds are the same; and this can be told with great exactness by a trained ear; or, if they differ slightly, the exact difference in frequency can be determined by the beats. Again, the vibrations may both be produced by resonance from a third vibration. Thus, if a vibrating tuning fork is held at the opening of a column of a gas whose length can be varied, resonance may be secured by noticing for what length the sound is most reënforced; and similarly with the column of a second gas. (A column of gas can vibrate, as has been shown, in many ways, breaking up into a different number of segments. So with a given tuning fork, resonance may be secured for several different lengths of the column of gas. The distance from node to node is the same, however, for all these different nodes of vibration; and it is the quantity L in the previous formula. If the frequency of the fork N is known, and this length L is measured, the velocity of the waves in the gas is obviously $2\,LN$.) Again, stretched cords or wires may have their lengths changed until they are in resonance with a given tuning fork by the following mechanical method: stretch the cord or wire from two pegs or over two knife edges which are fastened to a large box open at its ends; place loosely on the cord or wire a light saddle of paper; set the stem of the vibrating tuning fork on the box; if there is resonance, the wire or cord will be set in vibration and the paper saddle will be thrown off.

Kundt's Method. — We can also compare the velocity of compressional waves in a solid rod with their velocity in a gas or liquid by a method devised by Kundt. The gas is contained in a long tube, which is closed at one end by a tightly fitting piston and at the other by a very light one, which can move to and fro easily, but which nearly fits the tube. This last piston is attached rigidly to one end of the rod which is along the axis of the tube but projects beyond it, and which is clamped at its middle point. This rod is set in longitudinal vibration by stroking its free end with a damp cloth, and so the piston attached to its other end

FIG. 164. — Kundt's apparatus for measuring the velocity of waves.

vibrates and produces waves in the gas contained in the tube. The length of this column of gas is altered by means of the piston at its further end until it is vibrating with nodes and loops. The method of determining this condition is to sprinkle through the tube some light powder, such as is obtained from the finest cork dust; when nodes and loops are formed, the powder collects in ridges across the bottom of the tube, leaving, however, the nodes perfectly bare. The frequency of the vibration of the rod is the same as that of the column of air, because the latter is "forced" by the former; so, if L_1 is the distance from loop to loop in the rod, that is, its length if we neglect the mass of the vibrating piston; and if L_2 is the distance from node to node in the column of air, the velocity of *compressional* waves in the rod is to the velocity of waves in the gas as $L_1 : L_2$. (This method may also be used for a liquid instead of a gas by using suitable powder to mark the nodes.)

It should be observed that the piston on the end of the rod is at a loop for the rod, but a node for the gas; exactly as when the long elastic cord is set in vibration by the hand, as described on page 350, — the motion at a loop in the solid rod is extremely small compared with that at a loop in the gas.

The explanation of the formation of the transverse ridges of the powder depends upon the fact that, as the particles of gas away from the nodes vibrate back and forward between the particles of powder, their mean velocity depends upon the arrangement of the dust particles, and varies at different points, thus producing pressures in certain directions. (See page 169.) The full description of the process is long, and will not be given here. It may be found in Rayleigh's *Theory of Sound*, Vol. II, page 46.

Therefore, if we know by direct experiment, or otherwise, the velocity of waves in air, we may by Kundt's method determine the velocity of compressional waves in any solid material out of which a rod can be made; and then, by replacing the air in the tube by some other gas (or by a liquid), we may determine the velocity of waves in it. (This is the standard method for all gases which can be secured in a small quantity only, such as helium, argon, and the other new gases.)

The values of the velocity of compressional waves in a few substances is given in the following table:

Air	0° C.	33,140 cm. per second
Hydrogen	0° C.	128,600 cm. per second
Illuminating gas	0° C.	49,040 cm. per second
Oxygen	0° C.	31,720 cm. per second
Alcohol (absolute)	8°.4 C.	126,400 cm. per second
Petroleum	7°.4 C.	139,500 cm. per second
Water	8° C.	143,500 cm. per second
Brass		350,000 cm. per second
Copper	20° C.	356,000 cm. per second
Glass		500,000 to 600,000 cm. per second
Iron	20° C.	513,000 cm. per second
Paraffin	16° C.	130,400 cm. per second

CHAPTER XX

HUYGENS'S PRINCIPLE. REFLECTION AND REFRACTION

Huygens's Principle. — One of the most important theorems in regard to wave motion is in part due to Huygens, and it is called by his name. In its most general form it is complicated, and can be demonstrated only by the aid of the infinitesimal calculus. We shall give certain special applications of it; and, although the statements to follow are not rigorous, they are sufficiently so for all present purposes.

We cannot do better than to use Huygens's own language, as it appears in the translation of his *Traité de la Lumière* by Crew in *The Wave Theory of Light* (New York, 1900). Huygens's treatise was written in 1678, but was not published until 1690.

"In considering the propagation of waves, we must remember that each particle of the medium through which the wave spreads does not communicate its motion only to that neighbor which lies in the straight line drawn from the luminous point, but shares it with all the particles which touch it and resist its motion. Each particle is thus to be considered as the centre of a wave. Thus, if *DCF* is a wave whose centre and origin is the luminous point *A*, a particle at *B*, inside the sphere *DCF*, will give rise to its own individual [*secondary*] wave, *KCL*, which will touch the wave *DCF* in the point *C*, at the same instant in which the principal wave, originating at *A*, reaches the position *DCF*. And it is clear that there will be only one point of the wave *KCL* which will touch the wave *DCF*, viz., the point which lies in the straight line from *A* drawn through *B*. In like manner, each of the other particles, *bbbb*, etc., lying within the sphere *DCF*, gives rise to its own wave. The intensity of each of these waves may, however, be infinitesimal compared with that of *DCF*, which is the resultant of all those parts of the other waves which are at a maximum distance from the centre *A*.

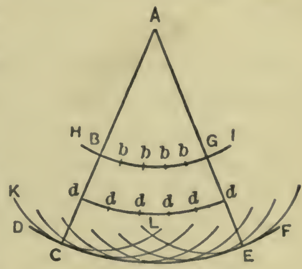


FIG. 165.

"We see, moreover, that the wave *DCF* is determined by the extreme limit to which the motion has traveled from the point *A* within a certain interval of time. For there is no motion beyond this wave, whatever may have been produced inside by those parts of the secondary waves which do not touch the sphere *DCF*."

We shall also quote Huygens in his explanation of reflection and refraction.

1. Reflection of Plane Waves by a Plane Mirror. — "Having explained the effects produced by light waves in a homogeneous medium, we shall next consider what happens when they impinge upon other bodies. First of all we shall see how reflection is explained by these waves, and how the equality of angles follows as a consequence. Let *AB* represent a plane polished surface of some metal, glass, or other substance, which, for the present, we shall consider as perfectly smooth (concerning irregularities which are unavoidable, we shall have something to say at the close of this demonstration); and let the line *AC*, inclined to *AB*, represent a part of a light

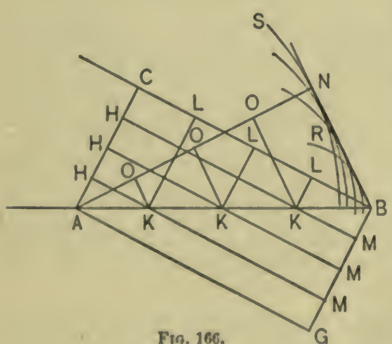


FIG. 166.

wave whose centre is so far away that this part, *AC*, may be considered as a straight line. It may be mentioned here, once for all, that we shall limit our consideration to a single plane, viz., the plane of the figure, which passes through the centre of the spherical wave and cuts the plane *AB* at right angles.

"The region immediately about *C* on the wave *AC* will, after a certain interval of time, reach the point *B* in the plane *AB*, traveling along the straight line *CB*, which we may think of as drawn from the source of light, and hence drawn perpendicular to *AC*. Now, in this same interval of time, the region about *A* on the same wave is unable to transmit its entire motion beyond the plane *AB*; it must, therefore, continue its motion on this side of the plane to a distance equal to *CB*, sending out a secondary spherical wave in the manner described above. This secondary wave is here represented by the circle *SNR*, drawn with its centre at *A* and with its radius *AN* equal to *CB*.

"So, also, if we consider in turn the remaining parts, *H*, of the wave *AC*, it will be seen that they not only reach the surface *AB* along the straight lines *HK* parallel to *CB*, but they will produce, at the centres, *K*, their own spherical waves in the transparent medium. These secondary waves are here represented by circles whose radii are equal to *KM*, that is, equal to the prolongations of *HK* to the straight line *BG*, which is drawn parallel to *AC*. But, as is easily seen, all these circles have a common tangent in the straight line *BN*, viz., the same line which passes through *B* and is tangent to the first circle having *A* as centre and *AN*, equal to *BC*, as radius.

"This line *BN* (lying between *B* and the point *N*, the foot of the perpendicular let fall from *A*) is the envelope of all these circles, and marks the limit of the motion produced by the reflection of the wave *AC*. It is here that the motion is more intense than at any other point, because, as has been explained, *BN* is the new position which the wave *AC* has assumed at the instant when the point *C* has reached *B*. For there is no other line which, like *BN*, is a common tangent to these circles . . .

"It is now evident that the angle of reflection is equal to the angle of incidence. For the right-angled triangles *ABC* and *BNA* have the side *AB* in common, and the side *CB* equal to the side *NA*, whence it follows that the angles opposite these sides are equal, and hence also the angles *CBA* and *NAB*. But *CB*, perpendicular to *CA*, is the direction of the incident ray, while *AN*, perpendicular to the wave *BN*, has the direction of the reflected ray. These rays are, therefore, equally inclined to the plane *AB*.

* * * * * * *

"I remark, then, that the wave *AC*, so long as it is considered merely a line, can produce no light. For a ray of light, however slender, must have a finite thickness in order to be visible. In order, therefore, to represent a wave whose path is along this ray, it is necessary to replace the straight line *AC* by a plane area . . . where the luminous point is supposed to be infinitely distant. From the preceding proof it is easily seen that each element of area on the wave [front], having reached the plane *AB*, will there give rise to its own secondary wave; and when *C* reaches the point *B*, these will all have a common tangent plane, viz., [a plane through *BN*]. This [plane] will be cut . . . at right angles by the same plane which thus cuts the [wave front at *AC* at right angles, *i.e.* the plane of incidence].

"It is thus seen that the spherical secondary waves can have no common tangent plane other than *BN*. In this plane will be located more of the reflected motion than in any other, and it will therefore receive the light transmitted from the wave *CH*.

* * * * * * *

"We must emphasize the fact that in our demonstration there is no need that the reflecting surface be considered a perfectly smooth plane, as has been assumed by all those who have attempted to explain the phenomena of reflection. All that is called for is a degree of smoothness such as would be produced by the particles of the reflecting medium being placed one near another. These particles are much larger than those of the ether, as will be shown later when we come to treat of the transparency and opacity of bodies. Since, now, the surface consists thus of particles assembled together, the ether particles being above and smaller, it is evident that one cannot demonstrate the equality of the angles of incidence and reflection from the time-worn analogy with that which happens when a ball is thrown against a wall. By our method, on the other hand, the fact is explained without difficulty.

"Take particles of mercury, for instance, for they are so small that we must think of the least visible portion of surface as containing millions, arranged like the grains in a heap of sand which one has smoothed out as much as possible; this surface for our purpose is equal to polished glass. And, though such a surface may be always rough compared with ether particles, it is evident that the centres of all the secondary waves of reflection which we have described above lie practically in one plane. Accordingly, a single tangent comes as near touching them all as is necessary for the production of light. And this is all that is required in our demonstration to explain the equality of angles without allowing the rest of the motion, reflected in various directions, to produce any disturbing effect."

The law of reflection in regard to the equality of the angles of incidence and reflection was known to the ancients, and was made use of by Euclid as early as 300 B.C. He also deduced some of the properties of concave mirrors. The law stating that the normals to the two waves and the surface lie in a plane was first given by the Arabian scholar Al Hazen, about 1000 A.D.

2. **Refraction of Plane Waves at a Plane Surface.** — Again quoting from Huygens: "In order to explain these phenomena on our theory, let the straight line *AB*, Fig. 10, represent the plane surface bounding a transparent body extending in a direction between *C* and *N*.

"By the use of the word plane we do not mean to imply a perfectly smooth surface, but merely such a one as was employed in treating of reflection, and for the same reason. Let the line *AC* represent a part of a light wave, whose source is so distant that this part may be treated as a straight line. The region *C*, on the wave *AC*, will, after a certain interval of time, arrive at the plane *AB*, along the straight line *CB*, which we must think of as drawn from the source of light, and which will, therefore, intersect *AC* at right angles. But during this same interval of time, the region about *A* would

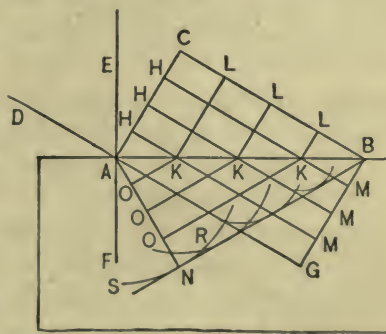


Fig. 167.

have arrived at *G*, along the straight line *AG*, equal and parallel to *CB*; and, indeed, the whole of the wave *AC* would have reached the position *GB*, provided the transparent body were capable of transmitting waves as rapidly as the ether. But suppose the rate of transmission is less rapid, say one third less. Then the motion from the point *A* will extend into the transparent body to a distance which is only two thirds of *CB*, while producing its secondary spherical wave as described above. This wave is represented by the circle *SNR*, whose centre is at *A* and whose radius is equal to $\frac{2}{3}$ *CB*. If we consider in like manner the other points *H* of the wave

AC, it will be seen that, during the same time which *C* employs in going to *B*, these points will not only have reached the surface *AB*, along the straight lines *HK*, parallel to *CB*, but they will have started secondary waves into the transparent body from the points *K* as centres. These secondary waves are represented by circles, whose radii are respectively equal to two thirds of the distances *KM*; that is, two thirds of the prolongations of *HK* to the straight line *BG*. If the two transparent media had the same ability to transmit light, these radii would equal the whole lengths of the various lines *KM*.

"But all these circles have a common tangent in the line *BN*; viz., the same line which we drew from the point *B* tangent to the circle *SNR* first considered. For it is easy to see that all the other circles from *B* up to the point of contact *N* touch, in the same manner, the line *BN*, where *N* is also the foot of the perpendicular let fall from *A* upon *BN*.

"We may, therefore, say that *BN* is made up of small arcs of these circles, and that it marks the limits which the motion from the wave *AC* has reached in the transparent medium, and the region where this motion is much greater than anywhere else. And, furthermore, that this line, as already indicated, is the position assumed by the wave *AC* at the instant when the region *C* has reached the point *B*. For there is no other line below the plane *AB*, which, like *BN*, is a common tangent to all these secondary waves. . . .

"If, now, using the same figure, we draw *EAF* normal to the plane *AB* at the point *A*, and draw *DA* at right angles to the wave *AC*, the incident ray of light will then be represented by *DA*; and *AN*, which is drawn perpendicular to *BN*, will be the refracted ray; for these rays are merely the straight lines along which the parts of the waves travel.

"From the foregoing, it is easy to deduce the principal law of refraction; viz., that the sine of the angle *DAE* always bears a constant ratio to the sine of the angle *NAF*, whatever may be the direction of the incident ray, and that the ratio is the same as that which the speed of the waves in the medium on the side AE bears to their speed on the side AF.

"For, if we consider AB as the radius of a circle, the sine of the angle BAC is BC, and the sine of the angle ABN is AN. But the angles BAC and DAE are equal, for each is the complement of CAE. And the angle ABN is equal to NAF, since each is the complement of BAN. Hence the sine of the angle DAE is to the sine NAF as BC is to AN. But the ratio of BC to AN is the same as that of the speeds of light in the media on the side toward AE and the side toward AF, respectively; hence, also, the sine of the angle DAE bears to the sine of the angle NAF the same ratio as these two speeds of light."

Since these speeds are properties of the media and not of the direction of the propagation of the waves, we have at once the law that the ratio of these sines is independent of the angle of incidence. It is evidently different for different media, and will be shown to be different for waves of different wave length in the case of ether waves in a material medium. This law of refraction was first discovered experimentally by Snell (1591–1626).

Refraction is a much more common phenomenon with ether waves than with air waves, and it will be discussed more fully in the section devoted to Light.

CHAPTER XXI

INTERFERENCE AND DIFFRACTION

Interference

Young's Experiments. — A most important phenomenon of wave motion, and one of particular interest historically because by means of its discovery Thomas Young in 1801 proved that light was due to waves, is what is called *interference.* In Young's own words: "*When two undulations, from different origins, coincide either perfectly or very nearly in direction, their joint effect is a combination of the motions belonging to each.*

"Since every particle of the medium is affected by each undulation, wherever the directions coincide, the undulations can proceed no otherwise than by uniting their motions, so that the joint motion may be the sum or difference of the separate motions, accordingly as similar or dissimilar parts of the undulations are coincident." (This principle is illustrated in the cut, which represents the "interference" of two trains of water waves.)

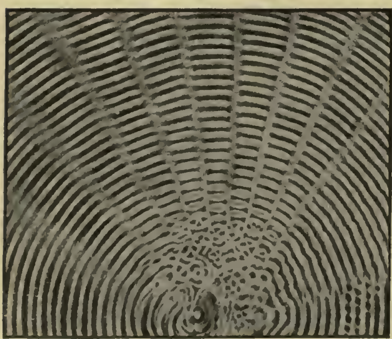


Fig. 168. — Interference of waves on the surface of a liquid, which are sent out by two point sources.

Two experiments may be described; both are due to Young, and both may be performed easily with home-made apparatus. We shall describe them as if the waves to be studied were light waves; but the same apparatus, suitably enlarged, would do equally well for air waves. Let there be trains of waves sent out by having some source placed near a long *narrow* slit in an opaque screen. If the slit is sufficiently narrow, the disturbances will proceed out from the slit in all directions, making a train of waves with a cylindrical wave front. A second opaque screen with two *narrow* slits, which are close together and both of which are parallel to the slit in the first screen, and at equal distances from it, is placed parallel to the latter. As the cylindrical waves reach these two slits, two cylindrical trains of waves are produced beyond the second screen. The importance of this arrangement lies in the fact that the two trains of waves thus produced are *identical;* that is, they have the same amplitude, the same wave length, and the same phase, because they are produced by disturbances in the same wave front at the same distance from the first slit; so, if the original source of the waves changes its character in any way, the two cylindrical waves from the two slits both change in the same manner at the same instant. Then, if we consider the effect at any point in the space which is traversed by the two trains of waves, it is receiving disturbances from both waves, and the effect produced is the sum of two, one due to each train.

Another mode of producing this result is to remove the screen with the two slits, and to place parallel to the slit in the first screen a narrow opaque object like a fine wire or small needle. As will be shown in speaking of diffraction (see page 389), disturbances are produced in the "shadow" exactly as if there were a source of waves along the edge of the obstacle; so, in the case of the wire or needle, the points in the shadow are receiving disturbances from two parallel line sources of waves along the two edges. These two disturbances are, then, in this case also due to two *identical* trains of waves.

If these waves are light waves of a definite color, and if, from a point in the medium traversed by the two trains of waves, one looks in the direction of the two slits (or of the wire), or, better still, if a magnifying glass is used in front of the eye, a series of black and colored lines parallel to the slits (or wire) are seen.

Similarly, if short sound waves are used, it is not difficult to prove by means of a sensitive flame (see page 192) that there are corresponding "bands of silence and sound," meaning that at points along a line parallel to the slits there are disturbances in the air, while along a neighboring line there are not.

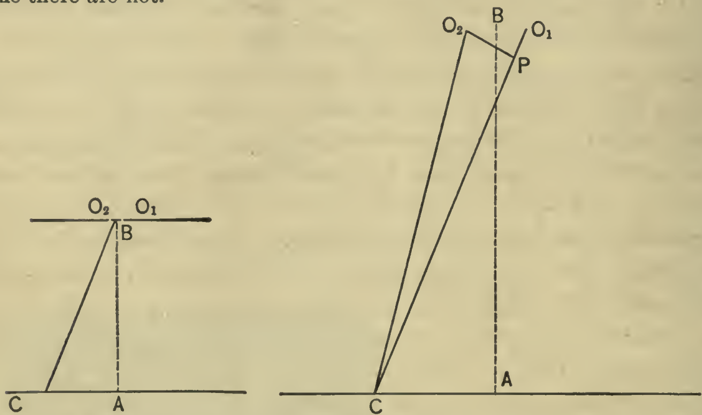


FIG. 169. — Diagram of Young's interference experiment. O_1 and O_2 are two sources of waves and AC is a screen on which the two trains of waves are received.

The explanation is not difficult. Let us consider the effect at various points on a screen parallel to the plane of the two slits. Let O_1 and O_2 be the traces of the slits on the paper, and $\overline{AC}$ that of the last screen. Let B be a point halfway between the slits; draw a line $\overline{BA}$ perpendicular to the screens, and let C be any other point on the screen. This point receives disturbances due to two trains of waves; but the lengths of the paths from C to the two sources O_1 and O_2 are not the same. This is shown on a large scale in part of the diagram.

From O_2 draw $\overline{O_2P}$ perpendicular to the line $\overline{O_1C}$. Since O_1 and O_2 are in reality extremely close together compared with the other distances in the apparatus, $\overline{O_1P}$ is the difference in path from O_1 and O_2 to C. If it amounts to a wave length exactly, or to any integral number of wave lengths, the disturbances reach C in the same phase, and so the effect is great; but, if this difference in path is exactly half a wave length, or any *odd* number of half wave lengths, the disturbances arrive at C in exactly opposite phases, and so there is no effect. At the point A the two paths are of equal length, so the effect is great; and as points near it are considered, constantly receding from A in either direction, the effect decreases, becomes zero when the difference in path is half a wave length, increases to a maximum when this difference is a wave length, decreases to zero again; etc. The effect at all points on a line through C parallel to the two sources is evidently the same; and so the screen is covered with a pattern of bands, or, as they are called, "interference fringes," alternately dark and bright.

The condition for a fringe where there is a maximum effect is, in terms of the figure, that

$$\overline{O_1P} = nl,$$

where n is any whole number 0, 1, 2, 3, etc.; and l is, as usual, the wave length. Similarly, the condition for a fringe of zero intensity is that $\overline{O_1P} = (2n+1)\frac{l}{2}$.

These conditions may be expressed in terms of the distance of the fringe from the central one at A; *i.e.* in terms of the distance $\overline{AC}$. Call this distance x; the distance apart of the two screens, *i.e.* $\overline{AB}$, a; the distance apart of the two sources $\overline{O_1O_2}$, b; and the angle (ABC), N. It is to be remembered that this angle is always small, and that the two sources are so close together that the lines drawn from C to the points O_1, B, and O_2 make practically the same angle with $\overline{AB}$. Then, referring to the diagram, the angles (ABC) and (O_1O_2P) are equal, $\overline{AC} = \overline{AB} \tan(ABC)$, and $\overline{O_1P} = \overline{O_1O_2} \sin(O_1O_2P)$; or, in symbols,

$$x = a \tan N,$$

or

$$x = \sqrt{a^2 + x^2} \sin N,$$

$$\overline{O_1P} = b \sin N,$$

and hence $\overline{O_1P} = \dfrac{bx}{\sqrt{a^2+x^2}}$. Therefore, if x is small compared with a, $\overline{O_1P} = \dfrac{bx}{a}$. If x gives the position of a fringe of maximum intensity, $nl = \dfrac{bx}{a}$, or $x = \dfrac{nal}{b}$. The position of the next similar fringe is given by $x_1 = (n+1)\dfrac{al}{b}$; so the distance apart of the two, or $x_1 - x$, is $\dfrac{al}{b}$. This shows that all the bright fringes are equidistant.

Similarly, if x gives the distance of a fringe where there is a zero effect,

$$(2n+1)\frac{l}{2} = \frac{bx}{a}, \text{ or } x = (2n+1)\frac{al}{2b}.$$

The distance of the next similar fringe is given by

$$x_1 = [2(n+1)+1]\frac{al}{2b};$$

so their distance apart, or $x_1 - x$, is $\dfrac{al}{b}$, the same as for the fringes of maximum intensity.

It is thus seen that the greater the wave length, the farther apart are the fringes; and that, if their distance apart, that of the screens, and that of the sources are known, the wave length may be determined. This matter will be referred to later.

It should be noted that in the formation of these interference fringes there is no destruction of energy; it is simply distributed differently from what it would be if the screen were receiving waves from two sources which had no permanent phase relation.

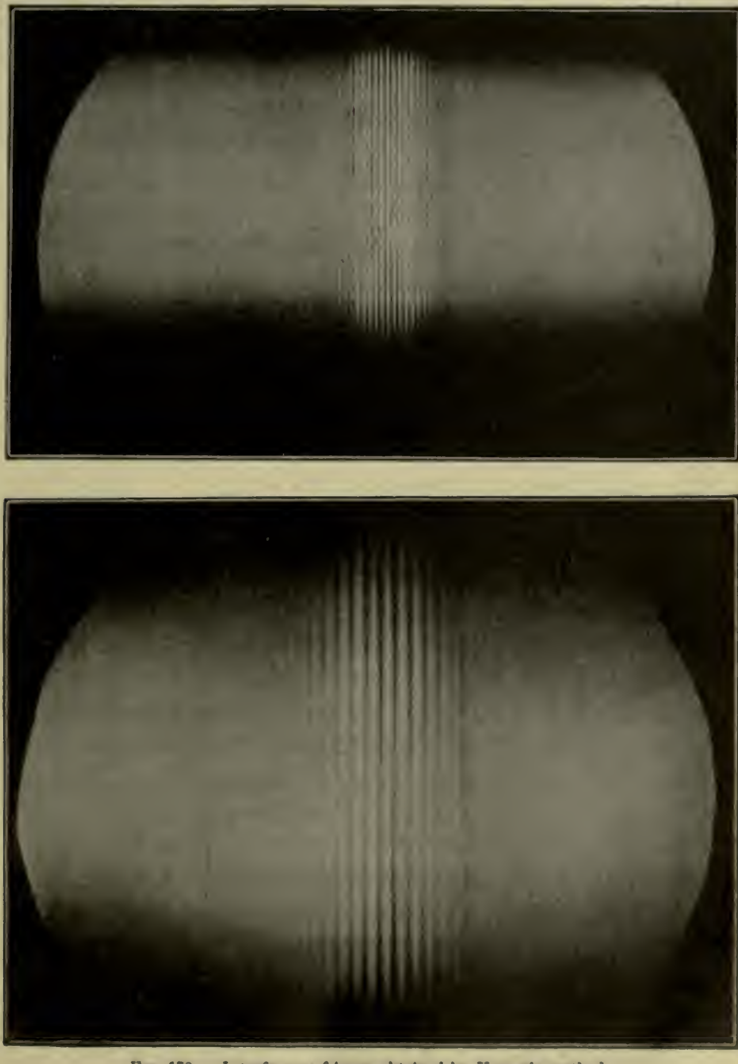


FIG. 170. — Interference fringes obtained by Young's method.

Other cases of interference will be described in the section devoted to Light, but all interference phenomena dealing with light can be reproduced with waves in the air.

Diffraction

Fresnel's Principle. — It is a well-known fact that, if an opaque obstacle is interposed in a beam of light, a shadow will be cast on any suitably placed screen, which is more or less sharply defined, depending upon the smallness of the source of the light. This is sometimes expressed by saying that "light travels in straight lines." In the case of sound, however, such an obstacle would not prevent a noise being heard behind it; in other words, there is no sound shadow with such an obstacle. The explanation of the difference in the two cases was given by Fresnel, making use of Huygens's principle. It may be well to quote Fresnel's own words in Crew's translation. Fresnel's great memoir on *Diffraction*, from which these quotations are made, appeared in 1810.

"I shall now show how by the aid of these interference formulæ and by the principle of Huygens alone it is possible to explain, and even to compute, all the phenomena of diffraction. This principle, which I consider as a rigorous deduction from the basal hypothesis, may be expressed thus: *The vibrations at each point in the wave front may be considered as the sum of the elementary motions which at any one instant are sent to that point from all parts of this wave in any one of its previous* * *positions, each of these parts acting independently the one of the other.* It follows from the principle of the superposition of small motions that the vibrations produced at any point in an elastic fluid by several disturbances

* I am here discussing only an infinite train of waves, or the most general vibration of a fluid. It is only in this sense that one can speak of two light waves annulling one another when they are half a wave length apart. The formulæ of interference just given do not apply to the case of a single wave, not to mention the fact that such waves do not occur in nature.

are equal to the resultant of all the disturbances reaching this point at the same instant from different centres of vibration, whatever be their number, their respective positions, their nature, or the epoch of the different disturbances. This general principle must apply to all particular cases. I shall suppose that all of these disturbances, infinite in number, are of the same kind, that they take place simultaneously, that they are contiguous, and occur in the single plane or on a single spherical surface. . . . I have thus reconstructed a primary wave out of partial [*secondary*] disturbances. We may, therefore, say that the vibrations at each point in the wave front can be looked upon as the resultant of all the secondary displacements which reach it at the same instant from all parts of this same wave in some previous position, each of these parts acting independently one of the other.

"If the intensity of the primary wave is uniform, it follows from theoretical as well as from all other considerations that this uniformity will be maintained throughout its path, provided only that no part of the wave is intercepted or retarded with respect to its neighboring parts, because the resultant of the secondary displacements mentioned above will be the same at every point. But if a portion of the wave be stopped by the interposition of an opaque body, then the intensity of each point varies with its distance from the edge of the shadow, and these variations will be especially marked near the edge of the geometrical shadow."

Rectilinear Propagation. — As a simple case, consider a train of plane waves advancing from left to right; let the paper be at right angles to them and let the trace on the paper of the wave front at any instant be given in part by AB. The effect at any later time at a point P in advance of the waves is determined, as already stated, by deducing the effects there owing to the secondary waves from each point of the wave front, and adding these geometrically.

It is evident that the effect at P of the secondary waves from any point Q on the wave front depends upon the length of the line $\overline{QP}$ for two reasons: the decrease in amplitude of spherical waves varies inversely as the distance, and the phase of the disturbance as it reaches P varies with it.

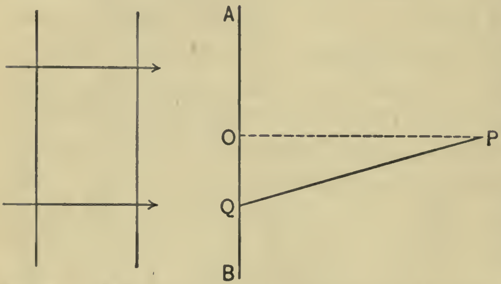


FIG. 171. — $\overline{AB}$ is a section of a plane wave advancing toward P.

Therefore, if O is the foot of the perpendicular let fall upon the wave front from P, *i.e.* its "pole," as it is called, and if a circle with a radius equal to $\overline{OQ}$ be drawn on the wave front around O, the secondary waves from each of the points on this circle will reach P with the same amplitude and in the same phase, because they start out with the same amplitude and in the same phase, and travel the same distance to reach P. But the *directions* of the displacements due to the separate secondary waves are not the same, and they must be added geometrically. (Since the phases are all the same, we have simply a case of vector addition.) Let us assume that the waves are longitudinal (the proof is similar, if they are transverse). Let Q_1 and Q_2 be two points at the end of a diameter of the circle round O; and let the displacement at P due to the secondary waves from Q_1 be represented by

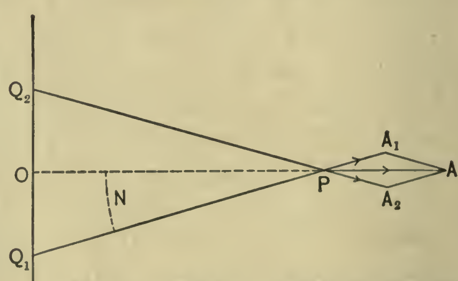


FIG. 172. — Diagram to represent the resultant action at P of all the secondary waves from points Q_1, Q_2, etc., on a circle around O.

$\overline{PA}_1$; then that due to the secondary waves from Q_2 is represented by $\overline{PA}_2$. Their resultant is $\overline{PA}$, a displacement perpendicular to the plane wave front. Similarly, the resultant displacement due to all the secondary waves coming from points in the circle around O, through Q_1, is proportional to $\overline{PA}$. But calling the angle (OPQ_1) the "inclination" of Q_1, and representing it by N, it is evident that $\overline{PA} = 2\,\overline{PA}_1 \cos N$; and therefore, as the inclination increases, the resultant displacement decreases.

The task of compounding the effects at P, due to the secondary waves from all the points in the wave front, is not at first sight a simple one; but Fresnel invented a most brilliant method. His plan is as follows: Calling the length of the line $\overline{OP}$ a, and the wave length of the waves l, describe around P as a centre a series of spherical surfaces of radii $a, a + \frac{l}{2}, a + l, a + \frac{3l}{2}$, etc.; these spheres will cut the wave front in a series of concentric circles around O, thus dividing the plane into a series of concentric zones or circular rings; and the effect at P due to the secondary waves from each zone is considered as a whole.

Fresnel's purpose in thus dividing the wave front into these particular zones was to secure the following condition: If we consider the disturbance at P from any point in any zone whose distance from P is b, this will be exactly opposite in phase to disturbances reaching P at the same instant from certain points in the two contiguous zones, the distances from which to P are $b + \frac{l}{2}$ and $b - \frac{l}{2}$; for, if two waves differ by half a wave length, one produces an effect exactly opposite in phase to the other; therefore, if the whole effect at P due to all the secondary waves from the points in one zone is in a direction which is called "positive," that due to the waves from the two contiguous zones will be in a negative direction. Consequently, calling m_1, m_2, m_3, etc., the magnitudes of the effects at P due to the first, second, third, etc., zones, the total effect at P due to the whole wave front may be written

$$I = m_1 - m_2 + m_3 - m_4 + \text{etc.}$$

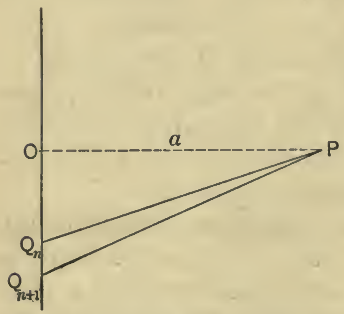


FIG. 173. — Q_n and Q_{n+1} are points on the opposite edges of the $(n+1)$st zone around O.

The numerical value of any m depends upon three quantities: the area of the zone, its mean distance from P, and its "inclination" — as just defined.

The area of a zone is easily calculated. Let the two radii of the spheres around P which determine it be defined by the equations:

$$\overline{PQ}_n = a + \frac{nl}{2},$$

$$\overline{PQ}_{n+1} = a + (n+1)\frac{l}{2}.$$

Then

$$\overline{OQ}_n^{\,2} = \overline{PQ}_n^{\,2} - \overline{OP}^2 = nla + \frac{n^2l^2}{4},$$

$$\overline{OQ}^2_{n+1} = \overline{PQ}^2_{n+1} - \overline{OP}^2 = (n+1)la + \left(\frac{n+1}{2}\right)^2 l^2.$$

Therefore the area of the circle through Q_n is

$$\pi\left(nla + \frac{n^2l^2}{4}\right),$$

and that of the circle through Q_{n+1} is

$$\pi\left[(n+1)la + \left(\frac{n+1}{2}\right)^2 l^2\right];$$

so the area of the zone between is the difference between these, or

$$\pi\left[la + l^2\frac{2n+1}{4}\right].$$

This may be expressed

$$\pi la\left(1 + \frac{l}{a}\,\frac{2n+1}{4}\right);$$

so if l is extremely small in comparison with a, this area is πla, and is therefore the same for all the zones. (In the case of ether waves which affect our sense of sight l is about $\frac{1}{50000}$ in., and so this assumption is justified; for waves in the air which affect our sense of hearing the wave length varies from about half an inch to about 20 ft., and this approximation cannot be used.)

If the wave length is so small that $\frac{l}{a}$ is a small quantity, the mean distances of two contiguous zones from P are also very approximately equal. In the general case, the area of each zone is slightly greater than that of the one inside it, and this fact would make its effect greater than that of the inner zone; but this is counterbalanced by the fact that the mean distance of the inner zone from P is less. In any case, the inclination of each zone is less than that of the contiguous one outside it, but by a very small amount; and therefore the value of m for any zone is greater than that for the outer zone. In symbols, this may be expressed $m_{n-1} > m_n > m_{n+1}$, where m_n is the effect of the nth zone. Therefore, the series $I = m_1 - m_2 + m_3 - m_4 + m_5 -$ etc., is one of terms which decrease in numerical value by extremely small amounts. If we rewrite this series in the form

$$I = \tfrac{1}{2} m_1 + (\tfrac{1}{2} m_1 - m_2 + \tfrac{1}{2} m_3) + (\tfrac{1}{2} m_3 - m_4 + \tfrac{1}{2} m_5) + \text{etc.},$$

we see that each of the terms with the exception of the first is extremely small. The last term in the series is one half the m for the last zone; and, if the wave front is not limited, its value may be assumed to be vanishingly small. Therefore, the sum of all the series except the first term is infinitesimal; and we may write $I = \frac{1}{2} m_1$. This means that the effect at P due to the secondary waves from any zone except the first is neutralized by one half the effects of the two contiguous zones; and so the total effect at P due to the whole wave front is one half that due to the central zone. The area of this zone is πla very approximately, if $\frac{l}{a}$ is a small quantity. (For visible ether waves l is not far from 0.00005 cm., and so if $a = 10$ cm., the area πla equals 0.0016 sq. cm.)

Spherical Waves. — The case of spherical or cylindrical waves may be treated in the same manner; and in some cases other modes of describing "Fresnel zones" are preferable. The result in them all is, however, the same. If from a point in advance of the wave front a perpendicular line is let fall upon it, the effect at this point, due to the whole wave front, is one half of that due to the first central zone around the point where this perpendicular meets the wave front, *i.e.* the "pole." (It is assumed, of course, that the velocity of a disturbance in the medium is independent of the direction of propagation; otherwise, referring to the previous cuts, the effects at P from points in a circle around O would not reach it in the same time, and so they would be in different phases. This case of non-isotropic media will be discussed later in speaking of Double Refraction.)

Let us, then, consider the propagation of a train of waves having a spherical wave front, spreading out from the point S. Let AB be a portion of the wave front at any instant; the effect at P at a later time will be one half that produced there by the central zone around O, where O is the pole of P. If the wave length is extremely small so that the area of this central zone is small, we may say that the effect at P is due to the disturbance at the *point* O; and so in turn, when the wave front reaches P, we may say that the effect at a point

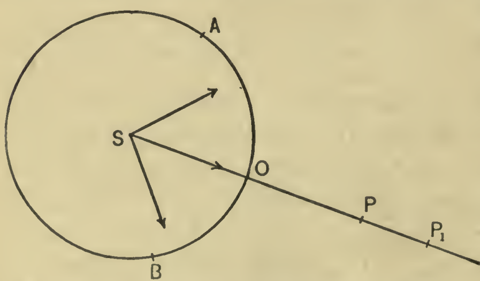


FIG. 174. — Case of spherical waves advancing toward P.

P_1, farther out on the line $\overline{SOP_1}$, is due to the effect at the point P, etc. In this sense, the disturbance due to a train of waves having a small wave length is propagated in a straight line. This line, $\overline{SOPP_1}$, is called a "ray." (In an exactly similar sense, the disturbance of an exceedingly thin "pulse" is propagated in a straight line.) This obviously explains the general phenomenon of the casting of shadows by opaque obstacles when a small source of light is used.

Diffraction past an Edge. — If the waves have a long wave length, the area of the central zone is not small, and the effect is not propagated from *point* to point. So with waves in the air, which produce sounds, there are no sharp shadows; neither are there in the case of light, if one examines the edge of the shadow with care. To see exactly what occurs, we shall consider one simple case, following Fresnel's treatment.

Let spherical waves from a point source S meet an opaque obstacle the section of whose edge is represented at A; we shall study the effects at various points on a screen back of the obstacle. Straight lines drawn from S through the points on the edge of the obstacle to the screen mark what is called the "geometrical shadow." Thus drawing the line $\overline{SAB}$, B is a point on the edge of the shadow. At any point P, outside this, the effect due to a wave front that just reaches the obstacle is found by considering the zones drawn

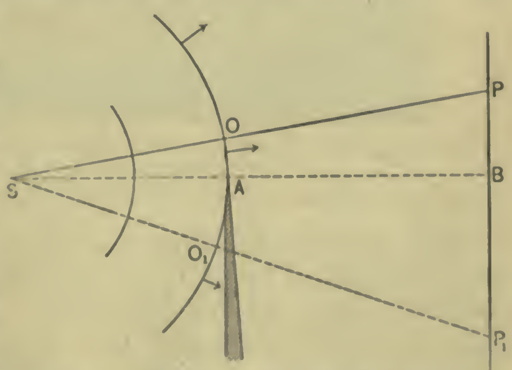


FIG. 175. — Diffraction past a sharp edge of an opaque obstacle. PBP_1 is a screen. B is the limit of the geometrical shadow.

around O, where this is the intersection of the wave front by the straight line $\overline{SP}$. It is seen that only a limited number of *complete* zones can be drawn, that have an effect; for the others will be cut off more or less by the opaque obstacle. If this number of complete zones is *even*, *e.g.* four, their effect may be written $m_1 - m_2 + m_3 - m_4$, and is therefore almost zero; while, if the number is *odd*, *e.g.* three, their effect may be written $m_1 - m_2 + m_3 = \frac{1}{2} m_1$, nearly, and so is great. Consequently, as the point P is taken farther and farther out from B, the intensity passes through a series of maxima and minima; and, when it is some distance out, the effect becomes uniform and is uninfluenced by the obstacle.

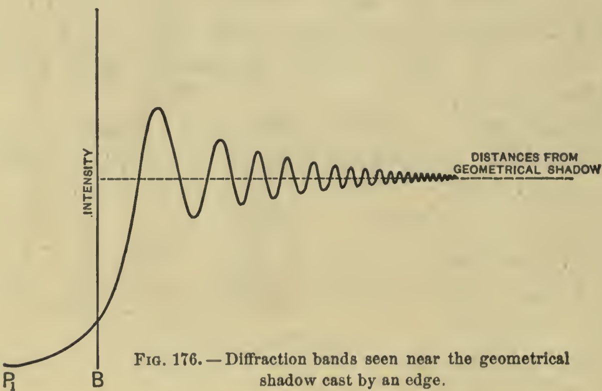


FIG. 176. — Diffraction bands seen near the geometrical shadow cast by an edge.

At any point P_1, *inside* the geometrical shadow, the effect may be considered as due to the zones around O_1, where the straight line $\overline{SP}_1$ meets the wave front. A number of these zones are behind the obstacle; but a certain one, say the nth, will produce an effect, because it emerges sensibly above the edge at A. The total effect at P may then be written $m_n - m_{n+1} + m_{n+2} \cdots$, which equals $\frac{m_n}{2}$ approximately; so there is an effect at P_1 which becomes less and less *continuously* as the point recedes into the shadow, corresponding to the increase in n and the consequent gradual decrease in m_n. Therefore, the disturbance from S penetrates into the geometrical shadow. The effect may be represented by a curve so drawn that its height above a horizontal line at any point indicates the intensity of the disturbance at that point. Thus, in Fig. 176, the points P_1 and B are the ones represented by the same symbols in the previous cut. It is seen that the effect in the geometrical shadow is as if there were a source of waves at the point A, the edge of the opaque obstacle.

FIG. 177. — Photograph of diffraction bands near the geometrical shadow of a sharp edge.

Naturally, everything depends upon the magnitude of the wave length of the waves. If this is small, as in light, the waves penetrate a very short distance into the geometrical shadow; or, better, the intensity of the waves decreases so rapidly inside the shadow that they can be perceived for only a short distance; and the region outside the shadow in which there are variations in the intensity, *i.e.* where there are "bands," is extremely limited. Both of these phenomena may be easily observed in the case of a shadow cast on a screen by any opaque obstacle if the source of light is small.

If we are dealing with air waves about 1 cm. long, all of the above phenomena, as described for light waves, are easily observed by the use of a sensitive flame; and, if the waves are several metres long, "sound shadows" may be observed if the obstacle is sufficiently large, *e.g.* a mountain. But we see that, if the waves are long and the obstacle is of ordinary size, the former will penetrate a great distance in the shadow.

This phenomenon of the peculiarities of a shadow produced by trains of waves incident upon an obstacle is called "diffraction." It was first described in the case of light by the Italian priest Grimaldi in 1666; but its explanation was given by Fresnel.

Diffraction through a Small Opening. — An important illustration of diffraction is afforded when a train of waves falls upon an opaque obstacle which has in it a single *small* opening. The general features of the phenomena may be deduced easily. Let the waves come from such a direction that their wave front is tangent to the plane of the opening. Draw the line $\overline{OB}$ perpendicular to this plane, and $\overline{OP}$ oblique to it; we shall consider the effect at various points on a screen perpendicular to $\overline{OB}$. The effect at B depends upon the number of zones

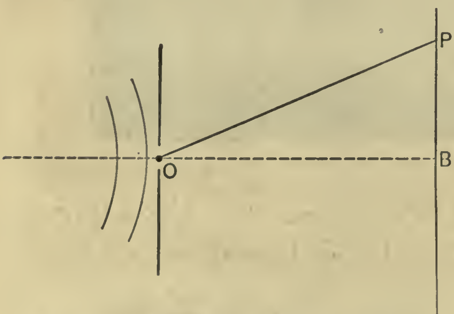


FIG. 178. — Diffraction through a small opening O.

that can be drawn in the *small* opening at O; if it is even, the effect is zero; if it is odd, the effect is practically as great as if the whole obstacle were removed, viz., $\frac{1}{2}\, m_1$. The effect at P in a similar manner depends upon the number of zones that can be drawn for it in the opening. The centre of the zones is the *pole* of P on the wave front, and is therefore far from the opening; and, since the edges of the zones get closer and closer as one recedes farther from the centre of the zones, P will have more zones — or rather portions of them — included by the opening than does B. So, if B has three zones in the opening, it will receive a maximum effect; and, if P is sufficiently far away from B, it will have portions of four zones included, and so will receive a minimum effect; farther out still, there will be a point for which there will be portions of five zones in the opening, and which accordingly receives a maximum effect; this is, however, much less than that at B, owing to the inclination of the zones and to the fact that only *portions* of zones are intercepted by the opening; etc. Consequently, if the opening is circular, the "diffraction pattern" on the screen will consist of concentric circular "bands" in which the intensity is alternately a maximum and minimum, but which rapidly fade in intensity as one recedes from their centre at B. It is also evident from the same reasoning that

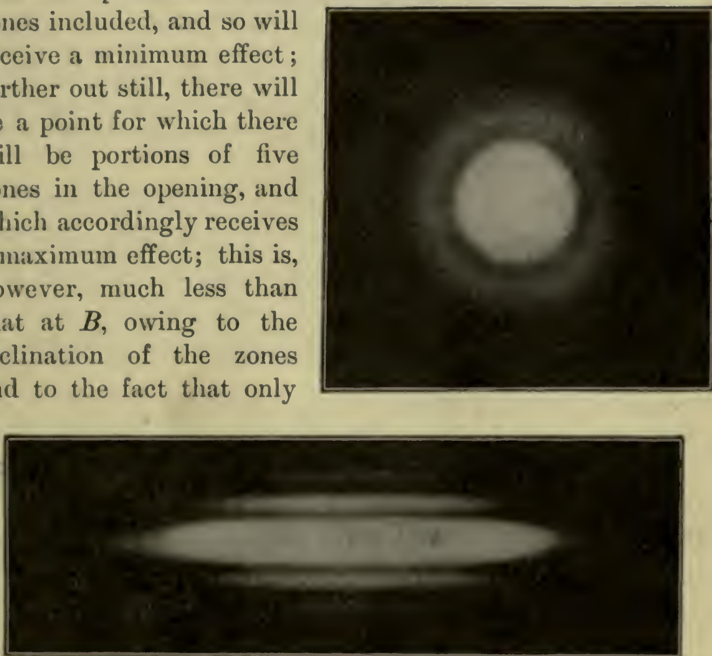


FIG. 179. — Photographs of diffraction pattern of a small circular opening and of a single narrow slit.

the smaller the opening, the farther apart are the bands; and, if the opening is sufficiently small, B will receive a maximum effect, and the first ring of minimum intensity recedes to an infinite distance; so all the points on the screen receive disturbances from the opening just as if it were a minute source of waves (see page 375).

This is a familiar fact, because if a *small* "pin hole" is made in a cardboard or in any *thin* opaque screen, and a light of any kind is placed behind it, it appears bright when viewed from any direction on the farther side. (This is, of course, entirely apart from any light *reflected* from the edges or sides of the opening.) Similarly, if a room has a small opening in it, and if a tuning fork is vibrating in the room, it may be heard outside in all directions from the opening.

The treatment just given of diffraction past an edge and through a small opening must not be regarded as rigorous. It is based upon many assumptions, which have been purposely omitted, and which are not all justified. On the whole, however, the results are sufficiently accurate for our present purposes.

Criteria of Wave Motion. — The existence of any of the phenomena that have been described in the foregoing articles — vibrations with nodes and loops, diffraction, interference — is proof positive of the presence of wave motion. This fact will be made use of immediately in describing and explaining the phenomena of Sound and Light.

SOUND

CHAPTER XXII

ANALYSIS OF SOUND

Fundamental Facts. — To one who has the sense of hearing the word "sound" has a definite meaning, as describing a certain sensation, but it is impossible to define it so as to convey an idea to one who is born deaf. The cause of this sensation can always be traced to some body that is vibrating rapidly. Thus, if we hear a sound from a tuning fork, a piano string, a metal bell, etc., it is a simple matter to prove that they are vibrating, and that, when the motion ceases, the sound does also. Again, it is a familiar fact that some time elapses between the instants when the vibration begins and the sound is heard, and between those when the former stops and the latter ceases; for there is a considerable interval of time between the instants when a distant gun is seen to be fired and when the sound of the report is heard, or when a distant steam whistle is seen to blow and when the noise is heard, etc. This proves that whatever causes the sensation requires time for its transmission through space. The fact that the production of the sensation depends upon the presence of a *material* medium between the vibrating body and the ear may be proved by suspending the vibrating body in a space from which the air may be more and more exhausted; as this is removed, leaving only the ether, the sound becomes less and less intense. Vibrating bodies will produce compressional waves in a surrounding fluid, provided the frequency of vibration is sufficiently great; and the fact that the sensation of sound is due to these waves may be proved most simply by allowing them to produce vibrations with nodes and loops, and showing that they cause sounds. Methods of doing this will be described in a few pages.

We shall discuss first the characteristics of different sounds and the physical cause of these differences, then describe a few typical musical instruments and some acoustic phenomena, and finally give the physical explanation of harmony in musical compositions, with a brief description of musical scales.

Noise and Musical Notes. — If we analyze our sensations of sounds, we are led at once to recognize two great classes which we call in ordinary language *noises* and *musical notes.* The latter have all the characteristics of periodic motion; they are continuous and uniform in character, and are pleasant to the ear. The former are discontinuous, with abrupt changes, and are often extremely disagreeable to the ear. Thus, the sounds due to a tuning fork, to a piano string, to the column of air in an organ pipe, etc., are musical notes. But the sounds heard when a piece of paper is torn, when a wagon rolls over cobblestones, when a slate pencil is sharpened, etc., are noises. We can study the nature of the vibrations of a body, as has been explained on page 319, and it is found that a musical note is always due to a periodic vibration; a noise, to an extremely complex motion, consisting of different vibrations which differ slightly in frequency and which are rapidly damped.

A confused vibration which causes a noise will produce a musical note, if the vibrating body is near a flight of steps; for, when the pulse reaches the first step, a reflected pulse is produced; and in a similar manner others are produced when it reaches in turn the other steps. Therefore there will be in the air a series of reflected pulses at exactly equal intervals apart; and, as they reach the ear, a musical sound is heard. Thus, if one claps one's hands near a staircase, the noise is first heard, but it is followed immediately by a musical note. The same phenomenon occurs if a noise is produced near a picket fence.

Simple and Complex Notes; Quality. — If we analyze our sensations of different musical notes, we can separate them into two classes: one we call "pure" or "simple"; the other, "complex." Thus, the note produced by vibrating metal plates like cymbals is complex, while that due to a tuning fork or to a stopped organ pipe is pure. In fact, all notes, with the exception of these last, are more or less complex. If we examine the corresponding vibrations, it is found that a pure note is always due to a simple harmonic vibration, and a complex note to a complex vibration.

Complex vibrations can be analyzed, in accordance with Fourier's theorem, into simple harmonic components whose frequencies are in the ratio 1 : 2 : 3 : etc., or into other components not so related. In a similar manner, if one listens attentively to a complex sound, various simple pure notes may be distinguished. (This statement that the human ear analyzes mechanically a complex wave into simpler components and hears the corresponding simple notes separately is known as Ohm's Law for Sound.)

If two complex notes differ, it is found that the corresponding complex vibrations differ; but the converse statement that two different complex vibrations produce two different complex notes is true only with one limitation. If the component parts of the two complex vibrations differ in their frequencies or in their amplitudes, the corresponding notes are different; but the differences in phase between the components may be different in the two complex vibrations or may be the same; they have no influence on the note. (This is seen to be in accord with Ohm's law, as stated above.)

Two different complex notes are said to differ in "quality." Thus, the complex notes produced by the vibrations of the column of air in an organ pipe, of a violin string, of a piano string, of the column of air in a horn, etc., all differ in quality; and it is by this property that we recognize the nature of the source of the sound. This quality depends upon the number and amplitudes of the other component vibrations besides the fundamental present in the complex vibrations; it does not depend, however, as has been already said, upon the relative phases of these component vibrations.

Pure notes are never used in music, because they lack what may be called "character," or individuality. Notes that are useful for musical compositions must be complex to a certain extent, seven or more components often being present.

Analysis of Notes. — This process of analyzing a complex sound or a complex vibration into its harmonic components is greatly helped by the use of resonators. Helmholtz in his epoch-making work used those of the form shown in the cut. (One of their advantages is that when the inclosed air is set in vibration, the motion is simple harmonic.) He constructed a set of them, and accurately determined the frequencies of the vibrations of their inclosed volumes of air. Then by bringing them in turn near a vibrating body, he could tell by holding one of the ends of the resonator near the ear whether the particular vibration that corresponded to that of the air in the resonator was present in the complex vibration of the body; for, if it were, the corresponding sound would be intensified. In this manner it is a simple matter to detect the components. In the best and most recent work on analysis of sounds, phonographs are used, and the traces are magnified.

Fig. 180. — A Helmholtz resonator.

Several other illustrations of resonance are worth mentioning. When a large seashell or a vase is held near the ear, the roaring that is heard is due to the resonance of the inclosed air produced by certain sounds in the room. The sound may be varied by partly closing the opening of the shell or vase. The passage leading from the outside of the head into the eardrum forms a small resonator, and its action is often noticed when one is listening to an orchestra, by the strong resonance of certain very shrill sounds like the buzzing of insects.

Helmholtz performed with his set of resonators the converse of the analysis of a complex sound; he produced one by means of the superposition of simple harmonic vibrations or of pure notes. He arranged in front of each resonator a tuning fork whose frequency of vibration was the same as that of the air in the resonator, and adjusted electro-magnets to these forks in such a manner that he could set them vibrating and maintain them in motion. He could also alter their amplitude. Then by making different forks vibrate he was able to produce different complex sounds, and in fact to imitate the sounds characteristic of different instruments.

In a complex sound, the component note corresponding to the fundamental vibration is called the "fundamental"; and the other notes "overtones." If the component vibrations form a harmonic series, the component notes are also called "harmonics."

Pitch and Loudness. — If we compare two simple notes, we recognize the fact that they may differ in two ways, in shrillness and in loudness. Thus, the notes of a piccolo are shriller than those from an organ pipe; and any note of an organ may keep the same shrillness and yet may vary in loudness. If we compare the corresponding vibrations, we find that in every case if one note is shriller than another, the frequency of its vibration, or rather the number of waves reaching the ear in a unit of time, is the greater. (This fact was first observed by Galileo.) Further, we find that if any one note decreases in loudness, its vibration decreases in amplitude, other things remaining unchanged.

We cannot measure the shrillness of a note, because we cannot imagine a unit of shrillness nor the idea of shrillness being made up of parts which can be compared. We can, however, give a number to the shrillness of a note by assigning it one equal to the frequency of the vibration, if the vibrating body and the observer are at rest relatively; or, more generally, we assign a number equal to the number of waves reaching the ear in one second. (See Doppler's Principle, page 345.) This number is called the "pitch" of the sound. Similarly, the pitch of a complex note is defined to be the pitch of its fundamental.

The loudness of a sound, either pure or complex, varies as the intensity of the waves producing it. It is thus seen why, when a sounding body approaches the ear, the loudness of the sound heard increases. We can measure the intensity of the waves, but we have no method of measuring the loudness, for this is a sensation, and not a physical quantity.

Audibility of Waves. — In order to produce waves in the air, the frequency of the vibration must exceed a certain limit, as has been explained ; otherwise the air flows, but is not compressed. But all waves in the air do not affect our sense of hearing ; for this sense depends upon disturbances being conveyed to the nerve endings from the external air by a mechanism whose parts are set in motion by waves of certain wave lengths, and not by others. Thus, it is found that waves whose wave number is greater than 20,000 per second or less than 30 do not in general produce sounds ; but, of course, these limits are only approximate, and vary greatly with different individuals. In musical compositions as played by orchestras the maximum range of pitch is about from 40 to 4000.

The Human Ear. — For a full description of the human ear reference should be made to some treatise on Physiology ; it is necessary to mention only a few details here. The ear consists of three parts : the external ear, which ends at the eardrum; the middle ear, which is connected with the throat and mouth by a tube, and in which there are three little bones with flexible connections, thus making a mechanism joining the drum to a membrane which closes one opening into the third portion of ear; the inner ear, which is entirely inclosed in the bone of the skull, and which consists of several cavities filled with a liquid. In one of these cavities there is a series of minute fibres of regularly decreasing length, with which the nerve endings of some of the auditory nerves are connected, and which are thought to play the part of resonators for musical notes. Other branches of the auditory nerve end in another cavity under conditions which have led several scientists to believe that their function was to respond to noises. In any case it is easy to trace a mechanical connection between the waves in the air and the nerve endings through the eardrum, the three bones, and the liquid in the inner ear. The student will find in the work of Helmholtz a full discussion of these various steps.

CHAPTER XXIII

MUSICAL INSTRUMENTS

We shall discuss only two types of musical instruments: stringed and wind instruments. The commonest stringed instruments are the piano, the violin, the violoncello, and the harp; and the commonest wind instruments are the organ, the flute, and the horn.

Stringed Instruments. — For present purposes we may regard the vibrations of the strings in any stringed instrument as being identical with those of a perfectly flexible cord, although in reality musical strings are far from being perfectly flexible, and their elasticity plays a part in addition to their tension. Only transverse vibrations are ever used. We saw on page 353 that the frequency of the fundamental vibration of a cord was given by the relation $N = \frac{1}{2L}\sqrt{\frac{T}{d}}$, where L is the length of the cord, T its tension, and d its mass per unit length; and that the partials had frequencies $2N$, $3N$, etc. This formula explains how a string may be "tuned" by altering its tension; how its frequency may be altered by shortening its length, as is done in violins, etc., by means of the fingers; and why the different strings of any one instrument are made of different densities.

When a string is struck at random or is plucked, the vibration is complex; but those components are absent which would have a node at the point struck or plucked; thus the quality of the note depends largely upon where this point is, as is shown in the use of violins. The difference between pianos as made by different makers lies to a great extent in the point of the strings which is struck, in the size and hardness of the "hammer," and in the duration of the blow; for these all influence the quality of the notes heard.

All stringed instruments, with a few exceptions, have the strings stretched between pegs which are fastened to a wooden board or box. Owing to the vibrations of the strings, and the resulting motion of the pegs, this board is made to vibrate; and, since these vibrations are "forced," they imitate more or less closely in character those of the strings. But, of course, there are differences depending upon the thickness, area, stiffness, etc., of the boards. Similarly, if there is a box or cavity, the inclosed air may be set vibrating. The vibrations of this "sounding" board or box affect the surrounding air much more than do those of the fine string; and so the sound we hear depends to a great extent upon the former vibrations. We see, therefore, the reasons why the violins of certain makers have such great value, owing to their skill in the construction of the wooden parts.

Wind Instruments. — We have given the theory of the vibrations of a column of air on pages 353–358, and have shown that in the case of a column open at both ends the frequency of the fundamental is $N = \frac{V}{2L}$, where V is the velocity of air waves, and L is approximately the length of the column; and that the partial vibrations have frequencies equal to $2N$, $3N$, etc.; whereas, a column which is closed at one end — a "stopped" pipe — has for the frequency of its fundamental $N = \frac{V}{4L}$, and for those of its partials $3N$, $5N$, etc. When a column is set vibrating, as a rule both the fundamental and partials are present; but the fundamental is in general much more intense than the others, and the partials decrease in intensity as their frequencies increase. So, when an organ pipe is blown gently, only the fundamental is heard and the note is almost pure; this is specially true of a stopped pipe because the vibration whose frequency is $2N$ is absent; and for this reason these pipes are not generally used in orchestras. With but little practice, one may learn to blow a pipe with such pressure as to make the fundamental or a particular partial the most intense.

We see from the formula that the frequency does not depend upon the material of the solid inclosing the column of air nor upon its cross section. (Of course the "correction" for the end by which L is increased by $0.57\ R$, as explained on page 354, is affected slightly.) If the length of the pipe is changed, the pitch of the corresponding sound is altered; and so pipes of different lengths are used in organs, and by making small changes a pipe may be "tuned" accurately. If the pipe is not of uniform cross section, but conical, there are marked differences produced — the position of the nodes and loops is affected; as is shown by the difference between an organ note and one produced by a horn. The action of "stops" or "pistons" in horns is to vary the length of the vibrating column of air.

If an opening is made in the side of the tube containing the column of gas, the vibrations must adjust themselves to this point being a loop; and if two openings are made, their distance apart determines the length of a vibrating segment and therefore the frequency of the vibration. This explains the effect of making and closing openings in a tube as is done in flutes and similar instruments.

Organ Pipes. — The column of air in a pipe or tube is set vibrating in several ways. In the ordinary organ pipe, a section of which is shown in the cut, there is a narrow passage leading to the bellows or wind chest, through which a blast of air is directed against a sharp lip forming the upper edge of a narrow opening at the bottom of the pipe. This blast at the beginning of operations sends a disturbance up the tube, which is reflected at the upper end and returns.

When it reaches the bottom, its effect is to deflect the blast out of the opening in the tube. When this effect ceases, the blast returns; and so there is an oscillation of the blast, which has a period determined by that of the stationary vibration of the column of air produced by the two trains of waves, the direct and the reflected. This bottom of the tube is approximately a loop, because it is open to the air. The vibration of the air would soon cease owing to loss of energy by friction and by the production of waves, were it not for the supply furnished periodically by the blast. The same method of producing vibrations is used in flutes and whistles, the lungs or mouth being the wind chest.

FIG. 181. — Section of an ordinary organ pipe.

Reed Pipe. — In another form of organ pipe, known as the reed pipe, the wind chest is connected directly with an elongated box; and into this is inserted an inner tube, in which there is a rectangular opening closed by a strip of stiff brass fastened at one end, called the "reed." When the pressure in the wind chest is sufficient, this spring door is pushed open, a blast of air passes, and the pressure being relieved slightly, the spring returns and, owing to its inertia, continues to vibrate in its own natural period. In this manner a series of puffs of air are delivered at intervals, determined by the frequency of the metal spring. There is always attached to the pipe a resonance tube of some kind, the air in which is set in vibration owing to the intermittent puffs. Without this box the sound is most complex, but with it the note becomes fairly simple. The instrument is "tuned" by altering the length of the reed by a clamp.

Horns. — In the case of horns, the vibration of the column of air is produced by means of the vibrations of the lips of the player. The column in a horn of fixed length can vibrate in only a limited number of ways; and the lips must be stretched to exactly the right degree so that, when they are set vibrating by air from the mouth being forced through them, their frequency is one of those to which the horn responds. If one is playing a horn of variable length, like a trombone, and a definite note is being produced, a change in the length changes the frequency of the vibration, and therefore requires a change in the frequency of the lips; but this change is produced almost automatically, owing to the reaction of the column of air itself.

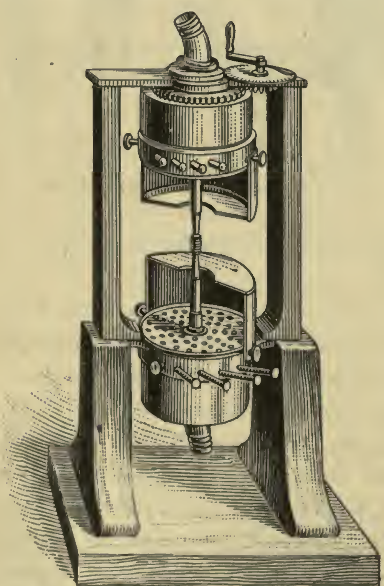


Fig. 182. — Helmholtz's double siren.

The Siren. — There is another acoustic instrument which, although not, strictly speaking, a musical one, should be described. It is called the "siren," because its action continues under water as well as above. In principle it is not unlike a reed pipe, inasmuch as it is designed to deliver a number of puffs of air at regular intervals, only with it this number can be varied at will and can be easily counted. As shown in the cut, there is a wind chest, which is closed on its upper side by a thick circular plate perforated with a definite number of holes, at regular intervals, around a circle concentric with the plate itself. These passages in the form of instrument pictured are not perpendicular to the plate, but are inclined slightly, so that the axis of the passage con-

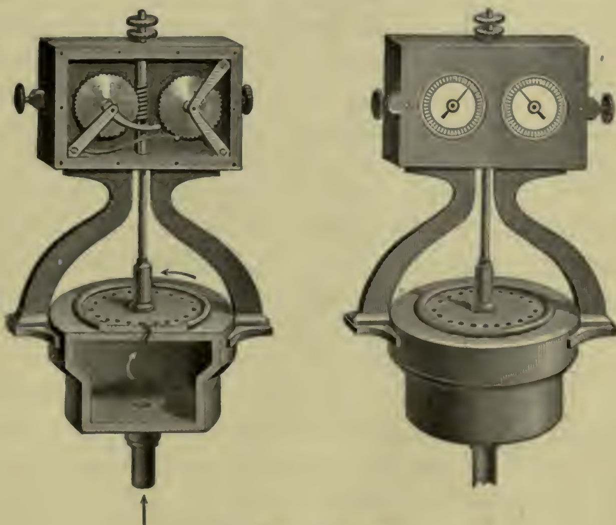


Fig. 182 *a*. — The siren.

sidered as a vector has a component parallel to the tangent of the circle through the holes at that point. Immediately over this fixed plate is a movable one, which can rotate on its axis of figure and which is identically like the fixed plate, except that its passages slope the opposite way. Therefore, if the movable plate is in such a position that its openings are over those in the fixed plate, the air rushing out from the wind chest will have a momentum against the sloping sides of the passages in the former plate, and will set it in rotation on its axis. Each time the openings in the two plates coincide, a puff of air escapes; and if there are n openings in each plate, there will be n puffs during each rotation. (This would be true, also, if there were n holes in one plate only and but one in the other; if there are n holes in each, the intensity of the puffs is increased.) So, if the rate of rotation is m turns per second, the number of puffs in a second is mn; and this is therefore the pitch of the resulting sound. The speed of rotation may be altered at will by regulating the pressure of the air in the wind chest. The number of openings in the plates is easily observed; and the number of revolutions in any interval of time is determined by using a mechanical counter, such as are seen on steam engines. (A screw thread is cut on the shaft of the rotating plate, so that a worm wheel is turned; this drives a train of cogwheels, which moves a hand over an indexed dial, like the face of a watch.) In other forms of this instrument, the passages in the plates are not slanting, and the movable plate is made to rotate by means of a mechanical or electric motor.

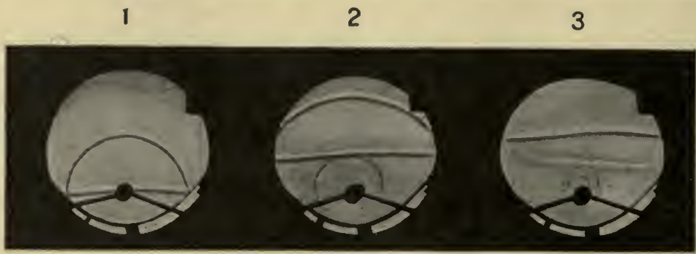
The simple siren as just described has been modified in two ways. One is to make in the plates several concentric rows of openings, which contain different numbers, *e.g.* 8, 10, 12; and thus, if all these are opened at one time, a complex sound is heard whose component simple vibrations have frequencies in the ratio 8 : 10 : 12. Again, two sirens connected with the same bellows may be arranged one over the other with their movable plates on the same shaft and facing each other; this enables one to produce two sounds whose pitches have a known ratio, that is, are at a known "interval" apart.

This instrument was invented by the German physicist Seebeck, and was improved by Cagniard de la Tour and more recently by Helmholtz. It is not, however, as much used now as formerly.

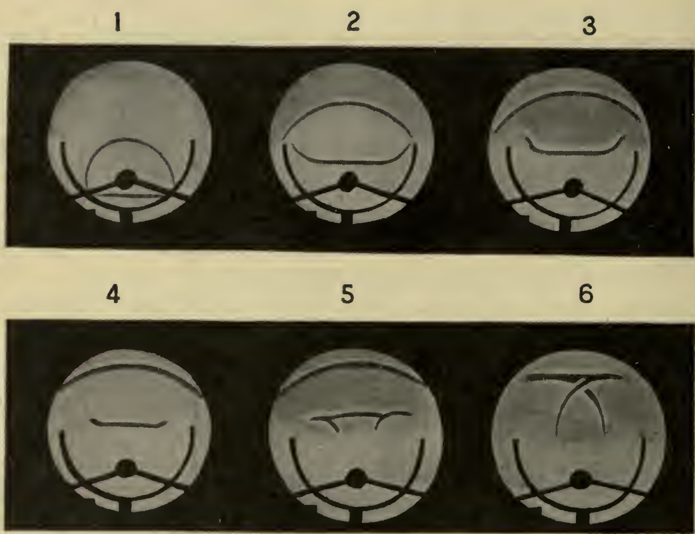
Phonograph. — Another acoustic instrument is the phonograph, which consists essentially of a hardened wax cylinder against whose surface presses a sharp point connected with a flexible membrane forming part of a mouthpiece. The cylinder is turned and advanced by clockwork; and, as sounds are produced near the mouthpiece, the point makes corresponding indentations in the wax, which are faithful reproductions of the displacements of the vibrating body. Then, if another point attached to a membrane is made by mechanical means to pass over these traces on a cylinder at the same rate as that at which the cylinder was turned originally, the membrane will be set in vibration, and its motion will therefore be very nearly the same as that of the one in the mouthpiece that caused the trace. These vibrations will then produce waves in the air which will affect the ear; and so the original sound is reproduced.

The Human Voice. — The human voice is due to the vibration of the vocal cords of the larynx and of the various movable parts of the mouth. Consonant sounds such as *b*, *c*, etc., are produced by vibration of the lips, the tongue, etc.; while vowel sounds owe their origin directly to the larynx. This consists of two stretched membranes which have free edges along a nearly straight line. These can be set in vibration by the air pressure in the lungs; and their frequency can be altered by voluntary changes in their tension. Owing to these vibrations, which are, however, very "damped," the air in the cavities of the mouth and throat is also set vibrating. When a definite vowel sound, such as *ah*, is produced, no matter what the pitch of its most prominent component, it is found by analysis that there are present one or more components of definite pitch. These are not partials of the fundamental vibration of the larynx reënforced by resonance, but are independent vibrations; and the vowel character of the sound depends upon them; that is, different vowel sounds have different permanent components.

Reflection and Refraction as Phenomena of Sound. — Since sounds are due directly to waves in the air, all the properties of wave motion may be studied and illustrated by sound



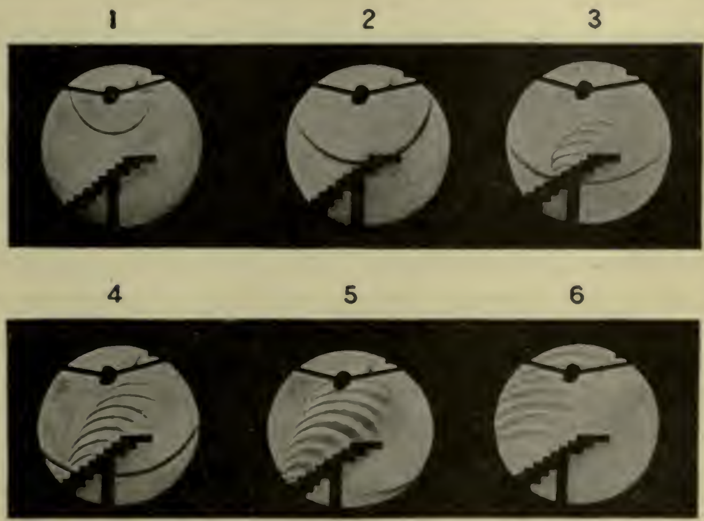
Parabolic mirror; centre of disturbance at its focus.



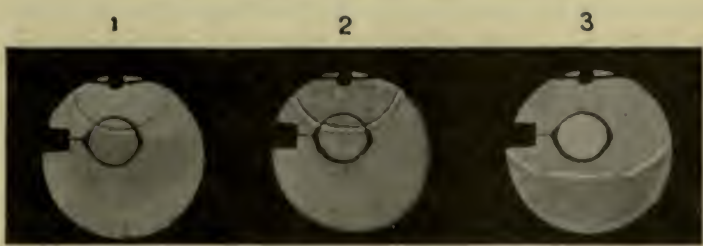
Spherical mirror; centre of disturbance at its focus. (Notice the effect at the edges of the wave-front.)

Fig. 183.

phenomena. Thus the reflection of waves is shown by echoes, by the use of concave mirrors, by sounding boards, etc. If a disturbance is produced near a curved wall, waves will spread out and will in part be reflected. Those waves falling very obliquely upon the wall may by reflection meet the wall again, be reflected again, etc. Thus a certain amount of the energy of the waves will follow around the



Reflection of a single pulse from a flight of steps, thus producing a regular series of pulses. A noise may thus by reflection cause a musical sound.



Refraction of a pulse by a lens containing a heavier gas.

FIG. 188 *a*.

wall. This is the explanation of "whispering galleries." Similarly, in "speaking tubes" and "speaking trumpets" the energy of the waves is confined within certain bounds, instead of spreading out in all directions.

An interesting case of irregular reflection is seen in the case of air waves passing through a *non-homogeneous* atmosphere occasioned by air currents of different density or by the presence of clouds. The waves may be reflected or scattered, as in the "rolling" of thunder; and many similar phenomena have often been observed with signals from fog horns, as shown by Henry and Tyndall. Sometimes the path of the waves is bent so as to rise in the air and then descend, causing regions or islands of silence. (The presence of fog as such does not give rise to these phenomena; for it does not make the atmosphere non-homogeneous.) One illustration of refraction has been given on page 331, where it was shown how winds may change the direction of a plane wave front. This, however, is what may be called "mechanical" refraction. True refraction, similar to that observed in Optics, may, however, be produced. Tyndall made a lens out of a soap bubble filled with nitrous oxide gas, which had all the properties with air waves that an ordinary glass lens has for ether waves.

A few photographs of pulses in the air, illustrating these and other properties of wave motion are added. They were obtained by Professor R. W. Wood, using a method devised by Toepler. These pulses are produced by the explosive action of an electric spark. The black knobs seen in the photographs are the centres of the disturbances; and the wave fronts may be distinguished clearly.

The Acoustic Properties of Halls. — Something should be said also in regard to the acoustic properties of halls that are used either for public speaking or for concerts. Great care must be exercised to avoid what is called "reverberation." This, if excessive, is a great objection. It is due, of course, to the echoing and reëchoing of the sounds, occasioned by the repeated reflection of the waves from the floor, walls, ceiling, seats, auditors, etc. The problem of investigating the exact conditions that determine or prevent reverberation was first undertaken by Professor Sabine of Harvard University in the year 1895. His method of studying it was to arrange an organ pipe at one point of a hall so that it could be blown and then stopped at any instant; stationing himself in turn at different positions in the hall, he would note how many seconds he could hear a sound after the organ pipe had ceased acting. It was found that the reverberation was the same practically at all points in the hall and that it was independent of the position of the pipe. For halls of the same *volume* the reverberation is the same; but, as the size of the hall increases, the reverberation increases also, other things remaining unchanged. It is decreased greatly by putting soft coverings on the floors, walls or seats, by making the walls less rigid, and by the presence of an audience. It was found that for practical purposes the reverberation should not be decreased below 2.3 seconds, otherwise the music was not fully appreciated by the audience. Professor Sabine was able to deduce a general formula which can be used to predict with great exactness the duration of reverberation in a hall when its dimensions and the materials used in its construction are known.

The acoustic property of a hall depends upon other things than reverberation, for its shape may be such as to focus the waves at particular points, etc. "Sounding boards" which reflect the waves down upon the audience are often used. Another effect which must be taken into account is that due to ascending and descending currents of air; for wherever there are changes in the homogeneous nature of the air, there are reflections of the waves.

CHAPTER XXIV

MUSICAL COMPOSITIONS

Combinational Notes. — When two instruments are sounding at the same instant, there are several interesting phenomena besides the production of the two sounds. If the two instruments are setting in vibration directly the same portion of air, as when a double siren is used, or if two wind instruments are blown by the same wind chest, other sounds are heard than the two corresponding to the instruments. Thus, if n_1 and n_2 are the frequencies of the two vibrations, other vibrations of frequencies, $n_1 + n_2$ and $n_1 - n_2$, are produced in the air. (The mathematical theory of this was given by Helmholtz.) The corresponding sounds are called "combinational," or "summational" and "differential" notes.

Beats. — If the two vibrations have frequencies which are quite close together, a different phenomenon is observed. Thus, suppose the frequencies are n and $n + m$, where m is a small number — not necessarily an integer. Then, when the instruments are sounded at the same time, the loudness of the sound heard fluctuates; it rises and falls at regular intervals. If $m = 4$, these intervals are a quarter of a second; or, in general, this is $\frac{1}{m}$th of a second. There are then said to be 4 "beats" (or, in general, m "beats") per second. This "beating" is due obviously to the fact that as the two trains of waves traverse the same medium before they reach the ear, there will be points at regular intervals apart where the compression of one train will neutralize the expansion of the other, and, at points halfway between these, the compressions of the two will coincide. Thus, if the waves have a velocity V, and if the difference in the wave numbers is an *integer* p, in a distance V in the direction of propagation of the waves, there are p points where the disturbances are almost, if not quite, neutralized and p points where the disturbance is abnormally great. This is shown in the figure,

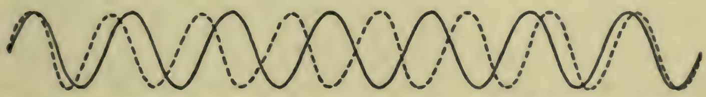


FIG. 184. — Diagram showing origin of beats. The two vibrations have frequencies whose ratio is 7 : 8.

where $n : n + p = 7 : 8$. Therefore, in the course of one second, as the waves enter the ear, it will happen p times that the sound almost vanishes and p times that it is abnormally loud. Thus, if we count the number of beats in a second, we can tell exactly the difference in the frequencies of the two vibrations. (If m is not a whole number, a distance greater than V must be chosen in the above treatment. If $m = 4.5$, a distance 2 V may be chosen, in which there are then 9 points where the disturbances neutralize each other, etc.) If the vibrations are complex, beats may occur owing to the proximity of the frequencies of any two of the component vibrations, or to that of the frequencies of any of these and those of the combinational ones. (The explanation of beats was first given by Sauveur, about 1700.)

It is recognized by every one that beats produce a disagreeable sensation, and for the same reason that a tickling feather or a rapidly flashing light do; namely, owing to fatigue of the nerves. When the beats become so rapid that they cannot be individually recognized, they cause a "roughness" in the sound which is unpleasant; but the degree of disagreeableness depends upon both the pitch of the sound and the number of beats per second. This matter was fully investigated by Helmholtz.

Harmony or Consonance. — It has been known to men of all nations since the earliest ages that there are certain combinations of vibrating bodies which produce sounds pleasant to the ear, such as that heard when two or three stretched strings of definite lengths are vibrated at the same time. This fact has nothing to do with the state of civilization or of musical cultivation; it is a property of the human ear. It was recognized by the Greeks before the days of Aristotle (probably as long ago as 500 B.C.) that, if two stretched strings of the same size and material and under the same tension, but of lengths in the ratio 1 : 2, were sounded together, the sound heard was agreeable; and also that, if there were three similar strings whose lengths were in the ratios 4 : 5 : 6, the same was true. Mersenne, in 1636, showed that the frequencies of the vibrations of stretched strings varied, other things being equal, inversely as the lengths of the strings. So the problem of explaining the consonance of the sounds produced by the two or three strings as just described became this: why should two notes produced simultaneously by strings making vibrations of frequencies n and $2n$, or three notes produced simultaneously by strings making vibrations of frequencies $4n$, $5n$, $6n$, cause a pleasant sensation?

Helmholtz's Explanation of Consonance and Discord. — The answer to these questions was first given by Helmholtz. He showed that in every case of consonance when two or more notes are produced simultaneously, that is, a "chord" consisting of two or more notes, beats were nearly, if not entirely, absent; and that in any other case, when two or more notes were sounded together, the degree of the discord could be predicted from calculations of the number of beats present and from a knowledge of the degree of their disagreeableness to the ear. Thus, when a note of pitch n is produced by a vibrating string, notes of pitch $2n$, $3n$, etc., may be heard in the complex sound; and, when a note of pitch $2n$

is produced, others of pitch $4n$, $6n$, etc., may be heard; further, the combinational notes all have pitches n, $2n$, etc.; so there are no beats, and the two complex notes of pitches n and $2n$ are therefore in harmony when produced by wind or stringed instruments. But, if the lower note had the pitch $n+5$, its partials would have the pitches $2n+10$, $3n+15$, etc., and these would beat with the other note of pitch $2n$ and with its partials. Therefore, if two complex notes of pitch $n+5$ and $2n$ are produced, they are discordant.

It is evident that the same explanation applies to the harmony of the chord consisting of the three notes of pitch $4n$, $5n$, and $6n$. In general, two or more notes are consonant, or approximate to it, if their pitches bear simple numerical relations to each other, such as 1 : 2, 1 : 3, 2 : 3, 1 : 4, 2 : 5. If these relations can be expressed, however, only in terms of large numbers, the notes are discordant. On this fact is based the construction of all musical "scales," that is, series of notes of different pitch, which are played in chords in musical compositions.

In all music, however, another question enters, namely, are two or more notes or chords played in rapid *succession* pleasant to the ear or not? If they are, they are said to form a "melody." These questions belong to the Theory of Music and cannot be discussed here; but, in general, if two notes are consonant, they will also form a melody.

Musical Scales. — Many musical scales have been devised and used; but only two will be discussed here. All modern musical compositions are based upon the idea of selecting some note as a "keynote," and using this as the starting point of the scale. If N is the pitch of the keynote, a note of pitch $2N$ is called its "octave"; and the interval between them is called "an octave." The "interval" between any two notes is defined to be equal to the ratio of their pitches. In the formation of any scale, a definite number of notes is selected in an octave, whose intervals apart do not differ too widely, and which when played in chords do not produce too discordant sounds. Let their pitches be N, O, P, Q, ⋯, $2N$. Then, in the octave above this, that is, in the interval from $2N$ to $4N$, the notes selected for the scale have the pitches $2N$, $2O$, $2P$, $2Q$, ⋯, $4N$; those in the octave above this, $4N$, $4O$, $4P$, $4Q$, ⋯, $8N$; etc. The notes selected for the octave below the original one have the pitches $\frac{1}{2}N$, $\frac{1}{2}O$, $\frac{1}{2}P$, $\frac{1}{2}Q$, ⋯, N, etc. So, in defining a scale, it is necessary to choose only the pitches of the keynote and the notes in its octave.

The Diatonic Scale. — The "diatonic" or "natural" scale consists of a series of notes whose pitches may be expressed as follows: If $24n$ is the pitch of the keynote, the notes in its octave have the pitches $24n$, $27n$, $30n$, $32n$, $36n$, $40n$, $45n$, $48n$. Thus it is seen that in the interval of an octave there are seven notes, counting only one of the end notes of the octave.

If an instrument with strings of fixed lengths, like a piano, is to be constructed to play music written on this scale, some definite keynote must be chosen, and a string must be selected of such a length and size that under suitable tension it will give this note; then other strings must be selected to produce the other notes of the scale. But suppose a musical composer did not wish to use the same keynote for all his pieces; suppose, for instance, that he wished to have as the key tone one whose pitch is $20n$. The diatonic scale of this is

$$20n,\ \tfrac{27}{24}\times 20n,\ \tfrac{30}{24}\times 20n,\ \tfrac{32}{24}\times 20n,\ \tfrac{36}{24}\times 20n,\ \cdots,\ 40n,$$

or,

$$20n,\ 22\tfrac{1}{2}n,\ 25n,\ 26\tfrac{2}{3}n,\ 30n,\ \cdots,\ 40n;$$

and it is seen that, if the instrument described above is provided with strings giving notes in the diatonic scale having $24n$ as the keynote, many of the notes required for this composition cannot be produced; for instance, $25n$, $26\frac{2}{3}n$, etc. For this reason, and also to make the intervals between two consecutive notes more nearly equal, the diatonic scale was altered by the introduction of five new notes in each octave: between $24n$ and $27n$, $27n$ and $30n$, $32n$ and $36n$,

$36n$ and $40n$, and $40n$ and $45n$; and corresponding strings were introduced. But still the scale was unsatisfactory when musical compositions were written in different keys; and so a final change was made, which solved this difficulty, but introduced another less important one.

The obvious advantage of the diatonic scale is that chords played on it are as nearly in harmony as is possible for a scale having seven notes in an octave, since the ratios defining the scale are as simple as possible, viz. $24:36 = 2:3$; $30:40 = 3:4$; etc. But slight variations in the pitch of a note may occur in a chord without noticeable discord; and, in any case, people grow accustomed to a chord or to a piece of music and are not sensitive to its perfect harmony.

The Equally Tempered Scale. — This fact was taken advantage of in the construction of the new scale. It is called the "equally tempered scale," because the intervals between two consecutive notes are the same throughout the scale. Twelve notes are introduced in an octave; and calling the pitch of the keynote n and the constant interval a, these notes and the octave have the pitches: $n, an, a^2n, a^3n, a^4n, \cdots, a^{12}n$. But since $a^{12}n$ is the octave of n, $a^{12} = 2$ and $a = \sqrt[12]{2}$, or 1.0595. The notes in the octave above and below this are formed as before by multiplying and dividing by 2; etc. It is evident, then, that if a piano or organ is made with strings or pipes corresponding to the notes of this scale using any definite keynote, musical compositions written in any key whose note is anywhere in the scale may be played on it.

Standard Pitch. — The keynote adopted for musical scales by the Stuttgart Congress in 1834 and later by the "Society of Arts" in England is one whose pitch is 264. In England, however, at the present time "concert pitch" is based upon a keynote whose pitch is 273. For scientific purposes, tuning forks are used as standards of frequency or pitch; and they are generally made in sets following the diatonic scale. The most accurate forks in general use are those made by the late Rudolph Koenig of Paris; and he adopted as the fundamental frequency, or key tone, 256.

As will be explained presently, suitable names and symbols have been given the notes in the various octaves of different scales. For instance, "Ut_3" is sometimes used as the name of the key tone above described. So it is seen that the pitch corresponding to a certain name or symbol in musical notation is not definite. Thus the note called Ut_4 had a pitch less than 500 in the early part of the eighteenth century; in the days of Handel (1750) the note which had this same name and symbol in musical notation was one whose pitch was between 500 and 512; and in our days this same symbol is given a note whose frequency varies from 512 to 546, as we have seen above, for $Ut_4 = 2\ Ut_3$. These facts are expressed by saying that there is a tendency as years go by for the pitch corresponding to a given symbol to rise.

The octave of the equally tempered scale starting from the keynote 264 is made up, then, of the notes whose pitches are as follows: 264, 279.6, 296.3, 314.0, 332.6, 352.4, 373.3, 395.5, 419.0, 443.9, 470.3, 498.3, 528. If this note is used as the keynote of the diatonic or natural scale, the notes in the corresponding octave have the following pitches: 264, 297, 330, 352, 396, 440, 495, 528. It is thus seen how far apart the scales are at certain points.

Violins, violoncellos, etc., do not have strings of invariable lengths, because the fingers of the player can alter them at will by "stopping" at any point; and so one can play with them on the natural scale music written in any key, if the five additional notes are introduced, as previously explained.

Musical Notation. — The seven notes in the octave of the diatonic scale are called *c*, *d*, *e*, *f*, *g*, *a*, *b*; and different octaves are assigned different types or marks. Thus, the octave from 132 to 264 is written as above; the one from 264 to 528, c', d', etc.; the next one c'', d'', etc.; the one from 66 to 132, *C*, *D*, etc.; that from 33 to 66, C_1, D_1, etc.

Using the tempered scale and 264 as the keynote, the notes nearest these to which names have been given are called by the corresponding names; thus 264 is called *c*; 296.3, *d*; 332.6, *e*; etc. The note between *c* and *d*, *i.e.* 279.6, is called "*c* sharp" or "*d* flat"; that between *d* and *e*, *d* sharp or *e* flat; that between *f* and *g*, *f* sharp or *g* flat; that between *g* and *a*, *g* sharp or *a* flat; and that between *a* and *b*, *a* sharp or *b* flat. *a* sharp is written *a*♯; *a* flat, *a*♭; etc. Strictly speaking, if a note is "sharpened," its pitch is increased in the ratio 25 : 24; thus, if the note has a pitch 264, its "sharp" is 275. Similarly, if it is flattened, its pitch is decreased in the ratio 24 : 25; thus the "flat" of 264 is 253.44. These in some cases differ appreciably from the notes which have received the names in the tempered scale.

Another system of notation is to call the notes making up an octave on the diatonic scale Ut, Re, Mi, Fa, Sol, La, Si, starting from the keynote, or a note which is a number of octaves above or below it; and to distinguish the different octaves by subscripts. Thus, adopting 264 as the key tone, the octave from 33 to 66 is written Ut_1, Re_1, etc.; that from 66 to 132, Ut_2, Re_2, etc.; etc. (In place of Ut, Do is now more often used.)

BOOKS OF REFERENCE

POYNTING AND THOMSON. Sound. London. 1899.

This is the best modern text-book and is a storehouse of facts.

TYNDALL. On Sound. New York. 1882.

This is a most interesting description of numerous experiments dealing with all the phenomena of Sound.

HELMHOLTZ. Sensations of Tone. (Translation by Ellis.) London. 1885.

This contains a description of all the experiments on which Helmholtz established his theory of harmony, and also a complete explanation of musical instruments and scales.

RAYLEIGH. Theory of Sound. 2 vols. London. 1896.

This is a mathematical treatise; but it contains excellent summaries of our experimental knowledge of vibrations, waves, and musical notes.

LIGHT

CHAPTER XXV

GENERAL PHENOMENA OF LIGHT

To any one who has the sense of sight, the word "light" conveys a definite meaning, which, however, cannot be put into words; but to a person who was born blind the word is unintelligible. The attention of all who have this sense of sight is attracted to many phenomena in nature, such as the colors of objects, the action of mirrors, the refraction produced by water, etc.; and the study of these forms that branch of Physics called "Light." It will be seen, as we go on, that we can subdivide this subject in certain definite ways.

Fundamental Facts

Light is Due to Ether Waves. — The statement has been made before several times that light is a sensation due to waves in a medium called the ether; but a brief summary of the facts on which this is based may be given again. Thomas Young showed as early as 1802 that interference phenomena, such as described in Chapter XXI, could be observed with light; Fresnel soon after performed numerous diffraction experiments; and Wiener and others have obtained evidences of "stationary waves." The experiments of Young and Fresnel may be repeated easily by any student. Thus it is established that light is a wave phenomenon. Again, Fresnel showed conclusively that these waves are *transverse*, because they admit of *polarization* (see page 313). The existence of a medium is proved by the fact of the wave motion; and, since we can see objects through spaces which are void of ordinary matter and through glass, water, etc., it is shown that this medium is one which fills all space known to us, even inside ordinary material bodies (see page 19). We often use the expressions "light passes" or travels, etc., meaning that the ether waves which produce light in our eyes pass or travel; similarly, we speak of "red light," etc., meaning those ether waves which produce in our eyes the sensation that we call "red." We also speak of "waves in air" or in water, etc., meaning that the waves are in the ether, but that this medium is modified by the presence of air or water.

"Velocity of Light." — The fact that light travels with a measurable velocity was first shown by Roemer, a Danish astronomer, in 1676, from observations on the satellites of Jupiter; and in later years experimental methods have been devised to measure this quantity accurately. These will be discussed in a later chapter. The value of this velocity in the pure ether is not far from 3×10^{10} cm. per second, or about 186,600 mi. per second. It was shown by Foucault by direct experiment that the "velocity in water" is less than in air, proving that the presence of minute particles of water influence the ether more than does that of similar "particles of air."

Sources of Light. — As "sources of light" we make use in general of the sun, or of the sky; of electric discharges through rarefied gases; of solid bodies raised to a high temperature, such as the carbon rods in an electric "arc light," or the filament inside an ordinary electric "glow lamp," or gas-, lamp-, and candle-flames, for, in fact, the luminosity of any flame is due to the presence in it of minute *solid* particles which are raised to a high temperature by the combustion of the gas, etc. These are all, in general, sources of what we call "white" light. We may, however, produce light of different colors, — red, green, yellow, etc., — by surrounding the ordinary source by a colored screen, such as a piece of colored glass; or we can in certain cases put salts in a flame and so color it — thus we can secure a brilliant yellow light by putting common salt (NaCl) in a Bunsen flame, thus making what is called a "sodium flame."

Types of Waves. — If the source of light is very small, we have approximately spherical waves; while, if they come from a distant source like a star, they are plane. We can have two kinds of spherical waves, those which are expanding away from a source, and those which are contracting toward a point. It is a familiar fact that an ordinary reading lens, or magnifier, may produce an image of the sun upon some opaque screen — thus acting as a "burning glass"; this means that the *plane* waves reaching the lens from any point of the sun are changed on passing through the lens into *spherical* waves which converge to a point on the screen. This process of converging waves is exactly the reverse of that of waves diverging from a point source. Thus, if a spherical wave front is concave when considered from portions in the medium toward which it is advancing, it will converge to a point; if it is convex toward that side, it will diverge more and more.

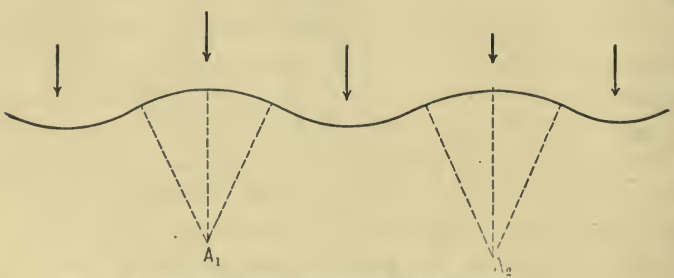


FIG. 185. — Diagram illustrating stellar scintillation.

This fact is illustrated in the familiar phenomenon of stellar scintillation. The waves coming from a star are naturally plane; but if the atmosphere is disturbed by ascending currents of hot air, the wave front is no longer plane, owing to the fact that the velocity of light in cold air is different from that in hot. Thus the wave front at any instant may have a "corrugated" form, as indicated in the cut.

Therefore the light will be concentrated in certain points A_1, A_2, etc., the centres of the concave portions of the wave front; and, as the heated portions of the air change their positions, these points move; so if the eye is at a point A_1 at one instant, the next it may be between two points, A_1 and A_2; etc. So the intensity of the light will increase and decrease intermittently. The same phenomenon is observed in the "shadow bands" seen at times of total eclipses of the sun.

Homogeneous Light and White Light.—In order to determine the wave lengths of these "light waves," it is simply necessary to use the interference method described on page 375. It is found that, if the source of light is white, the interference fringes or bands are all colored, with the exception of the central one, which is white. If, however, the source is colored, the bands are alternately black and colored; that is, in certain lines disturbances in the ether are entirely absent. In general, it will be observed that there are apparently two or more sets of fringes superimposed, each set having a definite color. It is possible, however, to secure such a colored source, that the bands are all of one color, separated by the dark fringes. (This is approximately the case with a sodium flame.) Under these conditions the light is said to be "pure," or "homogeneous." When the source is white, or when any ordinary colored source is used, we can analyze the complex interference pattern into series of fringes, each series having its own color and its own spacing.

It was shown on page 378 that, if the waves have a definite wave length, the fringes are evenly spaced at a distance apart which is proportional to the wave length. Therefore, when a source is homogeneous, it is emitting waves of a single definite wave length; and, in general, an ordinary source of light is emitting trains of waves of different wave lengths. Thus, the interference apparatus "disperses" the complex waves from the source into simpler trains, each having a definite wave length.

Connection between Wave Length and Color. — We can measure the wave length of the waves emitted by any homogeneous source, by using the formula deduced on page 378, viz., distance apart of fringes $= \frac{al}{b}$, where l is the wave length, a is the distance from the two sources to the screen, and b is the distance apart of the sources. In this way it is found that pure red light has a wave length greater than that of pure green; and this in turn is greater than that of pure blue. The longest waves that affect our sense of sight produce the sensation of red and have a length of about 0.000077 cm. (*i.e.* 770 $\mu\mu$); while the shortest produce the sensation of violet and have a length of about 0.000039 cm. (*i.e.* 390 $\mu\mu$). In between these limits there are all possible wave lengths, corresponding to which are all shades of color, ranging from the deepest red, through orange, yellow, green, blue, to the darkest violet. Waves shorter than 390 $\mu\mu$ can be observed by photography; and they have been obtained by Schumann as short as 100 $\mu\mu$. Waves longer than 770 $\mu\mu$ may be studied, as described on page 292, by various means; and they have been observed as long as 25,000 $\mu\mu$, *i.e.* 0.025 cm. (Electrical oscillations produce waves in the ether, which are as a rule very long; but, using minute conductors, waves as short as 0.6 cm. have been obtained. There is thus a gap between 0.6 cm. and 0.025 cm. which has not yet been investigated.)

The fundamental facts are, then, that light is due to transverse waves in the ether; that waves of different wave length produce different colors; that white light is, in general, due to a mixture of waves of all wave lengths which affect our sense of sight; that the velocity of these waves is less in ether inclosed in matter than in the pure ether. (It will be shown on page 433 that, in the former case, waves of different wave length have different velocities, which is not the case in the pure ether.)

General Properties of Light as Due to Wave Motion. — We described in Chapters XX and XXI certain general properties of wave motion which apply directly to light; viz., reflection, refraction, rectilinear propagation, interference, and diffraction. Each of these will be discussed more in detail later; but one or two points should be referred to here. (It should be remembered specially that we are now considering extremely short waves, viz., those whose wave length is not far from 500 $\mu\mu$.)

Rays, Shadows. — It was shown that, if the waves are short, the disturbance at any point in an isotropic medium due to a train of waves depends directly upon the disturbances at previous instants along a straight line drawn backwards perpendicular to the wave front; this line is called a "ray." If the rays are all parallel, we have plane waves, or a "beam" of light. If the waves are spherical, the rays are radii; and, by considering only those rays which are close together, we have a "cone" or a "pencil" of light. (Although this theorem in regard to rays was proved only for a *train of waves*, it may be shown to hold true for a "pulse" also.)

If the waves meet an opaque obstacle, that is, one which does not transmit light, a shadow is cast. If the source of light is a point, the shadow is that which we have called the "geometrical" shadow, if we neglect diffraction phenomena, as we shall for the time being. If, however, the source is large, like a flame or an illuminated piece of paper, the shadow phenomena are evidently not quite so simple.

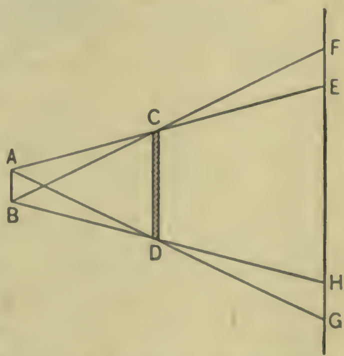


Fig. 186. — Diagram representing the shadow cast on a screen $\overline{FG}$ by an opaque object $\overline{CD}$, when the source $\overline{AB}$ is large.

Thus, if the large source of light is represented by *AB*, and the opaque obstacle by *CD*, the shadow cast by the waves from *A* on a screen *FG* is limited by *EG*, and that cast by the waves from *B* by *FH*. Therefore the region *EH* on the screen receives no light; and the region outside *FG* receives light from all points of the source; but the intermediate regions, *EF* and *GH*, receive light from only portions of the source. Therefore, the intensity of light on the screen fades away gradually toward the central region *EH*, where there is no light. The space back of the screen into which the waves do not penetrate is called the "umbra," while the partly illuminated space surrounding it is called the "penumbra."

An illustration is afforded by solar eclipses, where the sun is the source, the moon is the opaque obstacle, and the earth the screen. In the diagram, *S* represents the sun, and *M* the moon. The umbra and penumbra are indicated by dark spaces. If the earth enters this region, the eclipse is *total* for all points on its surface which are inside the umbra, and *partial* for points outside this but lying in the penumbra. If the earth just misses the umbra, it may happen that at certain points of its surface a ring of sunlight may be seen around the edge of the moon; this condition is called an "annular" eclipse.

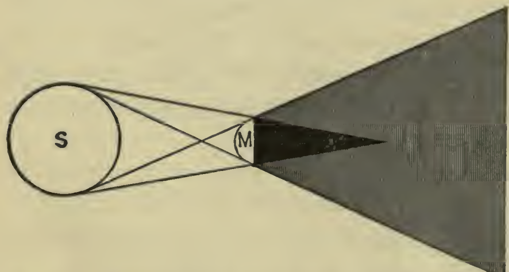


Fig. 187. — Diagram showing shadow cast in space by the moon, *M*, owing to the rays from the sun, *S*.

Another interesting case of rectilinear propagation is given by the formation of what are called "pin-hole" images. If a small hole is made in an opaque screen, any luminous object — *e.g.* a building in sunshine — situated on one side of it will produce on a screen on the other side an inverted image of itself, which is comparatively sharply defined. Thus if there is a small opening at *O* in the screen, and *A* is a point of an illuminated figure, there will be a cone of light from *A* passing through the opening. If this meets a screen, it will make at the point *B* a bright spot, which will have the shape of the opening. If the opening is small and the two screens are close together, the spot at B will be practically a point of light; and hence, as each point of the illuminated figure produces a point of light on the screen, there will be formed a well-defined image. As the images of the various points of the illuminated object overlap, the general appearance of the image is almost independent of the shape of the opening, if it is small. (The round or elliptical spots of light which are seen on the floors near curtained windows or under trees are images of the sun formed by minute openings in the curtains or leaves.) This is the principle of pin-hole photographic cameras, of the camera obscura, etc.

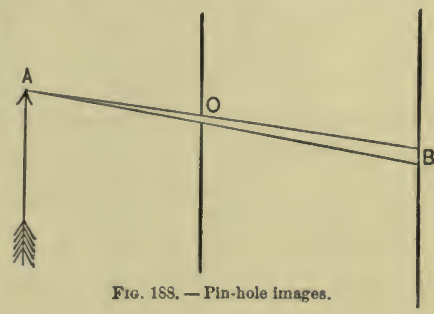


FIG. 188. — Pin-hole images.

Opacity; Transparency; Translucency. — A distinction is made between material media which are "transparent" and those which are "opaque." If an object when introduced between the eye and a source of light stops all the light, *e.g.* a board, a piece of tin, etc., we say that it is opaque, meaning that it does not transmit those ether waves which affect our sense of sight. An object may be opaque to some waves and not to others; thus, a piece of red glass is opaque to all visible waves except those which produce the sensation of red. Again, a given material body may be opaque to waves which do not affect our sense of sight, and may transmit all visible waves; and there may be media with just the reverse property. Opacity is due to the fact that the waves which are incident upon the body do not pass through it; they are either reflected back from it or are absorbed by it. Thus, a polished metallic surface is opaque largely owing to reflection, while a blackened surface is opaque owing to absorption.

If we can look through an object and see sources of light on the other side sharply defined, we say that it is "transparent." Here, again, the body may be transparent to some waves and not to others. Waves, then, when incident upon a transparent body, will under ordinary conditions be transmitted by the ether in the body, *maintaining a definite wave front*. This is true only if the transparent body has surfaces which are "smooth," in the sense that there are no inequalities comparable in size with the length of the waves. Thus, a window pane of glass, layers of water or alcohol, etc., are transparent. If the surface is rough, the waves suffer irregular reflection. Again, we will see later that in certain cases the waves incident upon the surface of a smooth transparent body are entirely reflected. (This is called *total reflection.*)

It must not be thought, however, that a body which is not transparent is necessarily opaque; a piece of opal glass or of oiled paper is not transparent, nor is it opaque. One cannot see objects through them, and yet they allow light to pass to a certain extent. Such bodies are said to be "translucent." What happens is this. When waves from any source fall upon a translucent body, they are broken up and scattered by it in such a manner that each point of the body becomes a new and independent source of waves. So when such a body is held between the eye and a source of light, the waves which reach the eye come directly from the points of the body, not from the source; and what the eye sees, then, is the surface of the body. A transparent body cannot be seen; it is only owing to dirt on it that we are able to see a surface of water or a window pane. The explanation of translucency will be given in the following pages.

Reflection and Refraction. — It has been shown that, if there are two media separated by a bounding surface, waves in one will be reflected at the boundary in general, if the velocity of the waves in the two media is different. The boundary surface must be large in comparison with the length of the waves, otherwise the waves will pass around the "obstacle."

If the surface is large, but possesses inequalities which are not small in comparison with the wave length, each minute portion of the surface acts like a separate reflector, and so the waves are scattered. If, however, the surface is "smooth," in the sense that any inequalities are extremely small, the reflection is regular, and the surface is called a "mirror." This may have any shape, but in practice the only forms used are plane, spherical, cylindrical, and parabolic surfaces. (These curved surfaces may, of course, be either concave or convex.) Similarly, in these cases there will be waves entering the second medium with a definite wave front. As already stated, these are called "refracted" waves. If the medium is transparent, they continue indefinitely, but if it is absorbing, they soon die down.

Diffuse Reflection. — When waves fall upon a roughened surface they are broken up and reflected irregularly, as just explained; they are said to be "diffused." Each point of the surface now becomes a new source of waves. This diffusion of light is taking place from all natural objects, all walls, pieces of furniture, etc.; and it is owing to it that we are able to see any object. If a body reflects regularly, we do not see it, but the source of light reflected in it.

Similarly, if a transparent body contains immersed in it a great number of minute foreign particles, it diffuses the light falling upon it. In one class of such bodies the minute foreign particles are opaque; whereas in another class they are transparent, but have irregular figures. In both cases the incident waves are broken up and suffer reflection to and fro from particle to particle; and finally the disturbances emerge in the form of spherical waves proceeding out from each point of the surface. (If the body is thin, these waves may emerge on both of its sides.)

Diffusive reflection is thus due to one of three causes: etching or roughness of the surface, the presence of opaque particles in a transparent medium, or that of minute irregular transparent particles in such a medium. Ground glass, ordinary unglazed paper, etc., are illustrations of the first class of bodies; opal glass, celluloid films, etc., of the second; and foam, thin starch water dried on glass, etc., of the third. All these bodies which reflect light diffusively, diffuse it also when they transmit it; that is, they are translucent.

If these foreign particles in a transparent medium are very minute, it may happen that the shorter waves only are affected, while the longer ones will be transmitted through the body. In this case the body would appear bluish when viewed in reflected light. This is the explanation of the blue color of the sky, of fine smoke, of a hazy atmosphere, of blue eyes, etc. In each of these cases there are minute particles existing as foreign bodies in a transparent medium, which diffuse the short waves, but allow the longer ones to pass.

Regular Reflection and Refraction. — We gave in Chapter XX the treatment of reflection of plane waves at a plane surface, quoting from Huygens. But, as noted at the time, this method is not rigorous. In order to make it so, we must follow the same plan as did Fresnel in discussing rectilinear propagation; namely, we must divide the wave front up into zones and deduce the effect of the secondary waves, thus combining Huygens's principle of secondary waves and Young's principle of interference. When this is done, we obtain Huygens's solution, and also learn the effect of reflection upon a *ray*. Thus, by Huygens's solution of the problem, plane waves, whose wave front at any instant is given in section by $\overline{AB}$, incident upon a plane surface MM, are reflected, and form the

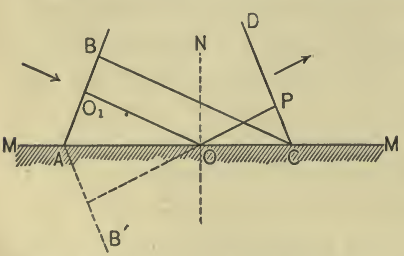


FIG. 189. — Reflection of plane waves by a plane mirror.

plane waves whose wave front at a later time is given by $\overline{DC}$, where the angles (BAC) and (DCA) are equal. If we draw a straight line $\overline{O_1O}$ perpendicular to $\overline{AB}$ and meeting the reflecting surface at O, and another $\overline{OP}$, perpendicular to the reflected wave front $\overline{CD}$, it is seen by geometry that the broken line $\overline{O_1OP}$ is the *shortest* line that can be drawn from P to the wave front $\overline{AB}$ by way of the surface, and so O_1 is the "pole" of P (see page 382); and, in drawing the zones around O_1, which are to be compounded in order to deduce the effect at P, it is seen that they are exactly the same as if the reflecting surface were removed and the wave front had the position $\overline{AB'}$ where the angles $(B'AC)$ and (BAC) are equal. From this fact the ordinary laws of reflection follow at once. The disturbance at P is due directly to that at O, and this in turn to that at O_1; consequently, the *ray* $\overline{O_1O}$ is turned into the ray $\overline{OP}$ by reflection. If a normal $\overline{ON}$ is drawn to the surface at O, the angles (O_1ON) and (PON) are equal; the former is called the "angle of incidence," the latter that of "reflection." The plane including the lines $\overline{O_1O}$ and $\overline{ON}$ is called the "plane of incidence"; it evidently includes also the reflected ray $\overline{OP}$.

Reflection of light may be demonstrated with ease by causing a beam of sunlight, or a beam from a lantern, to fall upon an ordinary looking-glass; for the path of the beams may be seen if dust or smoke is distributed through the air.

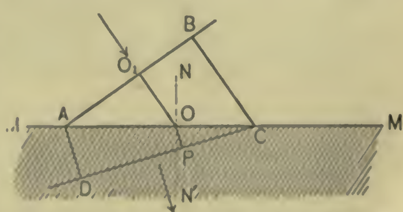


Fig. 190. — Refraction of plane waves by a plane surface.

Similarly, in the case of refraction of plane waves at a plane surface, Huygens's solution is that, if $\overline{MM}$ is the trace by the paper of the refracting surface, and $\overline{AB}$ that of the incident plane waves, the refracted waves are plane and their trace is $\overline{CD}$, where the angles (BAC) and (DCA) are connected by the relation

$$\frac{\sin (BAC)}{\sin (DCA)} = \frac{\text{velocity of waves in the upper medium}}{\text{velocity of waves in the lower medium}}.$$

This ratio is a constant for the two media and for waves of a definite wave number, and is, therefore, the same for all angles of incidence; it is called the "index of refraction" of the lower medium, with reference to the upper for this wave number. This proof of Huygens's, however, is not rigorous; but if Fresnel's method of treatment is followed, it may be made so.

If we draw a line $\overline{O_1O}$ perpendicular to the wave front of the incident waves at O_1, and continue it by the line $\overline{OP}$ drawn perpendicular to the wave front of the refracted waves, it may be shown by geometry that the *time taken* for a disturbance to pass from O_1 to P along this line is less than along any other line. (Huygens proved this.) Therefore O_1 is the pole of P on the incident wave front; and by constructing Fresnel's zones around O_1, we are led at once to Huygens's solution. The disturbance at P is due directly to that at O; and this, in turn, to that at O_1; so the incident ray $\overline{O_1O}$ has its direction changed at the surface into $\overline{OP}$. Drawing a normal $\overline{NON'}$ to the refracting surface at O, the angle (O_1ON) is the angle of incidence; and the angle (PON'), the angle of refraction. These angles are equal respectively to (BAC) and (DCA). Therefore we can state the law for a ray:

$$\frac{\text{the sine of the angle of incidence}}{\text{the sine of the angle of refraction}} = \text{the index of refraction.}$$

Calling the angle of incidence i, that of refraction r, and the index of refraction n, this relation is

$$\frac{\sin i}{\sin r} = n.$$

The other law of refraction is evident, viz., the refracted ray is in the plane of incidence.

Refraction may be studied with ease by allowing a beam of sunlight to fall upon the surface of cloudy water in a tank. It will be observed that, if sources of light of different color are used, the refraction is different; if the waves have the same angle of incidence, the angles of refraction are different. Thus the index of refraction varies with the wave number or color of the light. This is not easily shown in the experiment just described, unless the beam of light is made extremely narrow; because the differences in the refraction are not great, but by means of a "prism" or a lens this phenomenon is most apparent. If the incident light is white, each of the component trains of waves (see page 423) has its own index of refraction, and so the light is broken up or dispersed, forming a "spectrum." In the case of ordinary transparent media such as glass, water, etc., the waves having the shorter wave lengths are refracted more than those having longer ones; *i.e.* blue light is refracted more than green, green more than red. This proves that in these media short waves have a less velocity than do long ones. (In the pure ether all waves, so far as we know, have the same velocity.) This kind of medium is said to have ordinary or "normal" dispersion. In other media, it may happen that some waves are refracted more than others which have a shorter wave length; they are said to have "anomalous dispersion" (see Chapter XXX).

When waves are passing in the ether inclosed in any material medium, such as water, the minute particles of this medium are in motion also to a greater or less extent, owing to the waves; so, if there is a long train of waves, and if we consider any one point in space, the effect produced there by the reaction of the matter on each "wave" as it passes it is different from what it would be if the matter were at rest; as, for instance, if a sudden "pulse" came up to this point and passed. Since the velocity of a disturbance through the ether depends upon this reaction of the matter which incloses it, it is evident that the velocity of a train of waves is different from a pulse. Further, the method of Fresnel for considering wave motion presupposes the existence of a train of waves. Thus the laws of refraction apply only to trains of waves.

Geometrical Optics. — Other cases of reflection and refraction will be considered in the following chapters: spherical waves upon a plane surface, plane waves upon a spherical surface, and spherical waves upon a spherical surface. There are two modes of procedure possible: one is to study the changes in the *wave front* produced by reflection or refraction; the other is to study the changes in the direction of the *rays*, making use of the theorems just deduced; for in the case of incidence upon a curved surface, we may consider the reflection or refraction of a ray at any point as due to an infinitesimal portion of the tangent *plane* of the surface at that point. The application of this latter method makes up what is called the science of "Geometrical Optics." We shall use this in these chapters, but shall also outline in certain cases the demonstrations in terms of waves.

Real and Virtual Foci. — If spherical waves diverging from a point source are spherical also after reflection or refraction, we may have either of two conditions: the centre of the reflected or refracted waves may be in the medium in which the waves are advancing, *i.e.* the waves *converge* to a "focus"; or the centre of the waves may be in the other medium, *i.e.* the waves will *diverge* away from their centre. (Of course, in the former case, the waves, after converging to a point, will diverge again beyond it if no obstacle prevents.) The centre of the converging waves is called a "real" focus; it is said to be a "real image" of the source or "object." The centre of the diverging waves is called a "virtual" focus; it is said to be a "virtual image." There are cases, however, in which, even though the incident waves are spherical, the reflected or refracted waves are not.

Homocentric and Astigmatic Pencils. — Similarly, from the standpoint of rays, if we consider any incident pencil of rays proceeding from a point source, it will, after reflection or refraction, form another pencil with its vertex in the medium into which the rays are advancing, if there is a real focus, or one with its vertex in the other medium, if the focus is virtual. But there are cases when, after reflection or refraction, the rays do not form a cone. A pencil of rays which does form a cone is said to be "homocentric"; while one which does not is said to be "astigmatic." In this latter case, as we shall see, the rays of a homocentric pencil, after reflection or refraction, have as a focus (either real or virtual) not a point, but two short lines perpendicular to each other and a short distance apart; these are called "focal lines."

In describing the incidence of a pencil of rays it is simplest to give the direction of its central ray; so by speaking of "normal incidence" of a pencil we mean a case when the central ray of the small pencil is perpendicular to the surface at the point where the pencil meets it; and by "oblique incidence" is meant a case when the central ray of the pencil makes an angle different from zero with the normal to the surface at the point where this ray meets it. We shall see shortly that in all cases a pencil which is normal to a surface produces by reflection or refraction a homocentric pencil; and in nearly all cases an oblique pencil produces an astigmatic one.

Properties of a Focus from the Standpoint of Waves. — Since a wave front is the locus of the points reached by the disturbances at any one instant, we may consider the existence of foci from a different standpoint. If waves diverging from a point source converge after reflection or refraction to another point, we can draw various rays proceeding out from the former point and all meeting again at the latter. These rays have different paths; but the *time* taken for the disturbances to pass along all of them must be the same. Thus, if l_1 is the length of the portion of a ray in one medium in which the velocity of the waves is v_1, and if l_2 is the length of its portion in a second medium in which the velocity is v_2, the time taken for the propagation of the disturbance is $\frac{l_1}{v_1} + \frac{l_2}{v_2}$, or $\left(l_1 + \frac{v_1}{v_2} l_2\right)\frac{1}{v_1}$. But $\frac{v_1}{v_2}$ is the index of refraction, n, of the second medium with reference to the first; so this time is $(l_1 + nl_2)\frac{1}{v_1}$. The quantity $(l_1 + nl_2)$ is called the "optical length" of the ray. It is evidently equal to the distance in the first medium which the waves would advance in the time taken for the actual propagation in the two media. Then we may state that the optical lengths of all rays from the point source to the focus are the same. We shall make use of this principle in discussing lenses.

Another fact in regard to foci should be emphasized. Reflection and refraction are always produced by pieces of matter of limited size, *e.g.* looking-glasses, prisms, lenses, etc.; and so only a portion of the wave front undergoes the change. The effect at any point in the advance of the wave front must then be deduced by following Fresnel's method of combining the principles of Huygens and Young. The point at which the disturbance is greatest is the focus; but this does not mean that there is no disturbance at other points. The effect at these latter points must in each case be calculated; for some it is zero, and in no case does it approach in amount that at the focus. This fact is of great importance when we discuss the reflection or refraction of trains of waves from two points that are close together.

CHAPTER XXVI

PHOTOMETRY

Intensity of Sources of Light; Intensity of Illumination. — One of the most important practical questions in regard to sources of light deals with their illuminating power. Of course only part of the energy radiated by these sources is in the form of waves of such lengths as to affect our sense of sight; and we speak of the "luminous energy" or the "quantity of light" that is radiated. The quantity of light emitted by a source in a unit of time measures its "intensity." If the source is small compared with the distance at which the light is received, the waves may be regarded as spherical. In other cases, for instance a white wall or screen, this is far from being true. In any case, however, if by means of opaque screens with suitable openings in them the radiation is limited to a "beam," we can find the relation between the quantities of light falling on surfaces inclined at different angles to the direction of the beam. Thus, as shown in the cut, let the beam be that passing through two equal and parallel openings $\overline{AB}$ and $\overline{A_1B_1}$, of area A; and let the beam fall upon a screen whose normal makes the angle N with the direction of the beam. The illuminated surface on this screen, $\overline{CD}$, will have an area B, where

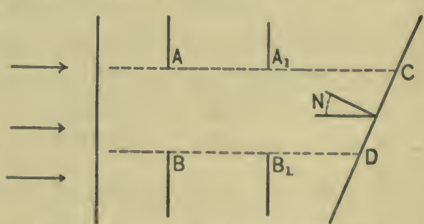


Fig. 191. — Incidence of a beam of light upon an oblique screen.

$B \cos N = A$. (See page 28.) If the screen is a diffusing one, this portion of it will appear bright when viewed from any direction. The "intensity of its illumination" is defined to be the quantity of light received per unit area in a unit time. Therefore, if Q is the quantity of light carried by the beam in a unit of time, the intensity of illumination, I, of a screen perpendicular to the beam is $\frac{Q}{A}$. That is, $I = \frac{Q}{A}$. The intensity of illumination of the oblique screen is $\frac{Q}{B}$. Calling this I_1, we have, then, $I_1 = \frac{Q}{B} = \frac{AI}{B} = I \cos N$.

If the source is small, we can, as has been said, consider the waves as spherical; and, if Q is the quantity of light emitted by the source in a unit of time, assuming that it radiates uniformly in all directions, the amount falling on a portion of the spherical surface of area A at a distance r is $\frac{QA}{4\pi r^2}$. If the illuminated surface is small and is oblique to the direction of propagation, let its area be B and let the angle made between a perpendicular to it and the direction of propagation be N; then the projection of this surface on a plane perpendicular to the direction of propagation has the area $B \cos N$. So in the above formula if A is this projected area, $A = B \cos N$; and the light received in a unit time by the oblique surface whose area is B is $\frac{Q}{4\pi} \frac{B \cos N}{r^2}$. The intensity of illumination of this oblique surface is then $\frac{Q}{4\pi} \frac{\cos N}{r^2}$.

This formula offers at once a method for the comparison of the intensities of two sources of light. The general method is to illuminate a portion of a diffusing screen by one source and contiguous portions of the screen by the other, and then to vary the distances of the sources until the two portions of the screen appear equally bright. When this is the case, the intensity of illumination must be the same for both portions; and, if the angles of inclination of the two sources are the same, $\frac{Q_1}{r_1^2} = \frac{Q_2}{r_2^2}$.

Luminosity of Sources of Light. — If the source of light is not a point but an extended surface, like the surface of a white-hot metal or of a diffusing screen, let us consider the radiation of any small portion of the surface, whose area we may call A. If this were a point source, and it emitted a quantity of light Q in a unit of time, the quantity received in that time on a screen of unit area at right angles to the direction of propagation at a distance r, would be, as we have seen, $\frac{Q}{4\pi r^2}$. Therefore the total amount of light actually received per unit area by a screen, *parallel* to the luminous surface and at a distance r from it, varies inversely as r^2, provided the area of the luminous surface and that of the portion of the illuminated screen considered are both small and face each other. That is, calling P the quantity of light received in a unit of time per unit area of the screen, $P = \frac{c}{r^2}$, where c is a constant depending upon the properties of the luminous surface. Evidently c varies directly as the area of this surface; so we may write $c = LA$, or $P = \frac{LA}{r^2}$, where L is a factor of proportionality. The quantity L is called the "intrinsic luminosity" of the luminous surface "for perpendicular emission." It is a constant for the given luminous surface.

The light received from a luminous surface by a parallel surface of unit area at a distance r is, as just shown, $\frac{AL}{r^2}$. The brightness of a luminous object as judged by our eyes depends upon the light received per unit area of the retina. Thus, if the light received by the eye, passing through the diaphragm formed by the pupil, is Q, and if the area of the image of the luminous object formed on the retina is a, the brightness of the object is proportional to $\frac{Q}{a}$. If the object is a surface parallel to the eye and at a distance r, and if B is the area of the pupil, this light Q entering in a unit of time is, from the above definition of L, $\frac{ALB}{r^2}$. The image formed has the area a; and it will be shown later in speaking of lenses that a is proportional inversely to r^2; or $a = \frac{k}{r^2}$, where k is a constant depending upon the construction of the eye. Hence the brightness of the luminous surface, $\frac{Q}{a}$ is $\frac{ALB}{r^2}$ divided by $\frac{k}{r^2}$, or $\frac{ALB}{k}$. For a given luminous surface and a given eye, this is a constant quantity; or, in words, a luminous surface appears equally bright at all distances from the eye.

Certain luminous objects, for instance a star, do not produce *images* on the retina, as do ordinary luminous bodies; they give rise to a diffraction pattern determined by the size of the pupil. (See page 390.) Therefore the above statement does not hold for them. A telescope will increase the brightness of a star, because it introduces more of its light into the eye, but will not increase that of the sun or the moon.

Experience shows that a luminous surface when viewed obliquely appears practically as bright as when viewed perpendicularly. Thus, a luminous spherical solid appears to our eyes like a luminous disk of uniform brightness. If the area of the oblique surface is A, and if it is inclined at such an angle that a perpendicular to it makes the angle N with a line drawn to the eye, $A \cos N$ is the area of the projection of this surface perpendicular to this line; and a surface of this area placed parallel to the eye appears of the same brightness as the one of area A placed at the angle N. Let L' be the intrinsic luminosity corresponding to the direction N; *i.e.* if the area of the surface is unity, the light received per unit area at a distance r in this oblique direction is $\frac{L'}{r^2}$. Then the light received by the eye from the oblique surface is $\frac{AL'B}{r^2}$; whereas, if the parallel surface of area $A \cos N$ were used, the light received would be $\frac{A \cos N \cdot LB}{r^2}$. But experience proves that these are practically equal; so $L' = L \cos N$. This is called "Lambert's Law."

The intrinsic luminosity in any direction of a small luminous surface is, in words, the quantity of light received per unit area by a screen perpendicular to this direction at a unit distance, divided by the area of the luminous surface. So, if the screen is placed obliquely to this direction, making an angle N_1 with it; and if its area is B, the light received on it in a unit of time from a luminous source of area A, making the angle N with this line referred to, and at a distance r, is

$$\frac{A \cos N \cdot L \cdot B \cos N_1}{r^2} \text{ or } \frac{LAB \cos N \cos N_1}{r^2}$$

As noted above, this statement is not absolutely true. We may regard L as a quantity which is not a constant factor but varies slightly with N.

A method is thus evident for the comparison of the luminosities of different sources of light. Each is surrounded by an opaque screen provided with a rectangular opening, or slit; the sources are so situated that a suitably placed diffusing screen receives light *perpendicularly* from these two openings; one portion of the screen receiving light from one source only, and contiguous portions receiving light from the other source only. The screen is now illuminated by light coming from the two rectangular openings as sources. Then by some means the conditions are so altered that the brightness of one portion of the screen is diminished until that of the two portions appears equal. This diminution in intensity may be secured in various ways: (1) remove one source to a greater distance, (2) alter the widths of the rectangular slits, (3) interpose between the more intense source and the illuminated screen an opaque disk which has certain sectors cut out, and cause this to revolve rapidly; by altering the size of the sectors the intensity may be diminished at will and in a known ratio. When the two portions of the screen are equally illuminated, they are receiving equal quantities of light per unit area. For simplicity let us assume that there is no interposing revolving disk. Let

A_1 be the area of one opening, L_1 its luminosity, and r_1 its distance from the illuminated screen; the light received by it per unit area per unit time is then $\frac{A_1L_1}{r_1^2}$. Similarly, giving A_2, L_2, and r_2 corresponding meanings for the second source, the light received by the screen per unit area per unit time from the second source is $\frac{A_2L_2}{r_2^2}$. Since, in the arrangement described above, these are equal, $\frac{A_1L_1}{r_1^2} = \frac{A_2L_2}{r_2^2}$; and so L_1 and L_2 may be compared.

Standards of Light; Photometers.—Various standard sources of light have been defined. One of these is the flame of a "standard candle," which is a sperm candle weighing one sixth of a pound and burning 120 grains per hour. Another is the flame of a "Hefner-Alteneck lamp" (which burns amyl acetate), when the height of the flame is kept at 4 cm. Still another is a surface of platinum when it is raised to such a temperature that it is on the point of melting.

An instrument for comparing the intensities or the luminosities of two sources of light is called a "photometer"; and the special science involved in the study is called "photometry." In Rumford's photometer, an opaque rod is placed close to the diffusing screen, so that two shadows are cast on it by the two sources; the shadow cast by one receives light from the other only; so, when the brightness of the two shadows is the same, the wished-for condition is obtained (care must be taken to have the angles of inclination the same). In Bunsen's photometer, a screen consisting of white unglazed paper, in the centre of which there is a small round or star-shaped grease spot, is placed between the two sources. Looking at this screen from either side, any portion is illuminated by the *transmitted* light from the source on the other side and also by the *reflected* light from the source on that side. The screen is moved until the grease spot and the other portions appear equally bright when viewed from either side; and then the above relation holds. For, let a be the proportion of light *reflected* by the unglazed paper, and b that reflected by the greased paper, and assume that there is no absorption. Then, if P_1 is the quantity of light per unit area incident upon one side of the screen, and P_2 that upon the other, the amount of light *reflected* per unit area by the unglazed portion on the former side is aP_1, and that received by transmission from the other side is $(1-a)P_2$; similarly, the light reflected by the grease spot is bP_1, and that received by transmission is $(1-b)P_2$. Hence, when the two portions are equally bright,

$$aP_1 + (1-a)P_2 = bP_1 + (1-b)P_2,$$

or

$$(a-b)P_1 = (a-b)P_2.$$

And therefore, since a does not equal b, $P_1 = P_2$. The best photometer in use to-day is one designed by Lummer and Brodhun. For full details of these and other instruments reference should be made to some treatise on Photometry, such as Stine, *Photometrical Measurements*, or Palaz, *Industrial Photometry*.

Naturally, the intensities of two sources of different color cannot be compared directly; and, in general, if any two sources are to be compared, their luminosities corresponding to each wave length should be investigated. This can be done by combining with a photometer a dispersing apparatus such as a prism. The complete apparatus is called a "spectrophotometer," the simplest and most accurate form of which is one devised by Professor Brace of the University of Nebraska.

CHAPTER XXVII

REFLECTION

WHAT is meant by regular reflection, and by a mirror, has already been explained; and the law of reflection for a ray has been deduced (see page 431). We will now consider several special illustrations.

Plane Waves Incident upon a Plane Mirror. — This is the case already discussed on page 367, and needs no further treatment here. There is one illustration of it, however, which may be described. It is that of a plane mirror which is being rotated when plane waves are incident upon it. Let the trace of the mirror by the paper at any instant be $\overline{MM}$, and let $\overline{PO}$ be any incident ray; draw $\overline{ON}$ perpendicular to the mirror at O; the reflected ray $\overline{OR}$ will make with the normal an angle (RON) equal to the angle (PON). Let the mirror now be rotated about an axis through O perpendicular to the plane of the paper; that is, about an axis parallel to the intersection of the plane wave front with the plane mirror. At the end of a certain time the mirror will have turned into the position indicated by M_1M_1; so, if $\overline{ON}_1$ is the position of the normal, the reflected ray will be $\overline{OR}_1$ where the angles (PON_1) and (N_1OR_1) are equal. The angle turned through by the reflected ray is

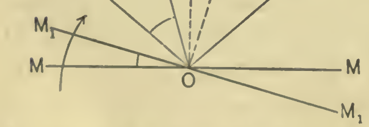


FIG. 192. — Rotating mirror: the incident ray is PO.

(ROR_1). This equals the difference between the angles (POR) and (POR_1), that is, *twice* the difference between the angles (PON) and (PON_1), or twice the angle (NON_1). But this is the angle of rotation of the mirror; so the reflected ray turns twice as fast as the mirror. This principle is made use of in many optical instruments: the sextant, which is used to measure the angle subtended at the eye of the observer by two distant points; the mirror attachment to a galvanometer; etc.

Spherical Waves Incident upon a Plane Mirror. — Let the sheet of paper be perpendicular to the plane of the mirror, and let the trace of the latter be MM_1. If O is the source of the waves, we may consider any two rays $\overline{OP}$ and $\overline{OQ}$. By reflection they become $\overline{PP_1}$ and $\overline{QQ_1}$, where the angles (MPO) and (M_1PP_1), and (MQO) and (M_1QQ_1) are equal. If $\overline{PP_1}$ is prolonged backwards, it will meet in the point O' a line drawn from O perpendicular to the mirror. Let this line meet the mirror in the point A. Since the angles (MPO) and (MPO') are equal, the triangles (OAP) and ($O'AP$) which have the side $\overline{AP}$ in common are equal; and so $\overline{AO} = \overline{O'A}$. The point O' is therefore as far below the surface of the mirror as O is above it, and is independent of the position of P; that is, its position does not depend upon the ray which we use to locate it; and therefore the prolongation backwards of the reflected ray $\overline{QQ_1}$ must also pass through O', as is evident at once from geometry. It follows that all the rays from O which are incident upon the mirror pro-

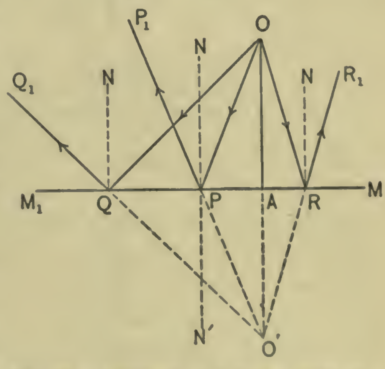


FIG. 193. — Formation of a virtual image by a plane mirror: O is the source, O' is the image.

ceed after reflection as if they came from O'. This, then, is the *virtual image* of O; and all pencils, normal or oblique, have their vertex at O' after reflection. Therefore an observer looking into the mirror will apparently see a source of light at O'. Similarly, any extended luminous object, AB, has a virtual image, $A'B'$, of the same size as itself and symmetrically placed with reference to the plane of the mirror. This explains the ordinary use of a looking glass.

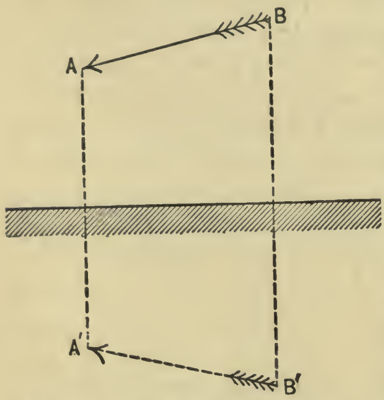


FIG. 194. — Image formed by a plane mirror.

If two plane mirrors are placed close together, several images of a point source are formed by the reflected rays. Thus, if the mirrors are at right angles, and the point source is at O, images are formed at O_1, O_2, and O_3; the first two are the ordinary images of O in the two mirrors; the third is the image in one mirror formed by those rays which fall upon it after reflection at the other. It may be shown in a similar manner that, if the mirrors make an angle N with each other such that, when expressed in degrees, $\frac{360}{N}$ is a whole number n, there are $n-1$ images. (This is the principle of Brewster's kaleidoscope, in which three mirrors are placed so that their section is an equilateral triangle, and broken pieces of colored glass serve as luminous objects. In this case $N = 60°$, so there are five images.)

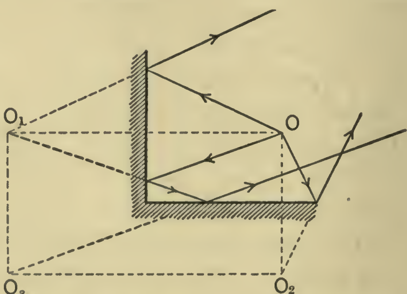


FIG. 195. — Images formed by two plane mirrors which are at right angles.

This case of reflection of spherical waves by a plane mirror may be stated simply in terms of waves. A point source emits waves which advance in the form of ever-expanding spherical surfaces. Let such a spherical wave front proceed out from the point source O; it will meet the surface at A, where $\overline{OA}$ is perpendicular to the plane surface. The waves will then be reflected. If there had been no reflecting surface, the wave front *would* have reached the position PQP' after a definite time; but owing to reflection, the disturbance has the wave front $PQ'P'$, where this spherical surface is simply PQP' inverted. This new wave front will therefore proceed back from the surface exactly as if it came from a point O' below the surface, where $\overline{OO'}$ is a line perpendicular to the surface and bisected by it.

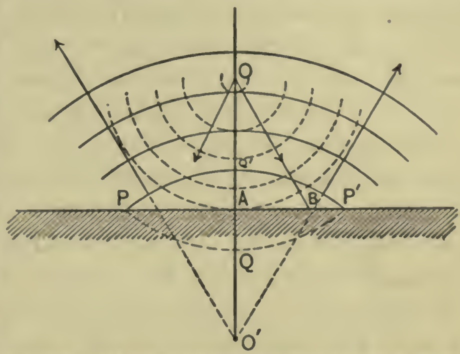


FIG. 196. — Reflection of spherical waves by a plane mirror.

Spherical Waves Incident upon a Spherical Mirror. — There are two cases to be considered, depending upon whether the *concave* or the *convex* surface is turned toward the illuminated object.

1. *Concave Mirrors.* — Let the centre of the spherical surface be C, and the point source be O; and let a section of the surface by a plane through these two points be PRM. The line $\overline{OCM}$, joining the centre of the spherical surface to the point source, is called the "*axis* of the mirror with reference to O." A normal pencil of rays from O has then the axis as its central ray. Let $\overline{OP}$ be any ray of such a pencil;

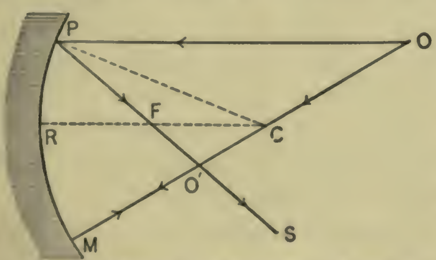


FIG. 197. — Formation of an image of a point source by a concave spherical mirror. O and O' are "conjugate foci."

that is, *the angle* (MOP), *and therefore the distance* $\overline{MP}$, *is supposed to be small.* The radius $\overline{CP}$ is normal to the surface at P; and therefore, if the line $\overline{PS}$ is drawn making the angle (SPC) equal to the angle (OPC), it is the reflected ray caused by the incidence upon the mirror of the ray $\overline{OP}$. The ray $\overline{OM}$ will be reflected directly back, because the line is perpendicular to the mirror. Hence O', the intersection of $\overline{PS}$ and $\overline{OM}$, is the point to which the two rays come. Further, *all* rays from O, after reflection at other points of the surface, have directions which pass through O', provided the surface is only slightly curved, and that only a small portion of the mirror around M is used. For, since the line $\overline{PC}$ bisects the angle $(O'PO)$,

$$\overline{PO} : \overline{PO'} = \overline{CO} : \overline{O'C};$$

or, putting

$$\overline{MO} = u, \ \overline{MO'} = v, \ \overline{MC} = r,$$

$$\overline{PO} : \overline{PO'} = u - r : r - v.$$

If, however, the above conditions as to the curvature of the mirror and the closeness of P to M are satisfied, the distance $\overline{PO}$ nearly equals $\overline{MO}$, and $\overline{PO'}$ nearly equals $\overline{MO'}$. That is,

$$u : v = u - r : r - v.$$

Hence for definite values of u and r, that is, for waves from a definite point source O falling upon a concave mirror with the radius r, the value of v, which determines the position of the image O', is independent of P. It is, in the case illustrated in the cut, a *real* focus of O. Conversely, if O' is a source of rays, they will after reflection converge to O. The two points are therefore called "conjugate foci."

The equation for v may be put in the form

$$\frac{u - r}{u} = \frac{r - v}{v},$$

or

$$\frac{1}{u} + \frac{1}{v} = \frac{2}{r}.$$

A simple geometrical method for determining O' is as follows: Draw $\overline{OCM}$ through the centre of the mirror C, draw $\overline{OP}$ to any point P near M, and $\overline{CR}$ through C parallel to it; draw a line $\overline{PF}$ so as to bisect $\overline{CR}$ at F; where this line intersects the line $\overline{OM}$ is the image O'. For, as has just been shown, O' lies on the line $\overline{OM}$ and on $\overline{PS}$, where the angles (SPC) and (OPC) are equal; and the intersection of $\overline{PS}$ with $\overline{CR}$ may be proved to bisect it. $\overline{CR}$ is drawn parallel to $\overline{OP}$, and F is its point of intersection with $\overline{PS}$. The angles (RFP) and (FPO) are equal, and (RFP) equals the sum of (FPC) and (FCP); hence (FCP) equals the difference between (FPO) and (FPC), *i.e.* (CPO). But (CPO) and (FPC) are equal (angles of incidence and reflection); therefore (FCP) equals (FPC); the triangle (CFP) is isosceles; and the sides $\overline{FC}$ and $\overline{FP}$ are equal. P is supposed to be close to M, and therefore to R; and so $\overline{FP}$ practically equals $\overline{FR}$. Consequently $\overline{FC}$ equals $\overline{FR}$. Q.E.D.

A special case is when u is infinite; that is, when the incident waves are plane, with their wave normal parallel to $\overline{OCM}$; hence $v = \frac{r}{2}$, or O' bisects the line $\overline{CM}$. This point is called the "principal focus" on the line $\overline{OM}$. Conversely, if a point source is at the middle point of $\overline{CM}$, *i.e.* if $u = \frac{r}{2}$, the reflected waves will be plane and will have their wave normal parallel to $\overline{CM}$.

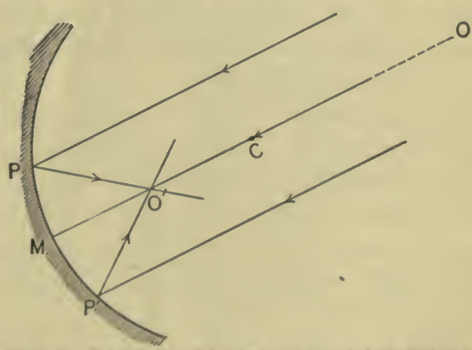


FIG. 198. — Special case: the source O is at an infinite distance

It is seen from the formula $\frac{1}{u} + \frac{1}{v} = \frac{2}{r}$, that, if $u < \frac{r}{2}$, $v < 0$. That is, if the point source O is between the surface and the principal focus, v is negative; and therefore O' is on the opposite side of the surface. This means, then, that it is a *virtual* focus; and so the rays from O, after reflection, *diverge* away from the surface.

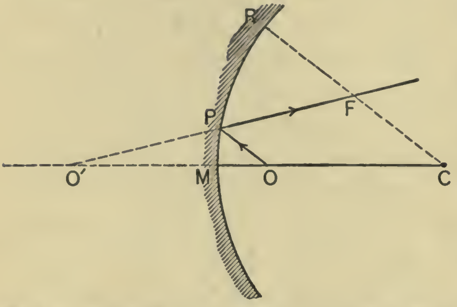


FIG. 199.—Special case: O is nearer to the mirror M than to its centre C. The image O' is virtual.

It is thus shown that a normal pencil of rays gives rise on reflection to another such pencil, having either a real or a virtual focus. If the pencil is oblique, this is not the case. Thus, if $\overline{OP}_1$ and $\overline{OP}_2$ are any two oblique rays which are close together, after reflection they will cross at a point F_1 off the axis; and we can deduce at once the main phenomena for the whole oblique *pencil* of rays from O, by rotating the plane figure in the cut through a small angle around $\overline{OM}$ as an axis. The two rays $\overline{OP}_1$ and $\overline{OP}_2$ will thus describe a cone having as its base a small rectangular area, of which $\overline{P_1P_2}$ is one of the sides. The point F_1 will describe a short straight line perpendicular to the plane of the paper. So all the rays making up this oblique pencil will, after reflection, pass through two small elongated areas which are practically short straight lines: one at F_1 perpendicular to the plane of the paper; and

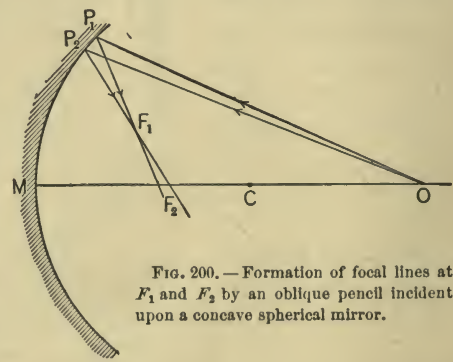


FIG. 200.—Formation of focal lines at F_1 and F_2 by an oblique pencil incident upon a concave spherical mirror.

one at F_2 in the plane of the paper at right angles to the reflected rays. These two lines are called "focal lines." Thus an oblique pencil gives rise to an astigmatic one. The phenomenon is called "spherical aberration," meaning that after reflection at the spherical surface the oblique pencil does not have a point focus. If we consider all the rays from O which fall upon the concave surface, it is seen that after reflection consecutive rays intersect at points which lie on a *surface* having a cusp at O', the image in the mirror of O for a normal pencil. This surface, a section of which is shown in the cut, is called the "caustic by reflection at a spherical surface." (Sections of this surface may be seen by looking at the reflection of a light in the mirror formed by a tumbler or cup containing milk or some opaque liquid. The cylindrical surface of the cup or glass forms the mirror, and the opaque surface is the screen on which the image is formed.) For the treatment of caustics by means of waves, the student should refer to the description of a method devised by Professor Wood of the Johns Hopkins University, which is given in Edser, *Light*, page 298.

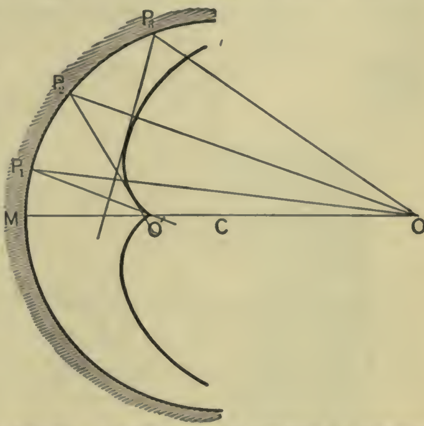


FIG. 201. — Caustic formed by reflection at a concave spherical mirror.

Formation of Images. — If there is an illuminated object $\overline{ON}$, its image will be $\overline{O'N'}$, as shown in Fig. 202, in which $\overline{OP}$ is parallel to $\overline{NCR}$, and where F bisects the line $\overline{CR}$. Each point of $\overline{ON}$ gives rise to an image at a point of $\overline{O'N'}$; and, if $\overline{ON}$ is perpendicular to the line $\overline{NCR}$, $\overline{O'N'}$ will be also, if $\overline{ON}$ is small. Two cases are illustrated in the cut. In one, the image is a "real" one, because the waves converge toward it after reflection; and if a screen is placed at $\overline{O'N'}$, a sharp *inverted* image will be formed on it; in the other it is virtual.

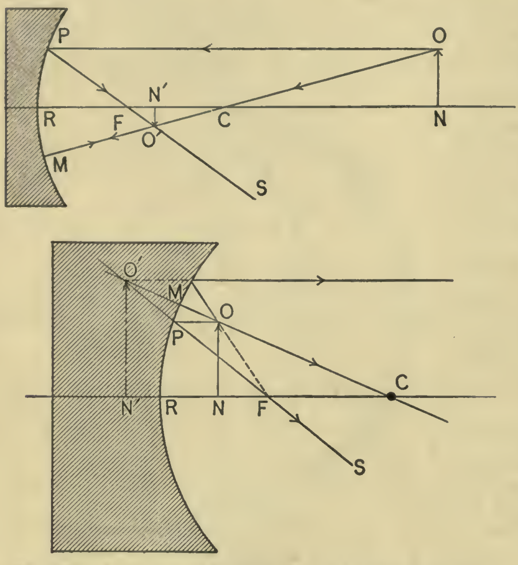


FIG. 202.—Formation of images by a concave spherical mirror.

The ratio of the *linear* magnitudes of the object and image, *i.e.* $\frac{\overline{ON}}{\overline{O'N'}}$, is evidently by geometry equal to the ratio $\frac{\overline{OC}}{\overline{O'C}}$, *i.e.* to $\frac{u-r}{r-v}$. Owing to the relation $\frac{1}{u}+\frac{1}{v}=\frac{2}{r}$, this ratio equals $\frac{u}{v}$. The ratio of the *areas* of the object and image is then $\frac{u^2}{v^2}$.

2. *Convex Mirrors.*—The same proof may be carried through for a convex mirror. Let O be the source, C the centre of the spherical surface, P any point of the surface near M, the intersection of $\overline{OC}$ with the surface; draw $\overline{OP}$ and $\overline{PS}$ so as to make equal angles with the radius $\overline{CPC'}$. Prolong $\overline{SP}$ backward and let it meet $\overline{OC}$ in O'; then O' is the virtual image of O. Further, by geometry,

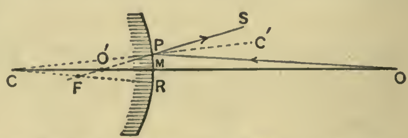


FIG. 203.—Formation of an image O' of a point source O by a convex spherical mirror.

$$\overline{PO}:\overline{PO'}=\overline{CO}:\overline{O'C}.$$

Call, as before, $\overline{MC} = r$, $\overline{MO'} = v$, $\overline{MO} = u$; and the equation becomes $u : v = u - r : r - v$, if P is close to M.

Hence, $$\frac{1}{u} + \frac{1}{v} = \frac{2}{r}.$$

If $\overline{CR}$ is drawn parallel to $\overline{OP}$, and if F is the intersection of $\overline{SO'}$ and $\overline{CR}$, it may be proved with ease that $\overline{CF}$ equals $\overline{FR}$, if the curvature of the surface is slight and P is near M. Hence the construction of images

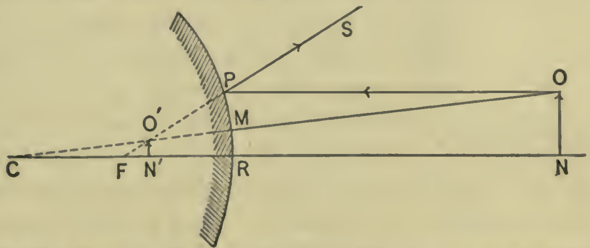


Fig. 204. — Formation of an image of an object ON by a convex spherical mirror.

is as shown in the cut: an illuminated object $\overline{ON}$ has a virtual image $\overline{O'N'}$. The distances of the object and image from the mirror are connected as before, by the relation

$$\frac{1}{u} + \frac{1}{v} = \frac{2}{r}, \text{ or } \frac{1}{v} = \frac{2}{r} - \frac{1}{u}.$$

Hence, so long as u is negative, *i.e.* so long as O is on the opposite side of the surface from C, or the waves diverge from a point on the *convex* side of the mirror, v is positive and the image lies on the same side of the mirror as the centre C.

A special case is when u is infinite, that is, when plane waves having the wave normal $\overline{OMC}$ fall upon the mirror; $v = \frac{r}{2}$, and the virtual image is therefore a point bisecting the line $\overline{CM}$. This is called the "principal focus."

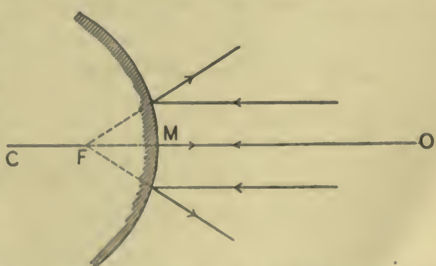


Fig. 205. — Special case: the source O is at an infinite distance.

Thus a normal pencil gives rise to a homocentric one; and it may be shown, by following the same method as was used for concave mirrors, that an oblique pencil produces an astigmatic one by reflection.

Plane Waves Incident upon a Parabolic Mirror. — This is a mirror whose surface may be imagined described by rotating a parabola around its axis; it is called a "paraboloid of revolution." If the rays are all parallel to the axis, they will after reflection *all* converge to the focus, *F*, of the parabola. (To prove this requires a knowledge of the analytical properties of the parabola.) In this case, then, there is no spherical aberration. Conversely, if a point source of light is placed at the focus *F*, all the rays which are reflected by the surface proceed out parallel to the axis. This is the reason why such mirrors are used in search lights, the headlights of locomotives, etc.

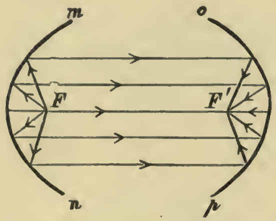


Fig. 206. — Parabolic mirrors.

CHAPTER XXVIII

REFRACTION

Plane Waves Incident upon a Plane Surface. — This is the case already discussed on page 432. The incident plane waves give rise to refracted waves, which are plane and in such a direction that the incident and refracted portions of any ray and the normal to the surface at the point of incidence are all in the same plane, viz., the "plane of incidence"; and, if N_1 and N_2 are the angles made with the normal by the two rays, $\frac{\sin N_1}{\sin N_2}$ is the same for all angles of incidence.

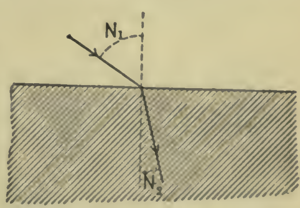


FIG. 207. — Refraction of a ray.

As already explained, this ratio, or the *index of refraction of the second medium with reference to the first*, as it is called, equals the ratio of the velocities of the waves in the two media. It should be noted that the waves are supposed to be homogeneous, *i.e.* to have a definite wave number. Calling v_1 and v_2 these velocities, and writing $n_{2,1}$ for this index of refraction, $n_{2,1} = \frac{\sin N_1}{\sin N_2} = \frac{v_1}{v_2}$. Conversely, the index of refraction of the first medium with reference to the second, $n_{2,1}$ equals $\frac{1}{n_{2,1}}$. Therefore, if $v_1 > v_2$, $\sin N_1 > \sin N_2$, and so $N_1 > N_2$; that is, the refracted ray is bent in closer to the normal than is the incident ray. This is the case illustrated in the cut. On the other hand, if $v_1 < v_2$, $N_1 < N_2$; and the refracted ray is bent away from the normal. This may be illustrated by the cut, if the arrows indicating the directions of the rays are considered reversed.

Total Reflection. — It is evident in this second class of refraction that, if the angle of incidence, N_2, is sufficiently increased, the angle of refraction, N_1, may finally equal 90°, *i.e.* $\frac{\pi}{2}$. So the refracted ray just grazes the surface. When this occurs, $\sin N_1 = 1$; and, therefore, according to the formula, $\sin N_2 = \frac{1}{n_{2,1}}$. This angle of incidence is called the "critical angle" for the two media and for waves of a definite wave number. If it can be measured, $n_{2,1}$, or the corresponding index of refraction, may be at once calculated. If the angle of incidence exceeds this critical angle in value, there is no refracted ray, for the sine of an angle cannot exceed unity, and the ray suffers *total reflection.* The velocity of ether waves in water is less than in air, as is shown by direct experiment, or by the fact that a ray in air incident upon a plane surface of water is bent toward the normal. So this phenomenon of total reflection will be observed if rays are incident obliquely upon a surface of water from below. This condition may be secured if one holds a tumbler of water in such a manner that the eye looks *up* through the glass at the surface of water and turns so as to face an illuminated object. If the direction in which one looks is sufficiently oblique to the surface, nothing is seen *through* it; for it acts like a plane mirror.

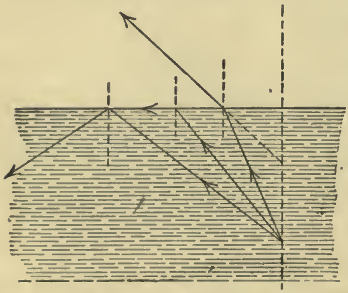


Fig. 208. — Total reflection.

A piece of apparatus that is often used to change the direction of a beam of light, called a "totally reflecting prism," consists of a glass triangular prism whose cross section is an isosceles right-angle triangle. Light incident normally upon one of the smaller faces suffers total reflection at the hypotenuse face and emerges perpendicular to the third face.

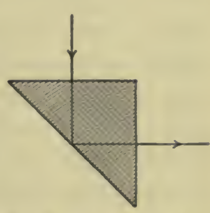


FIG. 209. — A totally reflecting prism.

Index of Refraction. — As defined above, the value of the index of refraction of a substance depends upon its optical properties with reference to some other medium. The natural medium to which all indices of refraction should refer is the pure ether; and so, when the words "the index of refraction" are used, this is implied. Thus, if v_0 is the velocity of waves in the pure ether, and $n_{1,0}$ the index of refraction of any substance, $n_{1,0} = \frac{v_0}{v_1}$; similarly, $n_{2,0} = \frac{v_0}{v_2}$ for a second substance, and therefore $n_{2,1} = \frac{n_{2,0}}{n_{1,0}}$. Experiments show that the index of refraction of air is very nearly unity, about 1.0003; and therefore the index of refraction of any solid or liquid substance, such as glass or water, with reference to air is practically the same as with reference to the pure ether. In actual measurements of indices of refraction, it is customary to determine them with reference to air and then to make the necessary correction. In the illustrations which follow, n will indicate the index of refraction of a substance with reference to air.

The fact that air refracts is shown by the unsteadiness of objects when seen through air rising from a hot stove or field; the different portions of the air, being at different temperatures, refract differently, and the effect is the same as if a pane of glass which is uneven and "streaky" is moved up and down in front of the eye. The refraction of air is important, too, in all astronomical observations.

The fact that the waves travel with different velocities in layers of air at different temperatures is shown by the fact that light is *reflected* by such layers. The explanation of mirage and similar phenomena depends upon this.

Special Cases. — The refracting matter is generally made into a figure with regular geometrical surfaces. There are three cases of special interest: (1) a "plate," which is a figure bounded in part by two parallel planes; (2) a "prism," which is a figure bounded in part by two non-parallel planes; (3) a "spherical lens," which is a figure bounded in part by two spherical surfaces, and which is symmetrical around the straight line joining their centres. We shall discuss briefly the path of a ray in passing through these various figures.

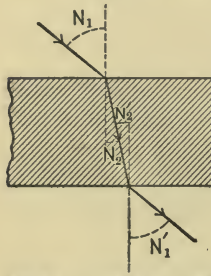


Fig. 210. — Refraction of a ray by a plate.

1. *Plate.* — Let the plate be placed with its parallel faces perpendicular to the sheet of the paper, and consider a ray incident in this plane. This is illustrated in the cut. If N_1 is the angle of incidence upon one plane face and N_2 that of refraction, the angle of incidence of the ray upon the other plane face, N_2', must by the laws of geometry equal the angle N_2; and the angle of refraction, or of emergence, out into the original medium, N_1', must equal N_1; for

$$\frac{\sin N_1}{\sin N_2} = n = \frac{\sin N_1'}{\sin N_2'},$$

and, since $N_2 = N_2'$, $N_1 = N_1'$.

Therefore the emerging ray is parallel to the incident one, but is displaced sidewise an amount depending upon the angle of incidence, the thickness of the plate, and its index of refraction.

It follows, then, that plane waves incident upon a plate emerge in the form of plane waves parallel to the incident waves. The case of spherical waves will be considered later.

2. *Prism.* — The straight line in which the two non-parallel surfaces meet (or would meet if prolonged) is called the

"edge," and the angle between them is called the "angle" of the prism. Let the prism be placed with its edge perpendicular to the plane of the paper, and consider a ray incident in this plane. This is illustrated in the cut. Call the angle of incidence upon the prism N_1; that of refraction, N_2; that of incidence upon the second face of the prism, N_2'; that of refraction out into the original medium, N_1'; that of the angle of the prism, A; and that between the directions of the entering and the emerging rays, D.

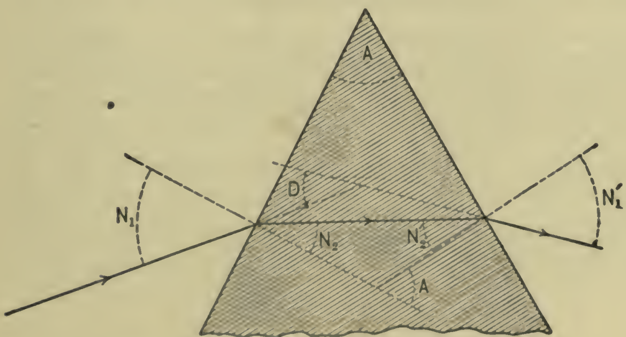


FIG. 211. — Refraction of a ray by a prism.

It is evident from the geometry of the figure that the angle between the two normals drawn to the two surfaces equals the angle of the prism; and that the following relations are true:

$$A = N_2' + N_2';$$

$$D = (N_1 - N_2) + (N_1' - N_2')$$

$$= N_1 + N_1' - A.$$

Further, N_1 and N_2, and N_1' and N_2' are connected by the general refraction formula; and so the value of D, the "deviation," as it is called, can be deduced in terms of A, N_1, and n.

If waves of different wave numbers (or colors) are incident upon a prism, it is observed that it deviates these waves to different degrees, thus showing that these waves have different indices of refraction. If n is large, the ray is refracted more than if n is small; and therefore the deviation is great, as is evident from the cut apart from the formula; as a result, if white light enters the prism, it is dispersed into a spectrum of colors. (With glass or water the shorter waves, *e.g.* the "blue ones," are deviated more than the longer ones, *e.g.* the "red ones.") (See Chapter XXX.)

It may be seen by actual experiment, and it may be proved by methods of the infinitesimal calculus, that as the angle of incidence is varied gradually from normal to grazing incidence, *i.e.* from 0° to 90°, the deviation gradually decreases, reaches a definite *minimum* value, and then increases; and, further, that this minimum deviation is obtained when the angle of incidence, N_1, equals that of emergence, N_1'; in other words, when the ray is symmetrical on the two sides of the prism. Call this angle of minimum deviation D; then, since $N_1 = N_1'$, $N_2 = N_2'$, and the two formulæ above become

$$A = 2N_2,$$

$$D = 2N_1 - A.$$

It follows that, since

$$n = \frac{\sin N_1}{\sin N_2},$$

we may write

$$n = \frac{\sin \frac{D+A}{2}}{\sin \frac{A}{2}}.$$

A and D may both be measured with accuracy; and so n may be obtained. (Reference for details of the method should be made to some laboratory manual.)

If homogeneous plane waves parallel to the edge of a prism are incident upon it, they will therefore emerge in the form of plane waves, but will be deviated through a certain angle. The case of spherical waves will be discussed later.

3. *Spherical Lens.* — Let the line joining the centres of the two spherical surfaces, the "axis" of the lens, as it is called, lie in the plane of the paper. The section of the lens will be one of the forms shown in the cut, for a plane surface is a

FIG. 212. — Different forms of lenses.

special case of a spherical one. These are called "double convex," "plano-convex," "concavo-convex," "double concave," etc. Consider a ray incident upon the lens. It is refracted into the lens at one point and out at another, exactly as if there were minute plane surfaces tangent to the lens at these points. These minute planes constitute a prism; and therefore a lens may be treated as a special case of a combination of a great number of prismatic faces. (Other lenses than spherical ones are often used, especially in the construction of spectacles.)

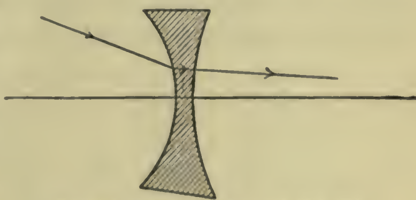


FIG. 213. — Refraction of a ray by a lens.

When plane waves are incident upon a lens, its different rays meet the surface of the lens at different angles of incidence and on emerging they are refracted out at different angles; so nothing can be said immediately in regard to the nature of the transmitted waves. Plane waves are a special case of spherical ones; and the theory of the action of a lens upon the latter will be given shortly.

The values of the indices of refraction of a few substances are given in the following table:

INDICES OF REFRACTION FOR YELLOW LIGHT OF WAVE LENGTH 0.0000589 CM.

Substance	Index	Temperature
Air	1.0002922	0° C. } pressure 76 cm.
Helium	1.000043	0°
Hydrogen	1.000140	0°
Nitrogen	1.000297	0°
Oxygen	1.000272	15°
Alcohol	1.360	15°
Chloroform	1.449	25°
Carbon bisulphide	1.624	25°
Water	1.334	16°
Rock salt	1.5441	24°
Flint glass	1.651	
Crown glass	1.524	

Spherical Waves Incident upon a Plane Surface. — Let the surface be perpendicular to the plane of the paper, and let the point source O lie in this plane. Let $\overline{OB}$ be any ray incident upon the surface at a point B near the foot of the perpendicular $\overline{OA}$ dropped from O upon the surface. Let $\overline{BC}$ be the refracted ray, and $\overline{BO_1}$ its prolongation backward until it meets the normal $\overline{OA}$. There are two cases to be considered, depending upon whether the velocity of waves in the first medium is greater or less than that in the second. In either case let n be the index of refraction of the second medium with reference to the first. If the waves in the

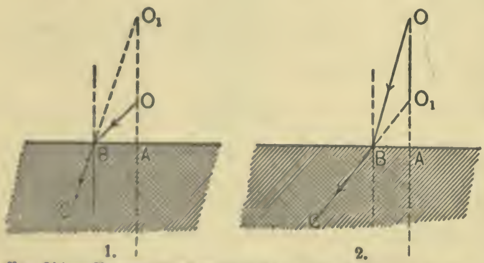


FIG. 214. — Formation of images of a point source O by a plane surface: (1) when $n > 1$; (2) when $n < 1$.

latter have a greater velocity than in the former, $n > 1$; if their velocity is greater in the former, $n < 1$. These two cases are illustrated by the two cuts.

In each the angle of incidence equals (BOA); and that of refraction (BO_1A). Therefore, since the sine of $(BOA) = \frac{\overline{AB}}{OB}$, and the sine of $(BO_1A) = \frac{\overline{AB}}{O_1B}$, the index of refraction, $n = \frac{O_1B}{OB}$. If B is extremely close to A, that is, if the ray $\overline{OB}$ is one of a normal pencil, we may replace the ratio $\frac{O_1B}{OB}$ by $\frac{O_1A}{OA}$. So $n = \frac{\overline{O_1A}}{OA}$ or $\overline{O_1A} = n\overline{OA}$. It follows, then, that O_1 is a point at a fixed distance from the plane surface for *all* the rays of the normal pencil provided the waves are homogeneous, so that n is a constant. If $n > 1$, O_1 is farther from the surface than O; if $n < 1$, it is nearer the surface. In other words, a normal pencil from a point source O gives rise to a pencil of rays by refraction whose centre is O_1, the virtual image of O. Conversely, if we imagine the directions of all the rays reversed, a normal pencil of rays in the second medium converging toward a point O_1 in the first will actually meet at a point O.

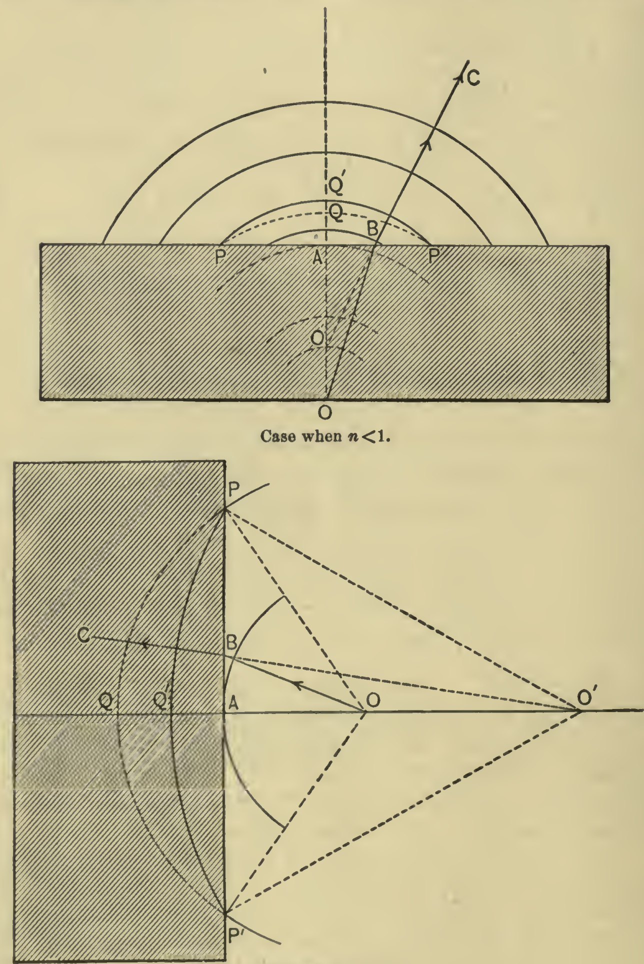
A luminous object in one medium will thus give rise to a virtual image of itself. This image will not be of the same size as the object; but their relative dimensions may be easily calculated.

The treatment of this case of refraction of spherical waves at a plane surface by the method of waves is as follows:

Let the source O be in the medium in which the waves have the less velocity, which may therefore be called the "slow medium," and let the waves meet the surface of the "faster medium" at A, as shown in Fig. 215. $\overline{OA}$ is therefore perpendicular to the surface. If the velocity of the waves in the two media were the same, the wave front would advance in a definite time to the position PQP'; but, since the new medium is "faster," the disturbance reaches Q' instead of Q, and the actual curve of the entering waves is $PQ'P'$, an arc of a circle whose centre O' is on the line $\overline{OA}$, but nearer the surface than O. O' is then the virtual image of O.

Consider also the case where the source O is in the "faster medium"; the spherical waves meet the surface at A, the foot of the perpendicular from O upon the plane surface.

The waves entering the other medium would have the form PQP' if the velocity were unchanged; but, since the medium is "slower," the



Case when $n < 1$.

Case when $n > 1$.

FIG. 215.—Refraction of spherical waves by a plane surface.

actual curve is $PQ'P'$, an arc of a circle whose centre is O', a point on $\overline{OA}$, but farther away from the surface than O. O' is, as before, the virtual image of O.

The distance between the point source and its image may be calculated from the formula already deduced: $\overline{O_1A} = n\overline{OA}$. For $\overline{OO_1} = \overline{OA} - \overline{O_1A} = \overline{OA}(1-n)$. Thus, if $n > 1$, the object appears to be farther from the surface than it really is; while if $n < 1$, the opposite is true. This is illustrated when one looks at a stone lying at the bottom of a pond of water; in this case O is at the bottom of the water, and O_1 is nearer to the surface by a distance $\overline{OA}(1-n)$, where $\overline{OA}$ is the actual depth of the water and n is the index of refraction of the air with reference to the water; *i.e.* it equals $\frac{1}{n'}$, where n' is the index of refraction of the water with reference to the air. (n' for water is approximately $\frac{4}{3}$; so $\overline{OO_1}$ is $\frac{\overline{OA}}{4}$.) A method is thus offered for the measurement of n, since both $\overline{OA}$ and $\overline{OO_1}$ can be measured.

If the pencil of rays from O is oblique, it forms by refraction an astigmatic pencil. Thus,

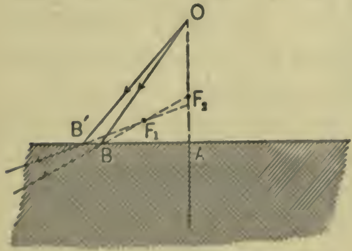


FIG. 216. — Formation of focal lines at F_1 and F_2 by the refraction of an oblique pencil at a plane surface.

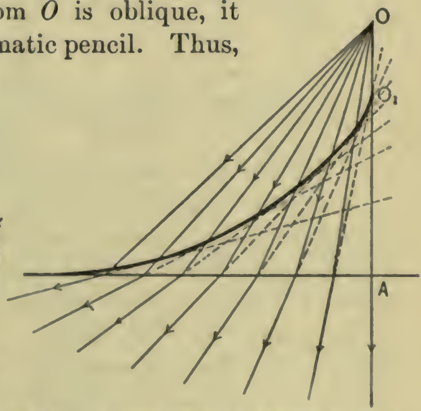


FIG. 217. — Caustic formed by reflection at a plane surface.

if $\overline{OB}$ and $\overline{OB'}$ are two oblique rays, they form by refraction two rays which if prolonged backward cross at F_1, a point off the perpendicular line $\overline{OA}$. Therefore, the whole oblique pencil gives rise to two virtual focal lines: one at F_1 perpendicular to the sheet of the paper: the other at F_2 along the line $\overline{OA}$. Further, if we consider all the rays from O falling upon the surface, they give rise by refraction to a virtual caustic surface with a cusp at O_1, the image of O for a normal pencil.

Special Cases

1. *Plate.* — We shall consider the plate made of a material in which the waves have a less velocity than in the surrounding medium, *i.e.* $n > 1$. Let O be the source of the spherical waves. The path of any ray is indicated by $\overline{OB}$, $\overline{BB'}$, $\overline{B'C}$. O' is the image of O in the first surface; and O'' is the image of O' in the second one. So, if the pencil is a normal one, all the rays leaving O will diverge after emerging from the plate as if they came from O''. Its position may be at once calculated.

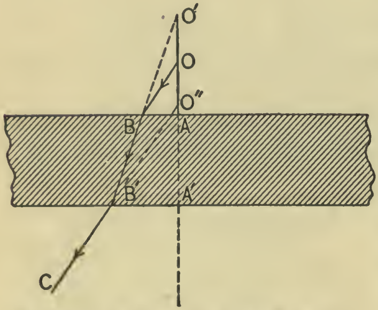


FIG. 218. — Formation of an image of a point source O by refraction through a plate.

2. *Prism.* — We shall consider the prism made of a material in which the waves have a less velocity than in the surrounding medium, *i.e.* $n > 1$. Let O be the source of spherical waves. The path of any ray is indicated by $\overline{OB}$, $\overline{BB'}$, $\overline{B'C}$. O' is the image of O in the first surface; and O'' is the image of O' in the second surface. It should be observed that, if the ray is one of the normal pencil

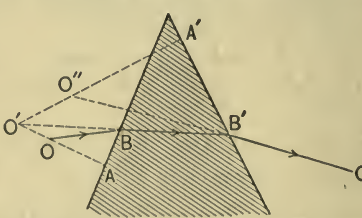


FIG. 219. — Formation of an image of a point source O by refraction through a prism.

for one surface it is oblique to the other when it is incident upon it. Therefore, in general, an incident pencil of rays will, after two refractions, emerge in the form of an astigmatic pencil; the rays will apparently come from two focal lines. It may be shown, however, that, if the pencil of rays is incident upon the prism at that angle which corresponds to minimum deviation, the virtual image from which the emerging rays apparently diverge is practically a point. So in this case the emerging pencil is homocentric.

Owing to the difference in the index of refraction for waves of different wave number, it is evident that, if there is at O a source of several different trains of waves, each of them will have a different virtual image O''. The greater the value of n, *i.e.* the shorter the wave length, so much the more is O'' displaced from O. The prism then disperses the light.

3. *Spherical Lens.* — It is evident that any ray which falls upon a lens suffers refraction twice, once on entering the lens and again on leaving it. We must therefore discuss, as a preliminary to the treatment of the lens, the refraction of a ray at a spherical surface. There are two cases, which we shall consider separately: refraction at a concave surface and at a convex one.

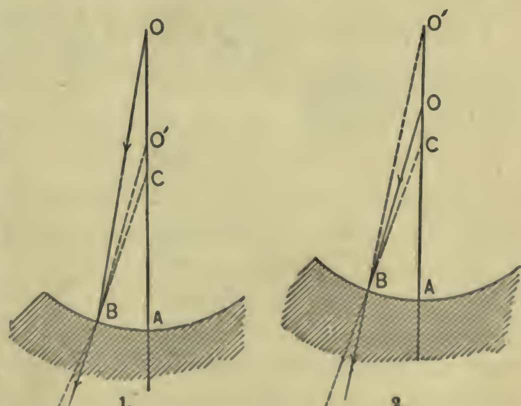


FIG. 220. — Formation of images of a point source O by refraction at a concave spherical surface: (1) when $n>1$; (2) when $n<1$.

a. Concave Surface. — Let the section made by a plane through the centre of the surface C and the point source O, *i.e.* through the "axis" $\overline{OA}$, be that shown in the cuts; let $\overline{OB}$ be any incident ray, and $\overline{O'B}$ the prolongation backward of the refracted ray; and draw $\overline{CB}$, which is therefore normal at B. (The two cuts illustrate respectively the cases when the index of refraction of the second medium with reference to the first is greater and when it is less than one.)

The angle of incidence is $(OBC) = N_1$.

The angle of refraction is $(O'BC) = N_2$.

Then by the well-known trigonometrical formula, which states that in any plane triangle the ratio of the sines of any two angles equals that of the lengths of the opposite sides, we have from the triangles (OBC) and $(O'BC)$:

$$\sin N_1 : \sin (BOC) = \overline{OC} : \overline{BC},$$

$$\sin N_2 : \sin (BO'C) = \overline{O'C} : \overline{BC}.$$

Therefore, since $\frac{\sin N_1}{\sin N_2} = n$, and $\frac{\sin (BOC)}{\sin (BO'C)} = \frac{\overline{O'B}}{\overline{OB}}$ (as is seen by dropping a perpendicular line from B upon $\overline{OA}$).

$$n = \frac{OC}{O'C}\frac{O'B}{OB}.$$

If B is very close to A, that is, if the ray is one of a normal pencil and if the curvature of the surface is small, we may replace the ratio $\frac{O'B}{OB}$ by $\frac{O'A}{OA}$; and so $n = \frac{\overline{OC}}{\overline{O'C}} \times \frac{\overline{O'A}}{\overline{OA}}$. This formula shows that, if a normal pencil from O is incident upon the surface, it gives rise to a homocentric pencil with its centre at O'; so O' is the virtual image of O. This can be expressed in a simpler mathematical form. Call the distances $\overline{OA}$, u; $\overline{O'A}$, v; $\overline{CA}$, r.

Then

$$\overline{OC} = u - r, \ \overline{O'C} = v - r;$$

and

$$n = \frac{u - r}{v - r}\frac{v}{u}.$$

Hence $$\frac{n-1}{r}=\frac{n}{v}-\frac{1}{u},$$

or $$\frac{n}{v}=\frac{1}{u}+\frac{n-1}{r}.$$

If $n<1$, it may happen, therefore, for a suitable value of O that v is negative. This means simply that O' is in on the opposite side of the surface from C; for in this formula the *positive* direction is that defined by r, namely from C to A.

b. Convex Surface.—The same treatment and the same formulæ as those above lead to the same result for a convex surface, viz., $n=\frac{OC}{O'C}\cdot\frac{O'A}{OA}$. And since $u=\overline{OA}$ and $r=\overline{CA}$, $\overline{OC}=\overline{OA}+\overline{AC}=\overline{OA}-\overline{CA}$. Hence, as before, $\overline{OC}=u-r$, and $\overline{O'C}=v-r$; and on substitution, we again have

$$\frac{n-1}{r}=\frac{n}{v}-\frac{1}{u}, \text{ or } \frac{n}{v}=\frac{1}{u}+\frac{n-1}{r}.$$

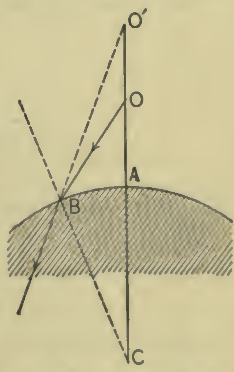


FIG. 221. — Formation of an image of a point source O by refraction at a convex spherical surface, case when $n>1$.

If O is on the opposite side of the surface from C, u has a negative value. For instance, if in the cut the *length* of the line OA is x, the value of u is $-x$.

If n' is the index of refraction of the first medium with reference to the second, $n'=\frac{1}{n}$; and the general formula becomes

$$\frac{1}{n'v}=\frac{1}{u}+\frac{1-n'}{n'r}, \text{ or } \frac{1}{v}=\frac{n'}{u}+\frac{1-n'}{r}.$$

Conversely, if a pencil of rays in the second medium is converging apparently toward a point O' in the first medium, they will actually meet at the point O.

If the pencil of rays is oblique, it gives rise to an astigmatic pencil; and if all the incident rays are considered, the image is a caustic surface. This phenomenon is said to be due, as in other similar cases, to spherical aberration.

We shall now return to the problem of the refraction produced by a spherical lens. It is evident from what has just been shown that a normal pencil from any point on the axis of the lens will give rise to a homocentric pencil emerging from the lens after the two refractions; the image of the source, though, may be either virtual or real. We shall deduce the formula for a double convex and for a double concave lens, and then show that, by a suitable agreement as to signs, one formula may be used for all lenses. We shall assume at first that the lens is so *thin* that a ray incident at any point of one surface emerges from the other surface at a point which is at the same distance from the axis as is the former.

a. Double Convex Lens. — Consider a section through the axis.

Let O be the point source.

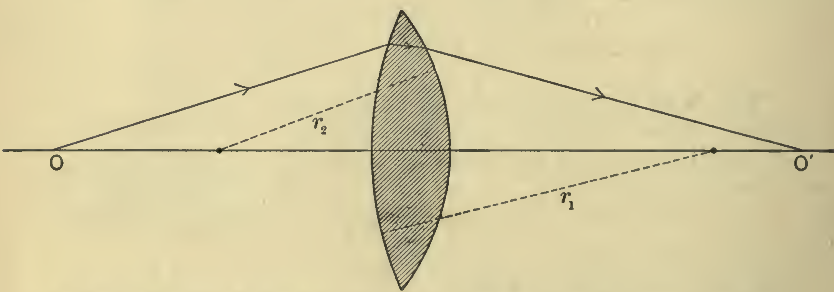


Fig. 222. — Formation of an image of a point source O by refraction through a double convex lens.

Let r_1 be the radius of the first spherical surface of the lens.

Let r_2 be the radius of the second spherical surface of the lens.

Let n be the index of refraction of the lens with reference to the surrounding air.

The formula for refraction at the first surface of the lens is, then, $\frac{n}{v_1} = \frac{1}{u_1} + \frac{n-1}{r_1}$, where u_1 and v_1 are *positive* if O and its image lie on the *right* of the lens, because the centre of the first surface of the lens lies on this side. The refraction at the second surface is from the lens out into the air; so the formula is $\frac{1}{v_2} = \frac{n}{u_2} + \frac{1-n}{r_2}$, where u_2 and v_2 are *positive* if the points to which they refer are to the *left* of the lens. But, if the lens is *thin*, $v_1 = -u_2$, for the rays incident upon the second surface are those diverging from the virtual image produced at the first; but a quantity u or v which is positive with reference to one surface is negative with reference to the other, since their centres are on opposite sides of the lens.

Therefore we have the two formulæ:

$$\frac{n}{v_1} = \frac{1}{u_1} + \frac{n-1}{r_1},$$

$$\frac{1}{v_2} = -\frac{n}{v_1} + \frac{1-n}{r_2}.$$

Hence, $$\frac{1}{u_1} + \frac{1}{v_2} = -(n-1)\left(\frac{1}{r_1} + \frac{1}{r_2}\right).$$

In this formula, u_1 is the distance O lies on the axis to the *right* of the lens (referring to the cut); and v_2 is the distance the image produced by the second surface of the lens lies to its *left*. Therefore, if we agree to call the distance the *point source* lies to the *left* of the lens u, and the distance the image lies to the *right* of the lens v, $u = -u_1$, $v = -v_2$. So the formula becomes

$$\frac{1}{u} + \frac{1}{v} = (n-1)\left(\frac{1}{r_1} + \frac{1}{r_2}\right).$$

The quantity on the right-hand side of the equation is a constant quantity for a given lens, and it is essentially positive if $n > 1$, as it is in all ordinary cases, *e.g.* glass, quartz, etc. lenses surrounded by air. We write this quantity $\frac{1}{f}$; and the formula then assumes the final form

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f}$$

b. Double Concave Lens. — In this case the formulæ for refraction at the two surfaces are, as before,

$$\frac{n}{v_1} = \frac{1}{u_1} + \frac{n-1}{r_1},$$

where u_1 and v_1 are positive if O and its image in the first surface lie on the *left* of the lens; and

$$\frac{1}{v_2} = -\frac{n}{v_1} + \frac{1-n}{r_2},$$

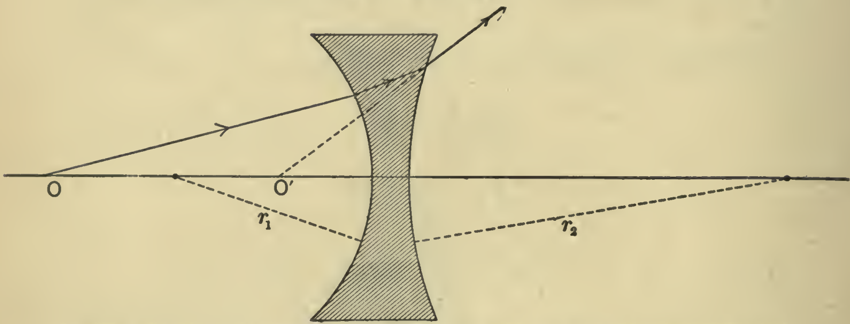


FIG. 223. — Formation of an image of a point source O by refraction through a double concave lens.

where u_2 and v_2 are positive if the points to which they refer are on the *right* of the lens. Hence,

$$\frac{1}{u_1} + \frac{1}{v_2} = -(n-1)\left(\frac{1}{r_1} + \frac{1}{r_2}\right);$$

or, introducing the same agreement as to signs as in the previous case, $u = u_1$, $v = v_2$; and hence

$$\frac{1}{u} + \frac{1}{v} = -\frac{1}{f}.$$

We can therefore use for both kinds — in fact, for all kinds — of lenses the one formula,

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f},$$

in which u and v have the meanings agreed upon above and f is positive for some lenses and negative for others. We shall discuss presently the properties of these two types of lenses.

The formula for a lens may be deduced in another manner, possibly more instructive, if we *assume* the fact that a point source on the axis gives rise to a point image.

The principle made use of in this second method is that the "optical" lengths of the paths from the point source to its image are the same along all rays. (See page 436.)

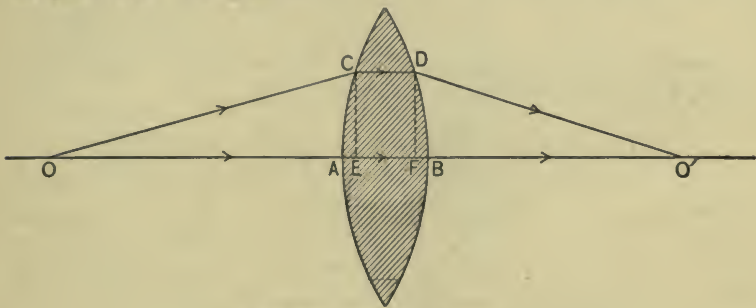


FIG. 224. — Refraction by a lens.

Consider the case of a *thin* double convex lens, as shown in the cut on an enormously magnified scale. Let O and O' be the source and its image, both on the axis; so $\overline{OA}$, $\overline{AB}$, $\overline{BO'}$ is the *normal* ray from O. Let $\overline{OC}$ be any incident ray making a small angle with the axis. Since the lens is assumed to be thin, the portion of the ray within the lens, namely $\overline{CD}$, is parallel to the axis; and the other portion is $\overline{DO'}$. Then the "optical" lengths of the two rays are

$$\overline{OA} + n\overline{AB} + \overline{BO'} \text{ and } \overline{OC} + n\overline{CD} + \overline{DO'}.$$

Since these are equal,

$$\overline{OA} + n\overline{AB} + \overline{BO'} = \overline{OC} + n\overline{CD} + \overline{DO'},$$

or

$$\overline{OC} - \overline{OA} + \overline{DO'} - \overline{BO'} = n(\overline{AB} - \overline{CD})$$

$$= n(\overline{AE} + \overline{FB}).$$

Paying strict attention to the signs of the lines, this may be written

$$\overline{OC} - \overline{OE} + \overline{AE} + \overline{DO'} - \overline{FO'} + \overline{FB} = n(\overline{AE} + \overline{FB}),$$

or

$$(\overline{OC} - \overline{OE}) + (\overline{DO'} - \overline{FO'}) = (n - 1)(\overline{AE} + \overline{FB}).$$

These three quantities may be expressed in a simple form provided the angles between the rays and the axis are small, as we have assumed to be the case. Thus, describing an arc of a circle $\overline{RS}$ around P as a centre, and drawing the two radii $\overline{PR}$ and $\overline{PS}$, and the perpendicular line $\overline{RQ}$ from R upon $\overline{PS}$, we have the following formula:

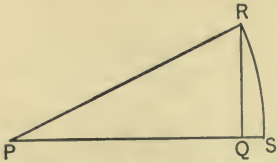


FIG. 225. — *RS* is portion of a circle whose centre is *P*.

$$\overline{PQ}^2 = \overline{PR}^2 - \overline{RQ}^2 = \overline{PR}^2\left(1 - \frac{\overline{RQ}^2}{\overline{PR}^2}\right).$$

Hence, if $\overline{PR}$ is large compared with $\overline{RQ}$, we have, on expansion by the binomial theorem and neglecting small terms,

$$\overline{PQ} = \overline{PR}\left(1 - \frac{1}{2}\frac{\overline{RQ}^2}{\overline{PR}^2}\right) = \overline{PR} - \frac{1}{2}\frac{\overline{RQ}^2}{\overline{PR}} = \overline{PS} - \frac{1}{2}\frac{\overline{RQ}^2}{\overline{PS}}.$$

Therefore, $$\overline{PR} - \overline{PQ} = \frac{1}{2}\frac{\overline{RQ}^2}{\overline{PR}};$$

or, $$\overline{QS} = \frac{1}{2}\frac{\overline{RQ}^2}{\overline{PS}}.$$

In the formula, as given above, for a lens, then,

$$\overline{OC} - \overline{OE} = \frac{1}{2}\frac{\overline{CE}^2}{\overline{OC}},$$

$$\overline{DO'} - \overline{FO'} = \frac{1}{2}\frac{\overline{DF}^2}{\overline{DO'}} = \frac{1}{2}\frac{\overline{CE}^2}{\overline{DO'}},$$

$$AE = \frac{1}{2}\frac{\overline{CE}^2}{r_1},$$

$$\overline{FB} = \frac{1}{2}\frac{\overline{DF}^2}{r_2} = \frac{1}{2}\frac{\overline{CE}^2}{r_2}.$$

Consequently, $$\frac{1}{OC} + \frac{1}{DO'} = (n-1)\left(\frac{1}{r_1} + \frac{1}{r_2}\right).$$

If the incident ray makes a sufficiently small angle with the axis, and if the curvatures of the two surfaces are small, we may replace $\overline{OC}$ by $\overline{OA}$, *i.e.* u, and $\overline{DO'}$ by $\overline{BO'}$, *i.e.* v; and we have the general formula, as before,

$$\frac{1}{u} + \frac{1}{v} = (n-1)\left(\frac{1}{r_1} + \frac{1}{r_2}\right) = \frac{1}{f}.$$

We shall now discuss the two types of lenses referred to above: for one f is positive; for the other, negative.

a. Lenses for which f is positive. — The general formula is $\frac{1}{u}+\frac{1}{v}=\frac{1}{f}$. Let us consider several special cases.

If the point source is removed farther and farther from the lens, but kept on the axis, u approaches an infinite value and the waves become plane as they reach the lens. When $u=\infty$, it is seen from the formula that $v=f$; and, since f is positive, this means that there is a *real* image at a distance f from the lens. This point is called a "principal focus" of the lens. We may express this fact in words by saying that plane waves advancing with their normal parallel to the axis are changed by the lens into converging spherical waves whose centre is at a distance f from the lens; or, rays parallel to the axis are bent by the lens in such a manner as to converge to a point on the axis at a distance f from the lens.

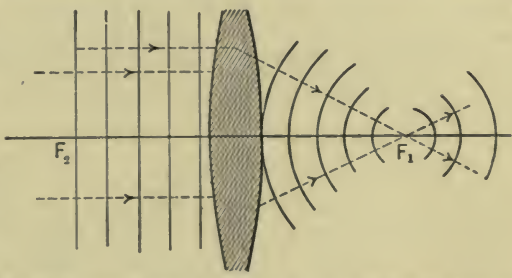


FIG. 226. — Special case: the source O is at an infinite distance on the axis.

Similarly, if $u=f$, it is seen from the formula that $v=\infty$. In words this states that, if the point source is on the axis of the lens at a distance from it equal to f, the diverging spherical waves are so converged by the lens as to become plane and to advance in a direction parallel to the axis; or, we may say that rays through a point on the axis at a distance f from it are so deflected by the lens as to become parallel to its axis. Conversely, we can have plane waves incident upon the lens from the other side, which will converge to a point at a distance f from it. There are thus two principal foci; one on each side of the lens, and at the same distance f from it, *if the lens is thin.* This distance is called the "focal length."

Therefore, if we consider the point source as at an infinite distance from the lens, its image is the principal focus on the other side; and, if the source approaches the lens, the image recedes from it, until, when the source reaches the principal focus, the image is at an infinite distance. When the source is between the principal focus and the lens, *i.e.* when $u < f$, it is seen from the formula that $v < 0$, so the image is on the same side of the lens as is the source, and is, therefore, virtual. When the source reaches the lens, *i.e.* when $u = 0$, $v = 0$ also; so object and image coincide. If u has a negative value, the physical meaning is that rays are converging on the lens apparently toward a point on the other side, which may be called a "virtual" source; and, in this case, as is seen from the formula, v is positive and less numerically than u; but it should be noted that v does not exceed f in value so long as u is negative. Therefore, the converging rays are converged still more, and form a real image.

It is thus seen that a lens for which f is positive always converges waves which fall upon it; for this reason it is called a "converging lens." The ordinary form of a converging lens is double convex; but any *thin* spherical lens thicker along the axis than elsewhere is a converging one.

We may arrange in tabular form the facts proved above in regard to u and v:

$u = \infty$	$v = f$
$\infty > u > f$	$\infty > v > f$
$u = f$	$v = \infty$
$f > u > 0$	$\infty > -v > 0$
$0 > u > -\infty$ . . .	$f > v > 0$

It is a simple matter to find by geometrical methods the position of the image of an object, for we know the effect of the lens upon *three* of the rays from any point source: A ray parallel to the axis is deflected so as to pass through the principal focus on the other side of the lens; a ray passing through the principal focus on the "incident side" of the lens will emerge on the other side parallel to the axis; a ray meeting the lens at the point where it is intersected by the

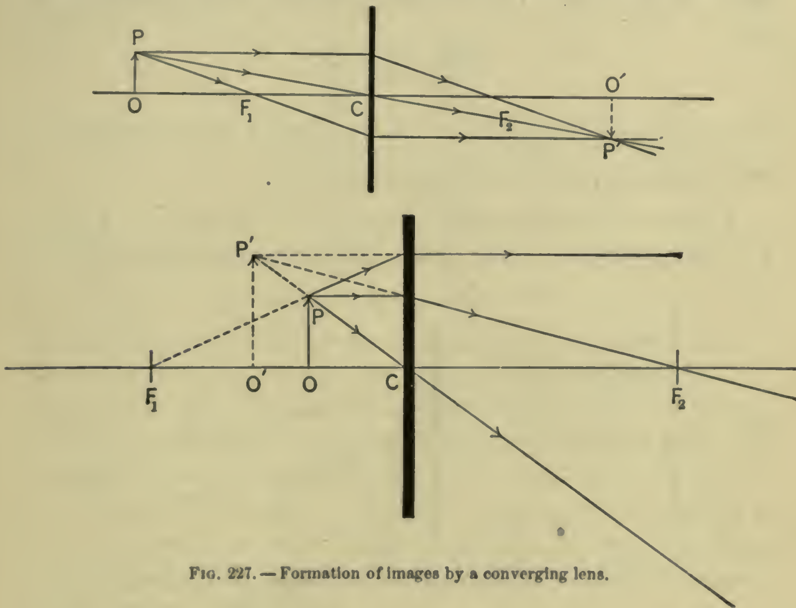


FIG. 227. — Formation of images by a converging lens.

axis, *i.e.* the "centre of the lens," keeps its direction unaltered, because at this point the two faces of the lens are parallel, and it is equivalent to an infinitely thin *plate*.

In showing graphically the formation of images by a lens, we shall represent the lens by a straight line, for simplicity. Thus, let F_1, F_2, and C be the principal foci and the centre of the lens, and P any point of the object; its image is at P'. Two cases are shown: if P is farther from the lens than the principal focus, the image is real; if P is between the lens and the principal focus, the image is virtual.

If the object is *small*, close to the axis, and perpendicular to it, as represented by $\overline{OP}$, its image $\overline{O'P'}$ is also perpendicular to the axis. So O' is the image of O; and, since these points are on the axis, $\overline{OC} = u$ and $\overline{CO'} = v$. Further, by similar triangles, $\overline{PC} : \overline{CP'} = u : v$. The ratio of the length of the line $\overline{O'P'}$ to that of $\overline{OP}$ is called the "*linear* magnification" of the lens. It is evident from the geometry of the cut that

$$\frac{\overline{O'P'}}{\overline{OP}} = \frac{\overline{CO'}}{\overline{OC}} = \frac{v}{u}.$$

Therefore, the magnification of the *surface* of any portion of the object perpendicular to the axis is $\frac{v^2}{u^2}$.

A case of special interest is when the source, P, is placed in a plane perpendicular to the axis at the principal focus, or

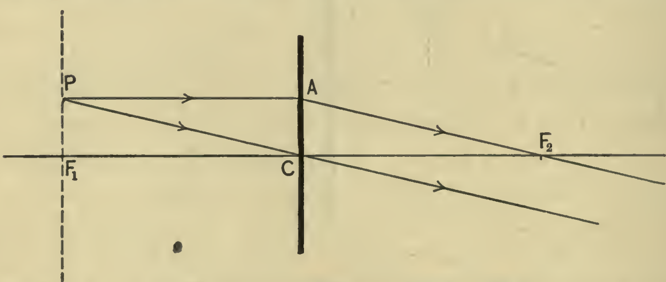


FIG. 228.—Special case: the point source P is in the focal plane of the converging lens.

in the "focal plane," as it is called. It is seen from the geometry of the cut that the two emerging rays which we can draw from known principles are parallel to the line joining P to the centre of the lens, C; for the triangles (PF_1C), (PAC), and (ACF_2) are all equal. Therefore, *all* the rays diverging from a point in the focal plane, which meet the lens, emerge parallel to the line joining this point to the centre of the lens; or, in other words, spherical waves diverging from such a point are made plane by the lens and proceed in the direction of the line referred to. Conversely, if we imagine these rays reversed, parallel rays incident obliquely upon the lens, but, at a small angle, converge to a point in the focal plane where this plane is intersected by a line through the centre of the lens parallel to the beam of rays; or, plane waves incident obliquely at a small angle upon a lens are brought to a focus at a point in the focal plane, as just defined.

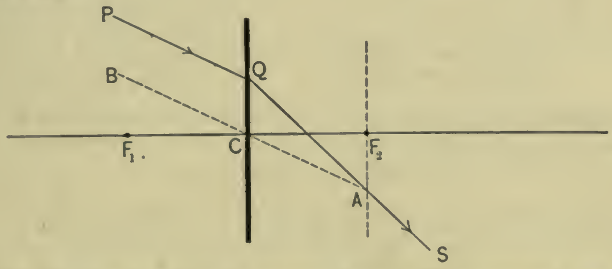


FIG. 229.—Construction for the refraction of any ray $\overline{PQ}$ by a converging lens.

This fact enables us to construct at once the refracted portion of any incident ray. Thus, let $\overline{PQ}$ be a ray incident at Q upon the lens, whose principal foci and centre are F_1, F_2, and C. Draw the focal plane through F_2, and a line $\overline{BC}$ through C parallel to $\overline{PQ}$; this meets the former in a point A. Draw the line $\overline{QS}$, joining Q to A; this is the emerging ray. For, if there were two parallel rays $\overline{PQ}$ and $\overline{BC}$, they would converge to A; and therefore $\overline{PQ}$ must produce a ray which passes through A.

Again, referring to Fig. 230, let there be two beams of parallel rays incident upon the lens, one having the direction $\overline{PC}$, the other $\overline{P_1C}$; they will be brought to a focus at Q and Q_1, in the focal plane. (This is the case when an image of a *distant* object is formed on a screen or photographic plate by a lens, for each point of the object sends out rays which are approximately parallel to each other when they reach the lens.) For a given "angular separation" of the two beams, *i.e.* (PCP_1) the distance $\overline{QQ_1}$ is evidently greater if the focal length of the lens is increased. If this angle is small, and if f is the focal length, the length of the line $\overline{QQ_1}$ equals the product of f by this angle, for the angles (PCP_1) and (QCQ_1) are equal. If these two rays come from two points

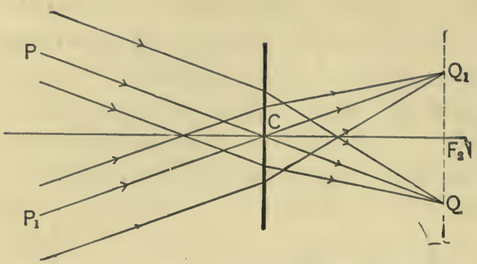


Fig. 230. — Formation by a converging lens of an image of an object at an infinite distance.

on the edge of the distant object, its image will be bounded by the two points Q and Q_1; and the *linear* dimensions of this image will vary directly as the focal length of the lens. Therefore the *area* of the image will vary as the square of the focal length. If a lens having a long focal length is used, the size of the image is great; but, if a photograph is to be taken, the time of exposure must be prolonged because the energy is distributed over a large area.

b. Lenses for which f is Negative. — The general formula is $\frac{1}{u}+\frac{1}{v}=\frac{1}{f}$. If the source is at an infinite distance, $u=\infty$, and therefore $v=f$; but since f is negative, v is also, and the image is a virtual one on the same side of the lens as the incident waves. This is called a "principal focus"; and its distance, f, from the lens is called the "focal length." Similarly, there is another principal focus on the axis on the opposite side of the lens, and at the same distance from it if the lens is thin.

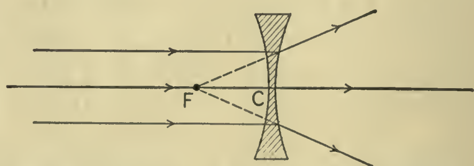


Fig. 231. — Special case: the point source is at an infinite distance on the axis.

If $u=f$, *i.e.* if the rays are converging apparently toward the principal focus on the opposite side of the lens, it is seen from the formula that $v=\infty$, *i.e.* the emerging rays are all parallel to the axis. We can, moreover, in a similar manner to that used in the previous case, discuss the relation between the positions of the object and image as the point source moves from $+\infty$ to $-\infty$. It is seen at once that the effect of the lens is to make the incident rays or waves diverge; for that reason lenses of this type are called "diverging lenses." All *thin* lenses which are thinnest at their centres are diverging.

We can also deduce graphically the position of the image of any point source, because we know the effect of the lens upon three rays: a ray parallel to the axis emerges in such a direction that if prolonged backward it would meet the axis at the principal focus; a ray pointed toward the principal focus on the other side of the lens emerges parallel to the axis; a ray through the centre of the lens retains its direction unchanged. A few cases will be drawn, the lens being represented as before by a straight line.

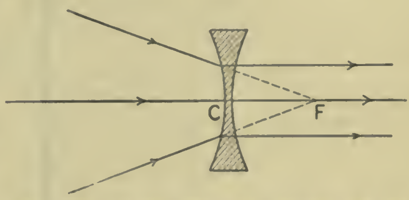


FIG. 232. — Special case: the incident rays are converging toward the principal focus on the farther side of the lens.

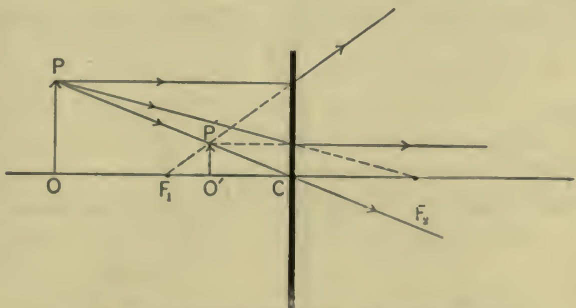


FIG. 233. — Formation of an image by a diverging lens.

A real source P gives rise to a virtual image P'; and a *small* object $\overline{OP}$ perpendicular to the axis has an image $\overline{O'P'}$ also perpendicular to the axis.

The linear magnification produced by the lens, that is the ratio $\frac{O'P'}{OP}$, equals as before $\frac{v}{u}$; and the surface magnification is $\frac{v^2}{u^2}$.

A case of special interest is when the virtual source is in the focal plane. See Fig. 235. Let this point be P; draw

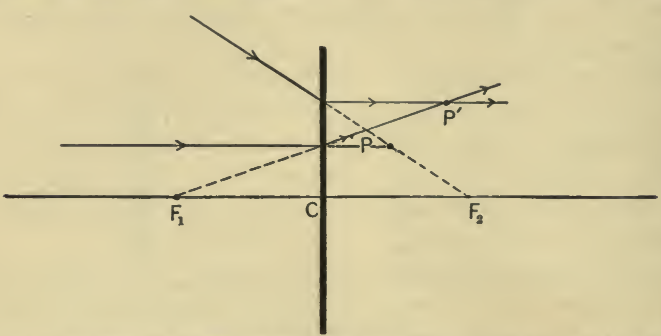


Fig. 234.—Formation of an image by a diverging lens. The "virtual source" P has a real image P', provided P is between the lens and the focal plane.

two rays pointed toward it, one parallel to the axis, the other through the centre of the lens; it is seen by geometry that the emerging rays are parallel, for the triangles (F_1CA), (CAP), (CF_2P) are all equal. Conversely, rays which are parallel to each other and inclined slightly to the axis diverge,

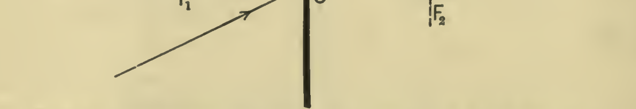


Fig. 235.—Special case: the virtual point object P lies in the focal plane.

after emerging from the lens, as if they proceeded from that point in the focal plane on the incident side where a line through the centre of the lens parallel to the rays meets the plane.

This furnishes us with a method for the construction of the emerging portion of any ray. Let $\overline{PQ}$ be a ray meeting the lens at Q; draw through C a line $\overline{CB}$ parallel to $\overline{PQ}$, and intersecting the focal plane at F_1 in A; draw through A the line $\overline{AQS}$; $\overline{QS}$ is the continuation of the incident ray. For, if there were two parallel rays $\overline{PQ}$ and $\overline{BC}$, they would diverge, on emerging, as if they came from A.

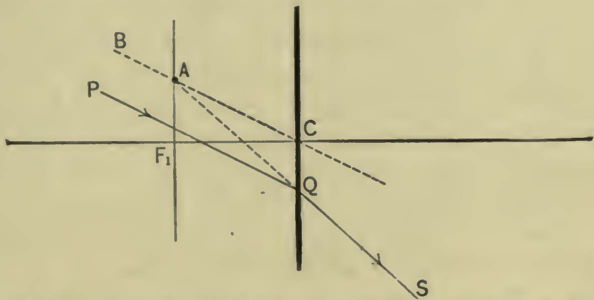


FIG. 236. — Construction for the refraction of any ray PQ by a diverging lens.

Spherical Aberration, etc. — It may not be useless to state again the assumptions made in the above treatment of lenses: the lens is supposed to be thin; the object must be small and close to the axis; the pencils must be normal; the lens must have surfaces whose radii are large in comparison with its dimensions. If any of these conditions are violated, the laws cease to hold. Reference should be made to some special treatise such as Lummer, *Photographic Optics*, for a full discussion of the general subject. *We have also assumed throughout that the waves were homogeneous, for* n *has been treated as a constant.*

Resolving Power of Lenses. — It has already been explained (see page 436) that when the waves from a point source fall upon a converging lens, the image of this source is the point to which by far the greater amount of the energy is brought, but that owing to diffraction energy is relieved by other points also. In the case of a lens with a circular edge — such as most lenses have — the diffraction pattern consists of a bright area, whose brightest point is the geometrical image, as just explained, and this is surrounded by rings alternately dark and bright, if the waves are homogeneous; that is, the light gradually fades, then increases gradually, etc. So, if there are two point sources close together, the line joining which is perpendicular to the axis, there will be two diffraction patterns, which will overlap; and the resultant effect is due to their superposition. It is evident that if the two point sources are so close together that the centres of their diffraction patterns almost coincide, it may be impossible to see these centres as *separate* bright points; but, if the centre of one pattern coincides with the first dark ring of the other, then the existence of two bright points may be recognized. There is thus a limiting value of the nearness of two point sources which can be perceived as such by the use of the lens. This quantity may be deduced by considering a simple case.

Let ACB be the cross section of a converging lens by a plane through the axis, and let O' be the image of O, a point on the axis. O' is the image of O because the disturbances along the different rays from the latter reach the former in the same time, and so are in the same phase of vibration; but, if a point near O' is considered, the different disturbances from O reach there in different phases, because they pass over different "optical" paths; and it may happen that the difference in path is such that owing to the disturbances arriving there in *opposite* phases they annul each other's action. If this is the case, this point is on a dark ring. If there is another point source at P, where OP is perpendicular to the axis, the disturbance produced by its waves at O'

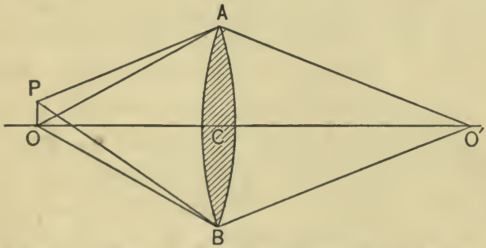


FIG. 237. — Diagram to illustrate the resolving power of a lens.

may be calculated in a similar manner. Draw $\overline{PA}$ and $\overline{AO'}$, $\overline{PB}$ and $\overline{BO'}$; these are the two lines from P to O' whose difference in length is the greatest. If this difference amounts to a whole wave length of the waves, it may be assumed that the resultant action at O' due to the disturbances from P which pass through one half the lens *differs in phase by half a wave length* from that due to those which pass through the other half of the lens; and therefore O' will be on the first dark ring of the diffraction pattern due to P. From what was said above, when this is the case, P is as close to O as it is possible for it to be and yet to be seen separate from it.

This condition may be expressed in an equation:

$$(\overline{PB}+\overline{BO'})-(\overline{PA}+\overline{AO'})=l,$$

where l is the wave length.

But O' is the image of O; and, therefore, drawing $\overline{OA}$ and $\overline{OB}$, we have, since the optical lengths from O to O' are equal,

$$(\overline{OB}+\overline{BO'})=(\overline{OA}+\overline{AO'}).$$

Subtracting this equation from the previous one, we have

$$(\overline{PB}-\overline{OB})+(\overline{OA}-\overline{PA})=l.$$

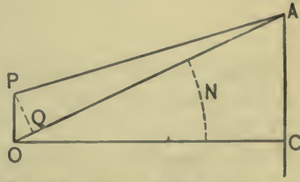


FIG. 238.—Portion of Fig. 237 greatly magnified.

These quantities may be expressed in a simpler form. Drawing the diagram on a larger scale, and dropping a perpendicular $\overline{PQ}$ from P upon $\overline{OA}$, it is seen by similar triangles that

$$\frac{OQ}{OP}=\frac{CA}{OA}, \text{ or } \overline{OQ}=\overline{OP}\,\frac{CA}{OA}.$$

But since $\overline{OP}$ is *extremely* small,

$$\overline{OQ}=\overline{OA}-\overline{PA}.$$

It may be shown in a similar manner that

$$\overline{PB}-\overline{OB}=\overline{OP}\,\frac{CA}{OA}.$$

Hence, substituting in the above formula, it follows that

$$2\,\overline{OP}\,\frac{CA}{OA}=l, \text{ or } \overline{OP}=l\,\frac{OA}{2\,CA}=l\,\frac{OA}{AB}.$$

This marks the minimum value of $\overline{OP}$ which allows us to recognize the *two* independent point sources O and P. The reciprocal of this, $\frac{1}{OP}$, or $\frac{2}{l}\,\frac{CA}{OA}$ or $\frac{1}{l}\,\frac{AB}{OA}$, is called the "resolving power" of the lens.

There are two cases of special interest: one when we are considering two luminous points of an object near at hand; the other when the two sources are at a great distance; one deals with the power of a microscope to "separate" fine details of a structure; the other, with the power of a telescope to separate the two components of a "double star" or to recognize the details of the surface of the moon or sun. The same formula applies to both, but it may be put in more convenient form.

1. Microscope.

Call the angle (AOC), N; then $\frac{\overline{CA}}{OA}=\sin N$; hence

$$\frac{1}{OP}=\frac{2\sin N}{l}.$$

Therefore, the larger the angle subtended at the object O by the lens, so much the greater is the resolving power, and therefore the lens should have a short focus, so that the point O may be close to the lens and yet may have a real image. Further, the shorter the wave length the greater this resolving power. Thus, fine details may be seen with a microscope if the object is illuminated with blue light, which are not distinguishable in red light.

2. Telescope. Since in this case the object is at a great distance, $\overline{OA}$ equals $\overline{OC}$, practically; and $\frac{OP}{OC}$ equals the angle subtended at the centre of the lens by the distant object. Call this angle M; then from the formula $OP=l\,\frac{OA}{AB}$ it is seen that $M=\frac{l}{AB}$. This is the smallest angle which two objects can subtend and yet be seen as separate points of light in a telescope. $\frac{1}{M}=\frac{AB}{l}$ is called the "angular resolving power." $\overline{AB}$ is the diameter of the lens; and we see the advantage then of using telescopes with large lenses (or mirrors) if one wishes to separate double stars, etc.

(This elementary treatment of the resolving power of a lens is due to Lord Rayleigh. A rigid treatment leads, as he and others have shown, to the formula $\overline{OP} = 1.22\, l \frac{\overline{OA}}{AB}$.)

In using high-power microscopes, it is customary to fill the space between the lens and the object with a liquid which has a large index of refraction — oil of cedar wood is generally used. (This oil has nearly the same index of refraction as the glass, so no light is lost by *reflection* at the surface of the glass; and, further, spherical aberration is avoided to a great extent.) Then the formula as deduced above must be modified, because the optical paths of the rays from P and O include portions in a medium whose index of refraction, n, is different from that of air. Thus,

$$(n\overline{PB} + \overline{BO'}) - (n\overline{PA} + \overline{AO'}) = l,$$

$$(n\overline{OB} + \overline{BO'}) - (n\overline{OA} + \overline{AO'}) = 0.$$

Hence, $$n\,(\overline{PB} - \overline{OB}) + n\,(\overline{OA} - \overline{PA}) = l;$$

or, $$2\,n\overline{OP}\,\frac{CA}{OA} = l;$$

and finally, $$\overline{OP} = l\,\frac{\overline{OA}}{2\,nCA} = \frac{l}{2\,n\sin N}.$$

Hence, $$\frac{1}{OP} = \frac{2\,n\sin N}{l};$$

so the resolving power of the microscope is increased in the ratio of $n : 1$. ($n \sin N$ is called the "numerical aperture" of the lens.) The greatest value N can have is $\frac{\pi}{2}$, for which the sine is 1; so the least possible value of $\overline{OP}$ is $\frac{l}{2\,n}$, or, very approximately, one half the wave length of the light used. No combination of lenses can therefore enable one to see separately two independent luminous points which are closer than this together.

Combination of Two Thin Lenses. — Let the two lenses be at a distance h apart; and let their focal lengths be f_1, f_2. The formulæ for the two lenses are:

$$\frac{1}{u_1} + \frac{1}{v_1} = \frac{1}{f_1}, \text{ and } \frac{1}{u_2} + \frac{1}{v_2} = \frac{1}{f_2},$$

where $u_2 = h - v_1$. To find the focal length of the combination on the right-hand side, with reference to Fig. 239, put $u_1 = \infty$. Hence, $v_1 = f_1$, $u_2 = h - f_1$; and

$$\frac{1}{h-f_1} + \frac{1}{v_2} = \frac{1}{f_2},$$

or,
$$\frac{1}{v_2} = \frac{1}{f_2} - \frac{1}{h-f_1} = \frac{h-f_1-f_2}{f_2(h-f_1)}.$$

Therefore, the focal length $v_2 = \dfrac{f_2(h-f_1)}{h-f_1-f_2}$.

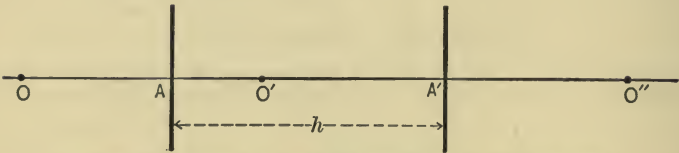


FIG. 239.—Formation of an image by a combination of two thin lenses.

A special case of this is when the two lenses are placed close together, *i.e.* when $h = 0$. Then the focal length equals $\dfrac{f_1 f_2}{f_1 + f_2}$; or, calling it f, $\dfrac{1}{f} = \dfrac{1}{f_1} + \dfrac{1}{f_2}$.

The reciprocal of the focal length of a lens (or of a combination of lenses) is called its "power." The unit of power adopted by opticians is that of a lens whose focal length is one *metre*; it is called a "diopter." To find the power of any lens, then, in diopters, its focal length in metres must be measured, and its reciprocal taken. A converging lens is called positive.

Thick Lens.—If the lens is of such a thickness that it cannot be assumed to be "thin," in the sense in which this word has been used above, the solution of the problem of refraction can be obtained, exactly as in the case of a thin lens, by considering the refraction of a pencil of rays at the two surfaces. But the distance from the first surface to the image produced by it no longer equals its distance from the second surface; that is, v_1 does not equal u_2 numerically (see page 471). If t is the thickness of the lens, $u_2 = -(v_1 + t)$; and, if this value is substituted in the equations, the final formula connecting the positions of the object and image may be deduced.

Chromatic Aberration. — Attention has been called repeatedly to the fact that the index of refraction of a substance is different for trains of waves of different wave number, showing that the velocity of these different waves in the substance is different. (It should be remembered that the *wave number* of a train of waves does not change as it passes from one medium into another, *e.g.* air into glass, for the number of "waves" reaching any point in a given interval of time must equal the number that leaves it. But, calling N this wave number, and V_1, l_1, V_2, l_2, the velocities and wave lengths in the two media, $V_1 = Nl_1$, $V_2 = Nl_2$; and since V_1 is different from V_2, l_1 is different from l_2. So, as the waves pass from one medium into another, their wave length changes.) In the case of ordinary dispersion the shorter waves are refracted more than the long ones; or, in other words, the index of refraction varies in an inverse manner from the wave length. Methods for the study of the connection between these two quantities will be described in a later chapter.

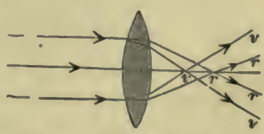


FIG. 240. — Chromatic aberration.

This fact that n varies with the wave length is of great importance in dealing with the theory of lenses; for this quantity enters into the two fundamental formulæ:

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f} = (n-1)\left(\frac{1}{r_1} + \frac{1}{r_2}\right)$$

and Linear Magnification $= \dfrac{v}{u}$.

Therefore, if waves of different wave length are used to illuminate a given object, or if the object itself emits such waves, not alone will the corresponding images be at different distances from the lens, but they will also be of different sizes. Since light waves of different wave length correspond to different colors, this phenomenon is called "chromatic aberration." For ordinary substances n is, as has just been said, greater for blue light than for green, etc.; and so, if there are two trains of waves corresponding to these colors, f is less for the blue ones than for the green, and consequently the principal focus for the former is nearer the lens, and the magnification is less. If the object emits white light, there will be a series of colored images of different sizes. This fact is, naturally, most detrimental to the proper use of an optical instrument, for sharp clearly defined images are desired. As will be shown in the next paragraph, it is possible by a combination of two lenses of different material to remedy this defect to a certain extent by causing any two definite trains of waves of different wave length to be brought to the same focus, but waves of all wave lengths will not be similarly affected. So, when white light is used, if two of its components are thus brought to the same focus, the other components will have different foci — they form what is called a "secondary spectrum." Thus chromatic aberration can be corrected only partially. The choice of the two trains of waves which shall be brought to the same focus is arbitrary; if the instrument is to be used for visual purposes, two trains are chosen which affect the eye most intensely; while, if it is to be used for photographic work, the two trains of waves chosen are those which act most intensely upon a photographic plate.

Achromatism. — Since a lens can be considered as made up of prisms, we shall first show how one prism may be so chosen as to neutralize the dispersive action of another prism for two given trains of waves. In speaking of the effect of a prism, it was shown that the deviation produced depended upon the material and angle of the prism, the wave number of the waves, and the angle of incidence. The difference in the deviations for two different trains of waves is called the "dispersion" of those waves; and by suitably choosing the material and angle of the prisms and the angles of incidence, it is evidently possible to secure the same dispersion for two definite trains of waves. When this is done, the dispersion of any two other trains is as a rule different; for owing to the difference in the material of the prisms there is no connection

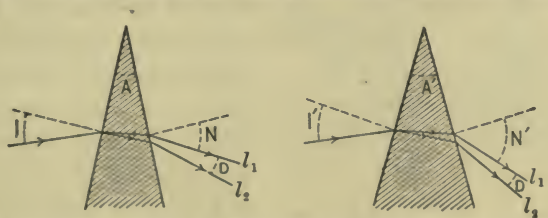


FIG. 241.—Two different prisms of different materials may produce the same dispersion of two rays l_1 and l_2.

between their dispersions in different parts of the spectrum. Thus suppose that, when white light is incident upon two prisms at angles I and I', as shown in the cut, a ray is dispersed in such a manner that two of its components, of wave length l_1 and l_2, have the same dispersion D. Let the angle between the emerging ray l_1 and the normal to the second face of the prism be N and N' in the two prisms. Then if a beam of parallel rays of wave length l_1 are incident at an angle N' upon the second face of the second prism, they will emerge, along the direction of the incident ray in the cut, *i.e.* making the angle I' with the normal; and, if a beam of parallel rays of wave length l_2 is incident upon the same face at an angle $(N' + D)$, they will emerge in the same direction as the former rays. This condition may be secured by inverting the first prism and placing it with its edge parallel to that of the other prism, but so inclined that the angle of incidence upon it of the ray l_1 from the second prism is N; for, as is seen from geometry, the angle of incidence of the ray l_2 is, under these conditions, $(N + D)$. Therefore a ray incident upon the second prism at the angle I' is dispersed by it; and two of its components, l_1 and l_2, fall upon the first prism, are refracted by it, and emerge parallel to each other, making an angle I with the normal to the last face of the prism. Therefore a beam of parallel rays of white light incident upon the second prism at an angle I' will be dispersed by the two prisms; but all the rays of wave length l_1 and l_2 will emerge parallel to each other in the direction defined by I. In practice it is found that it is possible to choose the prisms so that, when combined as described above, their adjacent faces may be parallel; but owing to the differences in the material and angles of the prisms the directions of the incident and emerging parallel rays are not necessarily the same. Therefore this double prism still *deviates* the two rays l_1 and l_2, but does not *disperse* them. Such a prism is called an "achromatic" one; or, either one of its component prisms is said to be "achromatized." (It is evident also that two prisms may be compounded which will not deviate one particular train of waves, but will deviate the others, thus producing dispersion.)

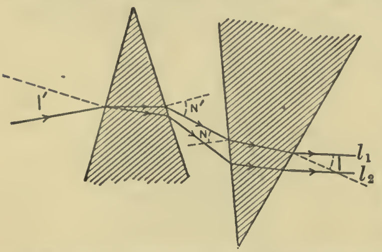


Fig. 242. — One prism may neutralize the dispersion of two rays produced by another prism.

It follows at once that, by combining two lenses of different materials, one diverging, the other converging, a double lens may be secured which will deviate two rays of definite wave length, but not disperse them. Such a lens is shown in the cut. The converging lens is as a rule made of "crown" glass; the diverging one, of "flint" glass. It is called an achromatic lens for the two definite trains of waves.

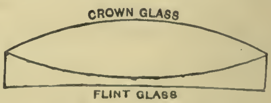


Fig. 243. — Achromatic lens.

If the component lenses are thin and have focal lengths f_1 and f_2, the focal length of the combination is given by $\frac{1}{f}=\frac{1}{f_1}+\frac{1}{f_2}$. Therefore, we can express the facts in regard to the achromatic combination by saying that f has the same value for the different values of n corresponding to l_1 and l_2. In the case of this combination of *thin* lenses, not alone are the focal lengths the same for the two trains of waves, but also the magnifications. This is not true if the combined lenses are thick. If a system of several lenses — not in contact with each other — is to be *completely* achromatic for two definite trains of waves, each of its component lenses must be achromatic also.

The first achromatic lens was made by John Dollond, a London optician, in 1758. Previous to this, as early as 1729, a method for achromatizing lenses had been discovered by Chester More Hall of Essex, England; and he constructed several such lenses, but never published any account of his work. Newton and others had maintained that if one prism corrected the *dispersion* of another, it would at the same time annul the *deviation*, and that therefore it would be impossible to make an achromatic lens. But when different kinds of glass were tested, it was found that the generalization drawn by Newton from his experiments on prisms of glass and water was not correct.

Achromatic Combinations. — It may be proved that if two thin lenses of focal lengths f_1 and f_2 of the same kind of glass are placed at a distance h apart, where $h=\frac{f_1+f_2}{2}$, the combination has the same focal lengths for waves of all lengths. The distance of the principal focus from the end lens of focal length f_2 is, as we have seen, $\frac{f_2(h-f_1)}{h-f_1-f_2}$ and the distance of the other principal focus from the other end is

$$\frac{f_1(h-f_2)}{h-f_1-f_2}.$$

There are two special combinations which are of interest because of their use as eyepieces in microscopes or telescopes.

1. *Ramsden eyepiece.* — In this there are two plano-convex lenses of the same kind of glass and of the same focal length placed with their convex faces toward each other. The condition for achromatism is that

their distance apart $h=\frac{f_1+f_2}{2}=f_1$, since $f_1=f_2$. Therefore, any scratches or dust particles on the surface of either lens would be seen if one were to look through the other with the eye accommodated for an infinitely distant object, as is the case when one is using a telescope. Therefore it is better to make the distance of the lenses apart slightly less, although by so doing the chromatic aberration is not exactly corrected. Thus, if $h=\frac{2}{3}f_1$, the distance of the principal focus from either end is

$$\frac{f_1(\frac{2}{3}f_1-f_1)}{\frac{2}{3}f_1-f_1-f_1}=\frac{f_1}{4}.$$

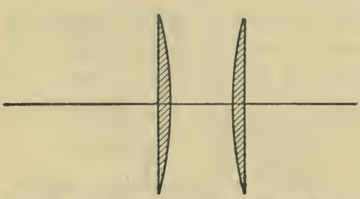


Fig. 244. — A Ramsden eyepiece: the focal lengths of the two lenses are equal.

In this case, any object placed at this distance from one of the lenses will be seen on looking through the other lens at apparently an infinite distance. In other words, this eyepiece has the focal properties of an ordinary converging lens of this focal length, but it is approximately achromatic.

2. *Huygens eyepiece.* — In this there are two plano-convex lenses, placed as shown in the cut, with focal lengths f_1 and f_2, where $f_1=3f_2$. In order to secure achromatism their distance apart should satisfy the equation $h=\frac{f_1+f_2}{2}=2f_2$. The distance of the principal focus from the first lens is given by $\frac{f_1(2f_2-f_2)}{2f_2-f_1-f_2}=-\frac{3f_2}{2}=-\frac{f_2}{2}$. This means that, if a ray is incident upon the first lens at such an angle as to be pointed toward a point *between the lenses* at a distance $\frac{f_1}{2}$ from the first lens, it will emerge from the second lens parallel to the axis. Therefore, since the object must be *virtual* in order to have an image at infinity, this eyepiece cannot be used, like a converging lens, to magnify ordinary objects.

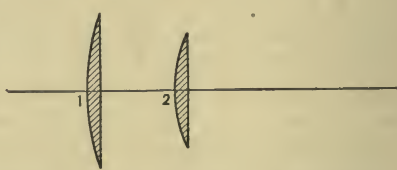


Fig. 245. — A Huygens eyepiece: the focal length of the first lens is three times that of the second.

CHAPTER XXIX

OPTICAL INSTRUMENTS

Optical instruments, so called, are pieces of apparatus so designed as to make use of light waves in order to form images of luminous objects. In general they consist of combinations of mirrors, lenses, and prisms; but some instruments do not have such parts. The number of these instruments is great, but in this chapter only the simplest and most generally important ones will be described.

Optical instruments may be divided roughly into three classes: "pin-hole," reflecting, and refracting. In the first, no mirror or lens is used; in the second, some form of mirror is the essential feature; in the last, some prism or lens, or a combination of prisms or lenses.

"Pin-hole" Instruments

These have been described in Chapter XXV, and nothing further need be said here. They all depend upon the "rectilinear propagation" of light. The pin-hole camera and the "camera obscura" are the only instruments of this class that are of importance.

Reflecting Instruments

The action of plane, spherical, and parabolic mirrors was explained in Chapter XXVII; and their use as looking glasses, in search lights, etc., was described. The great advantage of all forms of reflecting instruments over refracting ones is that they are necessarily free from chromatic aberration. Spherical aberration, however, can be avoided only by using normal pencils of light and mirrors with small curvatures.

One important application of a concave spherical mirror was first made by Newton, namely, to form a "telescope." When we look at a distant object, it subtends at the eye a comparatively small angle. Thus, if A and B are two points on opposite edges of the object, and if O marks the position of the eye, the *linear* angle subtended by the object is (AOB). We estimate the distance of an object from us by the angle which it subtends at our eyes; the nearer it is, the larger is this angle. The purpose of a telescope is to bring a distant object apparently nearer, by changing the direction of a ray $\overline{AO}$ to $\overline{A'O}$, and a ray $\overline{BO}$ to $\overline{B'O}$; so that the angle subtended by the rays from A and B is now $(A'OB')$, an angle greater than (AOB). The "power of the telescope" is defined to be the ratio of the angles $(A'OB')$ and (AOB).

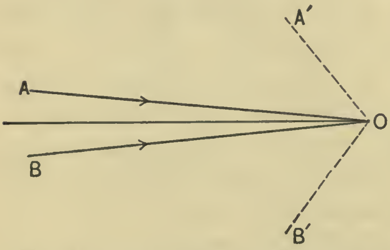


FIG. 246. — Diagram representing the effect of a telescope in increasing the angle between two rays $\overline{AO}$ and $\overline{BO}$.

In all telescopes either a concave mirror or a converging lens converges the rays so as to form a real image; and this is viewed by a lens, or a combination of lenses, called the "eyepiece." The simplest form of eyepiece is a converging lens, but a Ramsden or Huygens eyepiece is ordinarily used. In doing this the eyepiece is so placed that the image to be viewed comes at its principal focus (or just inside it), and it forms a virtual magnified image at apparently an infinite distance.

Reflecting Telescope. — In all forms of reflecting telescopes a concave spherical mirror is used to receive the rays from the distant object. The mirror is mounted in some framework or tube, so that it may be turned to point in any direction. In the cut, let C be the centre of this spherical surface, and $\overline{P_1C}$ and $\overline{PC}$ be lines drawn from two opposite points in the edge of the remote object. All rays parallel to $\overline{P_1C}$ are brought to a focus at a point F_1, on $\overline{P_1C}$, halfway between C and the mirror; and all rays parallel to PC will be brought to a focus at a point F, on PC, halfway between C and the mirror (see page 449). Therefore, there will be a real image of the object in the plane $\overline{FF_1}$.

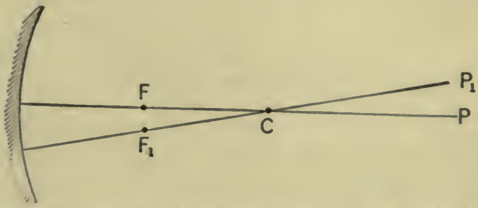


Fig. 247. — Diagram representing a reflecting telescope. P and P_1 are two points at an infinite distance.

In order to see the magnified image of this image produced by the eyepiece, some plan must be adopted which will render it unnecessary for the observer to stand in front of the mirror and so obstruct the view. Several have been tried: 1, the mirror is tipped slightly so that its centre lies outside the telescope tube, and the principal focus for parallel rays lies near the edge of this tube, and thus the image may be viewed — this arrangement was used by Herschel; 2, a small plane mirror (or a totally reflecting prism) is placed close to the principal focus, but between it and the mirror, and is so inclined as to turn the rays off sidewise through an opening in the side of the telescope tube; thus the image is formed outside or near the edge of the telescope — this was Newton's plan; 3, a small *convex* mirror is used in place of the plane mirror, and is so turned as to reflect the rays back through a small opening at the centre of the concave mirror and form an image at this opening — this arrangement is due to Cassegrain; 4, or, finally, a small *concave* mirror is placed beyond the principal focus of the large mirror, and is turned so as to form an image in an opening in the large concave mirror (like Cassegrain's method) — this plan was proposed by Gregory.

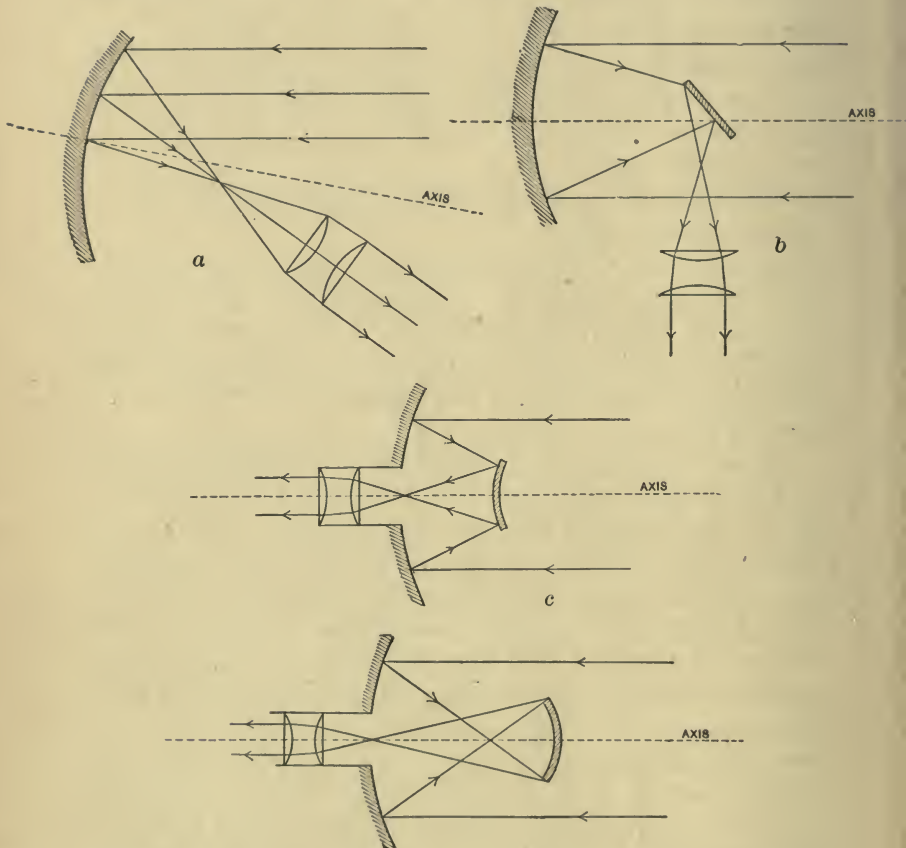


Fig. 248. — Different forms of reflecting telescopes: (*a*) Herschel; (*b*) Newton; (*c*) Cassegrain; (*d*) Gregory.

The power of the telescope is found directly from simple geometrical considerations. As shown in Fig. 249, parallel rays from the distant point *P* are brought to a focus at *F*.

halfway between C and the mirror; parallel rays from the distant point P_1 are brought to a focus at F_1. The straight line at A represents the eyepiece, which is placed so that F and F_1 lie in its focal plane; B is a point so chosen that $\overline{AB} = \overline{FA}$, and therefore F and B are the principal foci of the eyepiece. The rays from P_1, after reflection, converge to F_1, pass on through the eyepiece, and emerge as if they came from the virtual image of F_1, at an infinite distance in the direction $\overline{F_1A}$. Similarly, the reflected rays passing through F have a virtual image at an infinite distance on the line $\overline{FA}$.

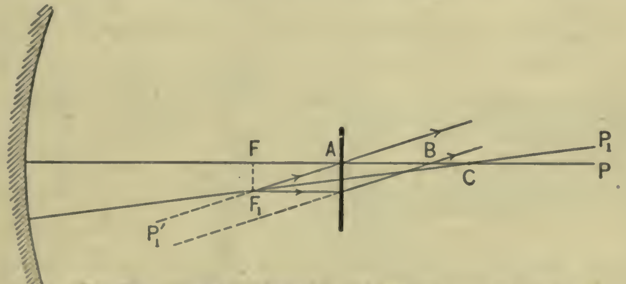


FIG. 249. — Diagram of formation of image by a concave mirror and eyepiece A.

Therefore, the angle subtended by the distant object itself is (P_1CP); while that subtended by the virtual image of this object formed by the eyepiece is (FAF_1). These angles may be compared by considering their tangents. The angle (P_1CP) equals (FCF_1), and its tangent equals $\frac{FF_1}{FC}$. The tangent of (FAF_1) equals $\frac{FF_1}{FA}$. These angles are, however, sufficiently small in general to allow us to replace their numerical values by those of their tangents; so the ratio of the angles (FAF_1) and (P_1CP) is $\frac{FC}{FA}$, that is, it is the ratio of the focal length of the mirror to that of the eyepiece. We see, therefore, the great advantage of having a telescope whose concave mirror has a long focal length.

The limit to the resolving power of a telescope is determined, as we have seen on page 487, by the diameter of the mirror; the larger it is, so much the greater is this power. A large mirror also gathers more light, and therefore enables fainter objects to be seen.

Newton constructed his first reflecting telescope in 1668. It had a diameter of one inch, and magnified thirty or forty times. He later made a larger instrument. Previous to this, such instruments had been designed; and the invention is attributed to Niccolo Zucchi (1586–1670) of Rome.

Refracting Instruments

All instruments containing lenses are subject to a certain amount of chromatic aberration, although this may be minimized by using only achromatic lenses and combinations. There is also always some spherical aberration; but this may be avoided largely by using only normal pencils of light, and exposing only the central portions of the lenses. The number of the defects which are possible with a lens is so great that some special treatise on the subject should be consulted. Only a few of the simplest refracting instruments will be described here.

Photographic Camera. — In this instrument a real image of an object, more or less remote, is formed on a plane photographic plate. This image is produced by a converging lens or system of lenses. If the camera is used for taking photographs of landscapes, the lens is ordinarily a single achromatic one, and a diaphragm with a circular opening of variable size is placed in front of it. If it is used for photographing buildings, special pains must be taken to avoid spherical aberration and the consequent distortion; two achromatic lenses symmetrical with reference to a plane halfway between them are used, the diaphragm being in this plane — this is called a "rectilinear" or "orthoscopic" lens. For portrait work, lenses of large diameter must be used in order to secure as much light as possible.

When photographs of distant objects are taken, the images are, in general, small; but by combining a diverging lens with the converging system, a larger image may be secured. Such a compound lens system is called a "teleobjective."

Projection Lantern. — This instrument consists of a lamp, or a strong source of light, which by means of lenses illuminates a drawing or photograph on glass, or some object which is transparent in parts, and of a converging system of lenses which throws a real image of this illuminated object upon a suitable screen. The lens system between the light and the object — called the "condenser" — consists, in general, of two plano-convex lenses with their curved surfaces in contact. Its function is to deflect down upon the object to be "projected" as much light as possible, so as to render it strongly luminous. Then by means of the "focusing lens" a real image of it is formed upon the screen. The focusing and condensing lenses must be achromatic, and the former must be corrected for spherical aberration also.

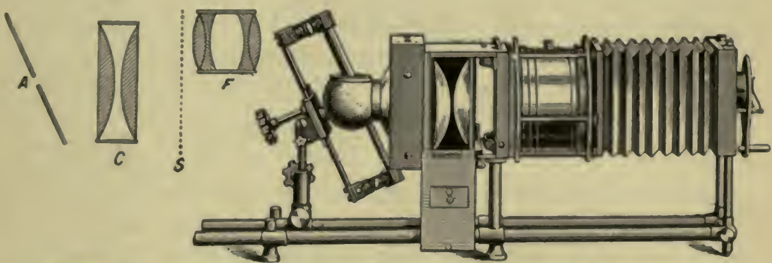


Fig. 250. — Projection lantern; (A) arc light; (C) condenser; (S) slide to be illuminated; (F) focusing lens.

Astronomical Telescope. — This instrument consists of a converging achromatic lens, called the "object glass," which forms a real image of the distant object in its focal plane, and this is viewed by an eyepiece. (In the cut a Ramsden eyepiece is represented.) The eyepiece is so placed that the image formed by the object glass comes at or just inside its principal focus, so the object is seen apparently at infinity. The power of this telescope is, therefore, like that of the

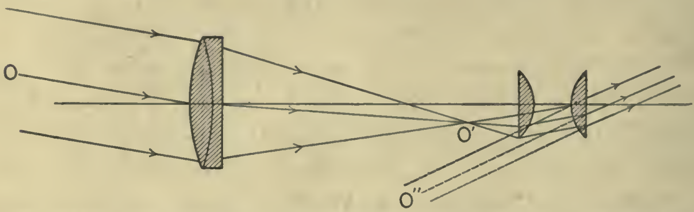


Fig. 251.—Astronomical telescope: O is source at an infinite distance; O' is image formed by object glass; O'' is virtual image formed at an infinite distance by the eyepiece.

reflecting instrument (see page 499), equal to the ratio of the focal length of the object glass to that of the eyepiece.

The resolving power of a telescope is, as we have seen, determined by the size of the object glass, as is also the quantity of light received. (See page 500.)

The astronomical telescope forms, as just explained, a virtual image of the distant object which is *inverted;* that is, if the object is a tree, the image will have the tree pointed down instead of up. This is a disadvantage if the instrument is to be used for purposes that are not purely astronomical.

Dutch Telescope (or Galileo's telescope). — This telescope is free from the disadvantage of the astronomical instrument

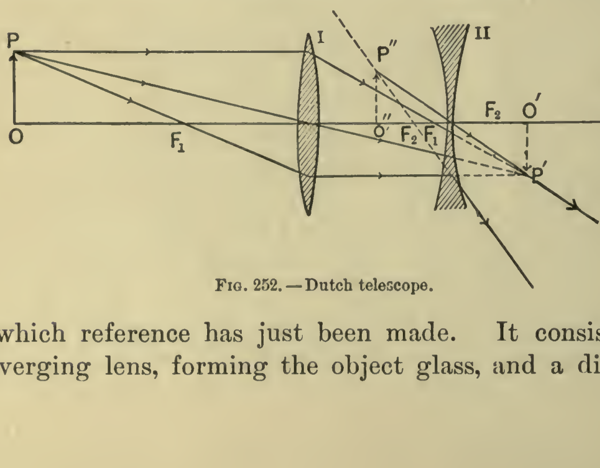


Fig. 252.—Dutch telescope.

to which reference has just been made. It consists of a converging lens, forming the object glass, and a diverging lens so placed that the principal focus on the side next the observer coincides with the principal focus of the object glass. Thus, parallel rays from a point of the distant object are converged by the object glass toward a point behind the diverging lens, which is in the focal plane of both lenses; these rays are diverged again by the second lens and emerge parallel to a line drawn from the centre of the second lens to the point in the focal plane toward which the rays were converging. The image of the distant object is therefore a virtual one at an infinite distance; but it is *erect;* that is, the image formed of a distant tree represents the tree in an upright position.

The first telescope was probably constructed by Hans Lippershey of Middleburg in the Netherlands in 1608; and Galileo, upon hearing of the invention, but without knowing any details of the construction, made an exactly similar instrument in 1609. Kepler was the first to suggest the use of a convex lens for the eyepiece. Galileo immediately used his telescope for observing the heavenly bodies, and made many most important discoveries.

Previous to the construction of the reflecting telescope by Newton and the invention of the achromatic lens by Dollond, the only means of minimizing the color effects produced by telescopes was to use lenses of great focal length, which were clumsy and offered many disadvantages. Huygens presented to the Royal Society of London a lens whose focal length was 123 ft.

Microscope.—This instrument is one designed to "magnify" a small object, that is, to increase the apparent distance apart of any two of its points which are close together.

a. Simple microscope. — As was said on page 494, in speaking of eyepieces, a single converging lens or a Ramsden eyepiece can be used as a microscope, the object being placed just inside the principal focus. There is thus formed an erect virtual image of the object (see page 477). The distance at which this image is formed is arbitrary; but it is generally chosen as about 25 cm. (or 10 in.) from the eye, because this is the distance at which people with normal eyes hold an object in order to see it most distinctly.

b. Compound microscope. — The magnification secured by a single lens is not great; and it is in general combined with another lens system, as shown in Fig. 253. The latter forms a real magnified image of the object, which is viewed and again magnified by the eyepiece. (In this instrument a Huygens eyepiece may be used, as is illustrated in the cut.) The lens system which is nearest the object is called the

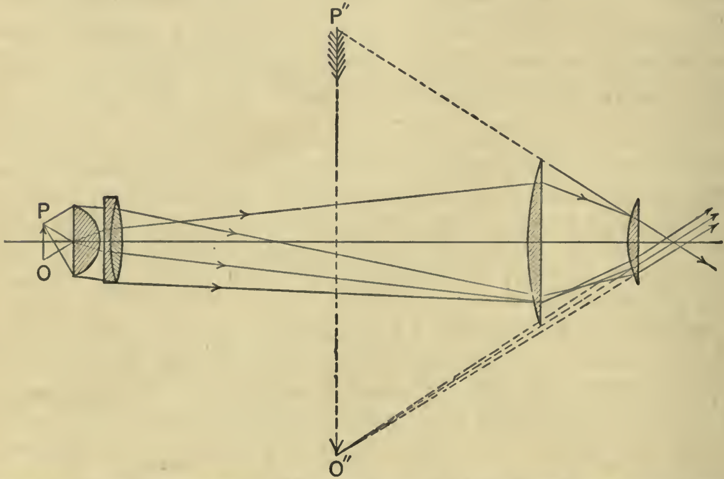


Fig. 253. — Compound microscope: $\overline{OP}$ is the object; $O''P''$, the virtual image formed by the eyepiece.

"objective." It consists of several lenses so chosen as to give an image as free as possible from chromatic and spherical aberration; and it is so constructed as to give as much light as possible to the image and at the same time to have a large value of $n \sin N$ (see page 487), so that the resolving power is large. The magnifying power can be deduced by simple geometrical methods.

The use of converging lenses as simple magnifiers was known to the ancients; but the compound microscope was probably invented by Zacharias Joannides of Middleburg in the Netherlands and his father, some years before 1610. It was invented independently by others also, among whom was Galileo. All the early microscopes had a concave lens for the eyepiece; and Franciscus Fontana of Naples was the first to suggest the substitution of a convex lens.

Spectrometer. — A spectrometer is an instrument primarily designed to measure the angle of deviation produced in the direction of a beam of light by reflection or refraction. It consists essentially of four parts. There is a substantial metal base to which is rigidly connected an upright metal cylinder, whose axis is called the "axis of the instrument." To this is attached a metal plate whose edge is divided into degrees, minutes, etc. Two bent metal arms are also attached to this cylinder by means of collars, so that they can turn around it as an axle; they carry metal tubes which lie in a plane perpendicular to the axis of the instrument and are thus movable in this plane. One of these tubes is a simple form of astronomical telescope; the other carries at one end an achromatic converging lens whose focal length is that of the tube, and at the other end it is closed by a metal cap in which there is a fine slit with straight parallel edges, which can be opened or closed — this is called a "collimator." If the slit is illuminated, it serves as a source of diverging waves which will emerge from the lens of the collimator in the form of plane waves.

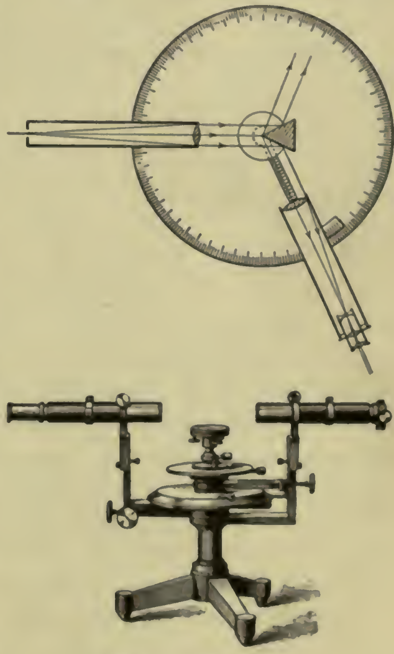


Fig. 254. — Spectrometer.

The telescope and collimator are turned until their axes intersect at the axis of the instrument; and the object which is to produce the deviation of the light — a plane mirror or prism — is placed on a platform in the middle of the metal plate referred to above, having the normals to its reflecting or refracting faces in the plane of the axes of the collimator and telescope. If now by means of the former a beam of parallel rays falls upon the mirror or prism, they will be deviated, and their new direction may be found by turning the arm carrying the telescope until there is formed in the centre of the field of view an image of the illuminated slit. In order to determine this condition exactly, it is customary to insert in the focal plane of the object glass a metal ring across which are stretched two *fine* silk fibres or spider lines, which are called the "cross hairs." These are made to cross exactly at the centre of the tube. Since they are at the focal plane of the object glass, they will be seen through the eyepiece at the same apparent distance as the object which sends the parallel rays into the object glass. The positions of the collimator and telescope may be noted by means of the divided scale on the edge of the plate, if suitable pointers or verniers are attached to them.

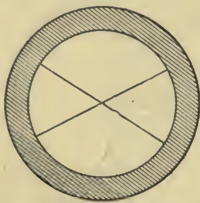


FIG. 255. — Cross hairs used in telescope.

With this instrument the laws of reflection and refraction may be verified; the angle of a prism and the angle of minimum deviation produced by it for any train of waves may be measured; etc. Therefore, the index of refraction of any substance which can be made in a prism may be determined; for $n = \frac{\sin \frac{A+D}{2}}{\sin \frac{A}{2}}$. (See Ames and Bliss, *Manual of Experiments in Physics*, pages 459–475.)

If the laws of reflection are to be studied, ordinary white light may be used to illuminate the slit; but, if the phenomena of refraction are to be observed, special precautions must be taken which will be described in the next chapter.

Effect of Diaphragms. — One fact that must be taken into account in the description of all reflecting or refracting instruments is that the wave front of the waves is always limited by certain apertures or diaphragms. Thus, in the case of a telescope, the only portion of the wave front that enters the instrument has the size of the object glass. Again, if there are diaphragms in the tubes, they may limit the cone of rays which proceed from any point of the object and enter the instrument. Further, the only rays that are of practical use in the case of an instrument that is used visually are those which have such a direction as they leave the eyepiece as to enter the pupil of the eye. The effect of the presence of these various circular openings is felt in many ways. The brightness of the image, the resolving power of the instrument, the contrast of the background, the spherical aberration, etc., all depend upon their size and position. For details in regard to the matter, reference should be made to some advanced text-book, such as Drude, *Optics*, or Lummer, *Photographic Optics*.

CHAPTER XXX

DISPERSION

In speaking of chromatic aberration, achromatic lenses, etc., it was noted that the index of refraction of a given substance varies with the wave length of the light which suffers refraction, and that, in general, the shorter the wave length, the greater is the index of refraction. In this chapter the method of determining the connection between these two quantities will be discussed.

Pure Spectrum. — Before any relation between index of refraction and wave length can be established, it is necessary to devise a method for securing homogeneous light of a

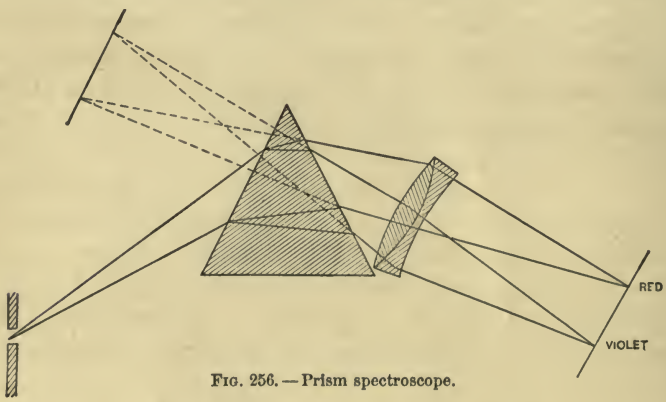


Fig. 256. — Prism spectroscope.

definite wave length. The method ordinarily adopted is to make use of the dispersive action of a prism. If waves of a definite index of refraction, emitted by a small source, fall upon one face of a prism, they suffer refraction and emerge on the other side, diverging apparently from a virtual image of the source. They do not, therefore, of themselves come to a focus ; but, if a converging lens is introduced, the rays may be focused upon a screen. This real image is due to the virtual image formed of the small source by the prism ; and it should be noted that the latter image is an astigmatic one, unless the pencil of incident rays meets the prism at the angle corresponding to minimum deviation. (See page 467.) The intensity of the effect can be increased if the point source is replaced by a series of such sources forming a line parallel to the edge of the prism. In general, a fine slit is made in an opaque solid, and the source of light, in the form of a flame, etc., is placed behind it. The image on the screen is in this case a narrow rectangle, practically a line, parallel to the slit.

If the source is emitting several trains of waves of different wave length, there will then be formed as many images of the slit as there are separate trains of waves. The light is said to be "analyzed"; or a "spectrum" of the light is said to be formed. If the slit is very fine and the adjustment for minimum deviation as exact as possible, the spectrum is said to be "pure," because in this case each train of waves affects only a very narrow rectangle on the screen, and so there is only a small amount of overlapping of the images. (Since the angle of minimum deviation is different for waves of different wave lengths, it is best to make the adjustment of the prism for the mean wave length of the light which is used. Further, since the focusing lens — even if achromatic — has a different focus for different waves, the screen must be curved so as to obviate the error, or one must be satisfied with a slightly impure spectrum.) In this manner the nature of the light — or, speaking in a more general manner, of the ether waves — emitted by any source can be investigated; and we shall discuss this matter in a later chapter.

Resolving Power. — Owing to the fact that the waves which leave the slit have passed through several "apertures" on their way to the screen (or to the eye), and that their wave front has been limited by this means to the shape of these apertures, the final image is not a *line*. These apertures are the edges of the lenses, of the prism, and of any diaphragms there may be in the tubes. There is, therefore, a diffraction pattern produced, exactly as described for a lens on page 484. The image of the slit, provided the light is homogeneous, is a broadened line accompanied on each side by a series of alternately dark and light fringes, whose intensity is much less than that of the central line. The distance apart of these fringes varies with the width of the aperture. So, if the source of light is emitting two trains of waves of wave length l and $l + \Delta l$, there will be two overlapping diffraction patterns; and the two trains of waves will produce two *distinct* images if the central line of one pattern is so far displaced by dispersion as to coincide with the first dark fringe of the other. If t_1 is the length of the shortest path of any of the rays through the prism, and t_2 is the length of the longest one, $t_2 - t_1$ is practically equal to the length of the base of the prism; call its value b. If n and $n + \Delta n$ are the indices of refraction of the two trains of waves, of wave length l and $l + \Delta l$, which are just resolved by the prism, it may be shown that $\Delta n = \frac{l}{b}$. In other words, if the prism has a thick base Δn is small, the image formed of the slit is extremely narrow, and the resulting spectrum is pure; but, if the prism is thin, the image formed is broad.

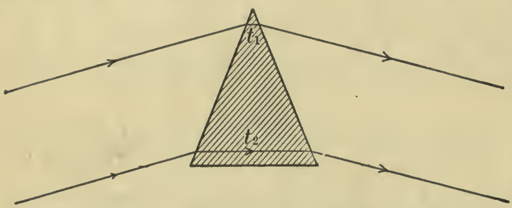


Fig. 257. — Diagram to illustrate the resolving power of a prism.

Spectroscope. — The essential parts, then, of an instrument to be used in order to form a pure spectrum are a slit, a prism (or other means for securing dispersion), and a converging lens. This is called a "spectroscope." In general, the arrangement is slightly different, a collimator and telescope being used as described for a spectrometer on page 505. When the slit is illuminated, plane waves emerge from the lens; these enter the prism and are dispersed by it; and, finally, as they leave the prism they enter the telescope which is focused for plane waves; and a series of colored images of the slit is seen by the observer looking in the eyepiece.

Measurement of Dispersion. — If the eyepiece is removed, and the images formed by the object glass are focused sharply on an opaque screen, and, if a fine slit, parallel to that in the collimator, is made in this screen, it will be illuminated by practically homogeneous waves, and will serve as a source of such waves. This entire instrument is called a "monochromatic illuminator." By varying the position of this second slit (or by turning the prism) the wave length of the waves transmitted through it is altered.

Methods will be discussed later which enable one to measure the wave length of any train of waves; and, granting that the wave lengths of the waves emitted by the source are known, we have thus a method for measuring the refractive index of any substance for waves of definite wave lengths.

The substance to be studied is made in the form of a prism, and is placed on a spectrometer table. The entire instrument is turned until the slit of its collimator coincides with the slit of the monochromatic illuminator. The angle of the prism and the angle of minimum deviation for different waves are measured, and from a knowledge of these quantities the indices of refraction may be calculated.

Fraunhofer Lines. — It will be shown later that, when sunlight is analyzed by a spectroscope, it is found to consist of trains of waves of varying lengths, forming what appears at first sight to be a "continuous" spectrum; that is, one in which waves of all wave lengths within certain limits are present. But, if examined with an instrument of fair resolving power, the solar spectrum is seen to be distinguished by the absence of a great many trains of waves, as is shown by the presence of black lines across the bright colored spectrum. (The fact that this absence of the waves is manifested by the presence of *lines* is due, of course, to the use of a *slit* illuminated with sunlight as the source of the light in the spectroscope.) These lines are called "Fraunhofer lines," because they were first carefully studied by him. He measured — by methods to be discussed later — the wave lengths of the waves which correspond to these lines; and to the most prominent of them he gave names, in the form of letters. (Thus the A line is at the limit of vision in the red end of the spectrum; B and C are also in the red; D_1 and D_2 are in the yellow, etc.; K is at the limit of vision in the violet end of the spectrum.) It is customary, therefore, in studying the dispersion of any prism to illuminate the slit of the spectrometer with sunlight, and then to note the angle of minimum deviation that corresponds to the waves in the immediate neighborhood of the various strong Fraunhofer lines. Thus, n_E means the index of refraction for waves whose wave length is that of the E line, etc.

Dispersive Power. — The differences in the refractive index of various substances for different waves is shown in the following table, in which the columns give the values for the A, B, C, etc., lines:

	A	B	C	D	E	F	G	H
Water, 16° C.	1.330		1.332	1.334		1.338		1.344
Carbon bisulphide, 10°	1.616		1.626	1.635		1.661		1.708
Crown glass	1.528	1.530	1.531	1.534	1.537	1.540	1.546	1.551
Flint glass	1.578	1.581	1.583	1.587	1.592	1.597	1.606	1.614
Rock salt, 17°	1.537	1.539	1.540	1.544	1.549	1.553	1.561	1.568

It is evident from this table that not alone is the index of refraction different for different substances, but that also the dispersion of any two rays, *e.g.* $(n_D - n_C)$, may be the same for two substances and the dispersion of two other rays may differ; thus for crown glass and rock salt $n_B - n_A = 0.002$, but for the former $n_H - n_F = 0.011$, while for the latter it equals 0.015. Further, what is called the "dispersive power" is different. This is defined to be the ratio $\frac{n_H - n_A}{n_D - 1}$. (In practice, the n in the denominator is taken as n_C, or n_D, or n_E, etc., whichever happens to be best known; for the error thus introduced is negligible.) Its value for the above substances is therefore:

Water	0.042	Flint glass	0.061
Carbon bisulphide	0.145	Rock salt	0.057
Crown glass	0.043		

The dispersive power of a substance is proportional to the length of the spectrum produced by a prism made of it, which has a small angle. For,

$$n = \frac{\sin \frac{A + D}{2}}{\sin \frac{A}{2}};$$

and if A and therefore D are both small, we may write $n = \frac{A + D}{A}$; so $n_H - n_A = \frac{D_H - D_A}{A}$, and $n_D - 1 = \frac{D_D - 1}{A}$. Consequently, the dispersive power equals $\frac{D_H - D_A}{D_D - 1}$; but $D_H - D_A$, the difference in the minimum deviations of waves of wave lengths A and H, is practically the angle subtended at the prism by the length of the spectrum when the prism is in its position of minimum deviation. Therefore, prisms having the same angle but made of different substances produce spectra of different length and of different dispersion for the same waves; and the average deviation for the whole spectrum may be different also.

It is owing to these facts in regard to dispersion that it is possible to construct an achromatic lens, or to combine two prisms in such a manner as to produce no deviation of a definite train of waves, but to disperse the other waves.

Direct-vision Spectroscope. — The spectroscope that has just been described suffers from the disadvantage that the collimator and telescope are not in the same direction, and therefore in some cases it is difficult to support the instrument rigidly. As has been just explained, however, it is a simple matter to choose two prisms of different materials which, when combined with their edges parallel but turned in opposite directions, will not deviate one particular train of waves, but will the others. The dispersion produced in this way by two prisms is not great; but, if several are thus combined, it may be made as great as desired. If such a series of prisms is combined with a collimator and telescope, the apparatus is all along one straight line. Such an instrument is called a "direct vision spectroscope."

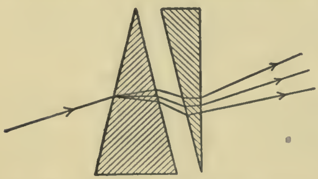


FIG. 258. — A direct vision spectroscope.

Normal and Anomalous Dispersion. — The table given above shows that in all the cases cited the index of refraction increases as the wave length decreases; this is called "normal dispersion." There are, however, many substances with which this is not true. As will be shown later, every substance absorbs more or less completely ether waves of certain wave lengths; and some substances absorb only waves of wave lengths that lie within a narrow limit, *e.g.* only waves having wave lengths that differ but slightly. In such cases the index of refraction for waves whose wave lengths do not differ much from this wave length of the "absorption band" does not follow the above law connecting it with the wave length. Such dispersion is called "anomalous." There may, of course, be several such absorption bands, introducing further complication.

Dispersion Curves and Formulæ. — The various facts in regard to dispersion, both normal and anomalous, can be best shown in a diagram whose axes are indices of refraction and wave lengths, or by a formula. Thus, for normal dis-

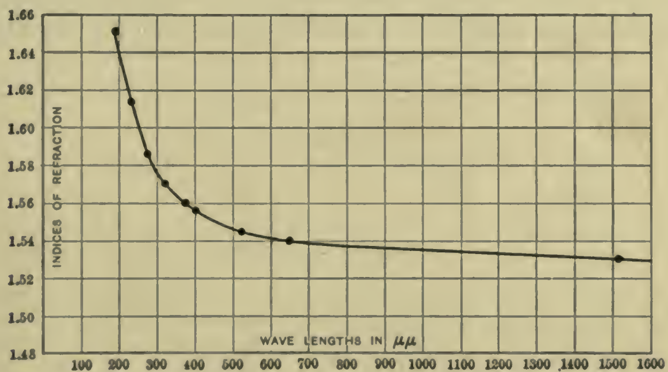


FIG. 259. — Dispersion curve for quartz.

persion, the connection between index of refraction and wave length is given by a curve like that shown in the cut, which is the "dispersion curve" for quartz. This may be also expressed by a formula, in which n is the index of refraction and l the wave length, viz.:

$$n^2 = -Al^2 + B + \frac{C}{l^2} + \frac{D}{l^4}.$$

(In the neighborhood of an absorption band this formula ceases to be true.)

If there is a single absorption band whose wave length is l_0, the dispersion curve has a general form as shown in the cut, which is the actual curve for cyanine as determined by

Professor R. W. Wood. This substance has an absorption band at about wave-lengths 600 $\mu\mu$.

The formula which expresses the facts outside the absorption band is $n^2 = A + \frac{B}{l^2 - l_0^2}$. It is seen from this or from the curve that for wave lengths not far from the absorption band, if $l_1 > l_0 > l_2$, $n_1 > n_2$; that is, the refractive index of the longer waves is the greater. So, if a prism is made of this substance, it will deviate the long waves near the absorption band more than the short ones; and the resulting

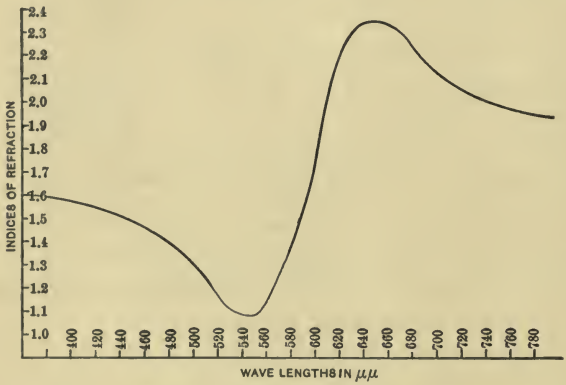


FIG. 260. — Dispersion curve of cyanine.

spectrum will appear, in general terms, as if it were divided in two parts by the absorption band and as if these two halves were shifted toward and across each other.

Whenever there is an absorption band, there is anomalous dispersion, and many substances show these phenomena in the visible spectrum; such are the aniline dyes, the vapors of sodium and other metals, thin layers or films of metals, etc. It will be seen later that all these substances have other optical phenomena which are closely connected with this fact.

Rainbows. — An interesting illustration of dispersion is furnished by the phenomenon in nature called the "rainbow." After a rain shower, if the sun is not far from the horizon, and the rain clouds have passed in such a direction that the observer is between them and the sun, a series of colored arcs or circles may be seen on looking away from the sun. These colored arcs are in the following order: violet on the inside shading off to red on the outside, then a dark space, and another arc colored red on the inside and violet on the outer edge. They are all portions of circles whose centres coincide at a point lying in the prolongation of the line joining the sun with the eye of the observer. There are often also "supernumerary" bows inside the primary one.

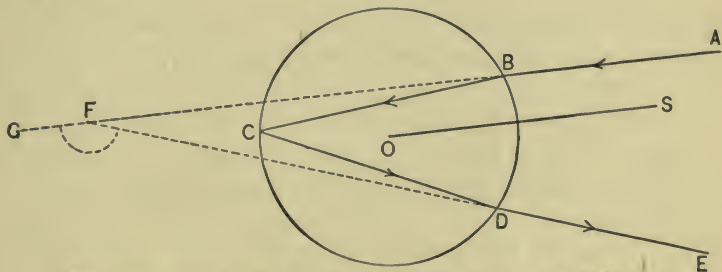


Fig. 261.—Rainbow: refraction of a ray by a drop of water, single reflection.

A complete explanation of these phenomena requires a consideration of the size of the drops, the nature of sun-light, the size of the sun, etc.; this was first given by Sir George Airy. An elementary, but imperfect, theory was given by Descartes; an outline of which is as follows. Consider a raindrop as a sphere of water, and draw the paths of the rays incident upon it from the sun. Certain rays will enter the drop, suffer reflection once and be refracted out, as shown in the cut, in which O is the centre of a drop, OS is the direction of the sun, and $\overline{AB}$, $\overline{BC}$, $\overline{CD}$, $\overline{DE}$ represent the portions of a ray. The deviation of the ray is shown in the cut by the angle (GFE), where F is the intersection of the prolongations of the incident and emerging portions of the ray, $\overline{AB}$ and $\overline{DE}$. By means of higher mathematics it may be shown that if there is a homogeneous beam of rays, all parallel to $\overline{OS}$, falling upon the drop, the angles of deviation all exceed a definite value. In other words, there is a *minimum* value of the deviation, and this fixes a certain direction with reference to the line $\overline{OS}$. If this minimum angle is N, draw the line $\overline{AB}$ making such an angle with $\overline{SA}$. Then of all the

Fig. 262.—Rainbow: N is angle of minimum deviation.

rays parallel to $\overline{OS}$ falling upon the drop, none are so deviated as to emerge outside the angle (SAB); some emerge in the directions $\overline{AB_1}$, AB_2, etc. This minimum angle is different for different wave lengths; for violet it is about 140°, for red about 138°, etc. Therefore, if the observer looks up at the rain cloud in such a direction that his line of sight makes an angle of 180° – 140°, or 40° (or less), with a line joining his eye to the sun, he will receive violet light from the rain-

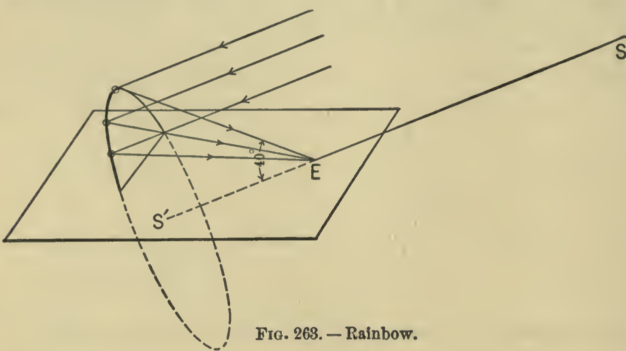


FIG. 263. — Rainbow.

drops. If he looks up at an angle greater than 40°, he receives no violet light at all. Therefore, the observer will see a violet arc, each point of which subtends at his eye an angle of 40° with the line drawn from the sun.

Similarly, with the other colors, there is a red arc corresponding to the angle 180° – 138°, or 42°, which is sharply defined on its outer edge, and to see which one must look higher up in the sky than was necessary in order to see the violet arc.

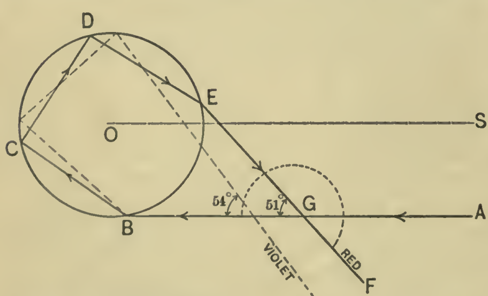


FIG. 264. — Rainbow: refraction of a ray by a drop of water, double reflection.

But some of the rays from the sun suffer two reflections in the raindrop, as shown in the cut. As before, we may show that there are angles of minimum deviation for the different colors, which give rise to a red bow at the angle 51° and a violet one at 54°, these bows being sharply defined on their lower edges.

CHAPTER XXXI

INTERFERENCE OF LIGHT

Interference Fringes. — The general phenomena of interference of waves were described in Chapter XXI, page 374, and the special ones dealing with light were discussed in Chapter XXV, page 420. It was shown that the simplest mode of illustrating interference was to place two *identical* sources of light close together and to allow them to illuminate a screen or to enter the eye directly. If the two sources are parallel slits, the interference pattern is a series of parallel colored fringes; if homogeneous light is used, these are alternately bright and dark, and at regular intervals apart proportional to the wave length; if white light is used, the fringes are a superposition of different sets, each of which is due to a different color, thus proving that white light is equivalent to a superposition of waves of different wave length. Unless the two sources are identical, there is no permanent phase relation between the two sets of waves emitted by them, and so these cannot interfere.

In demonstrating the interference fringes which are produced by the two identical sources, a converging lens is always used, the arrangement being as shown in Fig. 265. O_1 and O_2 are the two sources; L is the lens; M is a screen placed at a distance from the lens equal to its *focal length for parallel rays.* The two sources are emitting rays in all directions; let $\overline{O_1A_1}$ and $\overline{O_2A_2}$ be two parallel rays. After refraction by the lens, they will meet at the point B in the screen, where CB is a line drawn through the centre of the lens parallel to the incident rays. (See page 479.) If the difference in length of the *optical* paths of these rays is half a wave length (or an odd number of half wave lengths), there will be complete interference at B. This difference in path is found exactly, as on page 377, by drawing a line from O_2 perpendicular to $\overline{O_1A_1}$. For, if B were a source of waves, two of its rays would be $\overline{BA_1O_1}$ and $\overline{BA_2O_2}$; and the position of the wave front at any instant would be a plane perpendicular to $\overline{A_1O_1}$ and $\overline{A_2O_2}$. So, when the wave front reached O_2, it would be at P on the line $\overline{A_1O_1}$, where $\overline{O_2P}$ is perpendicular to $\overline{A_1O_1}$. In other words, the optical paths $\overline{O_2A_2B}$ and $\overline{PA_1B}$ are equal in length; and so the difference in length of the optical paths from O_1 and O_2 to B is $\overline{O_1P}$.

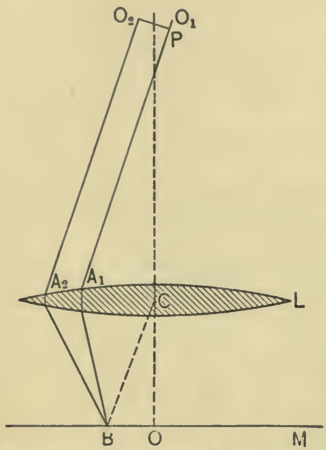


FIG. 265. — Young's interference experiment.

By means of the lens, then, the interference fringes are focused on the screen. The central image is at O, and the distance $\overline{OB}$ can be expressed at once in terms of the difference in path of the rays from O_1 and O_2. Call this difference D. Then

$$D = \overline{O_1P} = \overline{O_1O_2} \sin (O_1O_2P).$$

But the angles (O_1O_2P) and (BCO) are equal; and sin (BCO) equals $\frac{\overline{BO}}{\overline{BC}}$. Hence

$$\overline{BO} = \overline{BC} \sin (BCO) = \overline{BC} \sin (O_1O_2P) = D\frac{\overline{BC}}{O_1O_2} = \frac{D}{O_1O_2}\sqrt{\overline{OC}^2 + \overline{BO}^2}.$$

If the focal length of the lens, $\overline{OC}$, is called f, and if it is large compared with $\overline{BO}$, we have the relation

$$BO = \frac{D}{O_1O_2} \cdot f.$$

The condition for complete interference is that

$$D = (2n + 1)\frac{l}{2};$$

and for this value $\overline{BO} = \frac{(2n+1)l}{2}\frac{f}{O_1O_2}.$

Therefore the distance between two dark fringes is $\frac{lf}{O_1O_2}$. There is thus offered a method for the measurement of the wave length of light, as the various quantities in the formula with the exception of l may all be measured directly.

We shall now describe several methods of securing two identical sources.

Young's Method. — This consists, as has been already described, in illuminating two parallel slits in an opaque screen by light from another slit parallel to them and at the same distance from each. The slits must be sufficiently narrow to diffract the waves in all directions. (See page 392.)

Fresnel's Biprism. — This consists of a combination of two identical thin prisms which are placed base to base, as shown by the cross section in the cut. An opaque screen, with a slit in it, is so placed that the slit is parallel to the two edges of

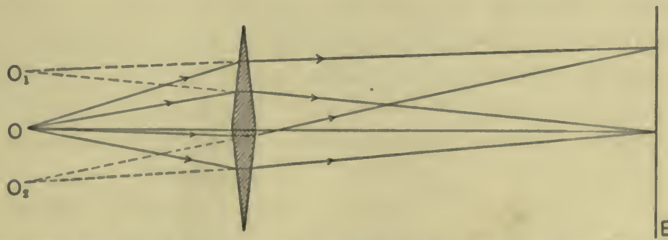


FIG. 266. — Fresnel's biprism: two virtual images O_1 and O_2 of the source O are produced.

the biprism and at equal distances from them. This slit, if illuminated, will emit waves, which will suffer refraction and deviation by the two halves of the prism. If the slit is at O, as shown in the cut, one half will form a virtual image at O_1, the other at O_2. So the waves as they emerge from the biprism will come apparently from the two sources O_1 and O_2; and there are then two *identical* trains of waves, which will interfere, and may be focused on a screen by a lens, as described above. Therefore, if the distance $\overline{O_1 O_2}$ is known (and it may easily be determined by experiment), the wave length of the light may be measured.

Fresnel's Mirrors. — These are two plane mirrors which are carefully adjusted until their faces are *slightly inclined* to each other, but are in actual contact along a line.

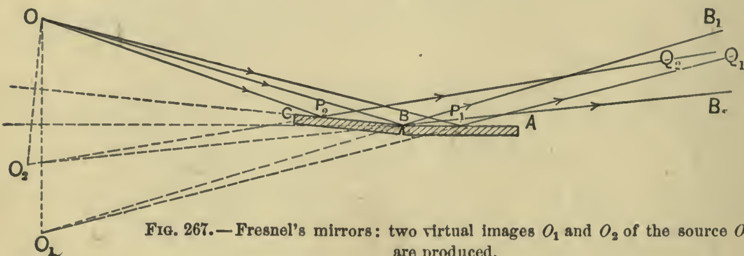


FIG. 267.—Fresnel's mirrors: two virtual images O_1 and O_2 of the source O are produced.

A slit O is placed parallel to this line; and therefore two virtual images of it, O_1 and O_2, are formed by the two mirrors. Let B be the line of contact of the two mirrors. Then a pencil of rays $P_1 OB$ falling upon the first mirror will be reflected into the pencil $Q_1 O_1 B_1$; and the pencil BOP_2 falling upon the second mirror will be reflected into $B_2 O_2 Q_2$. Therefore there will be a region included in the angle $(B_1 B B_2)$ which is traversed by two trains of waves coming from identical sources.

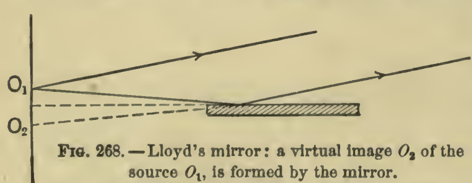


FIG. 268.—Lloyd's mirror: a virtual image O_2 of the source O_1, is formed by the mirror.

Lloyd's Mirror. — In this arrangement a slit is placed parallel to a plane mirror, at some distance from it, but only a slight distance above its plane. There will be a virtual image formed by the reflected rays; and so any point above the mirror will receive waves *directly* from the slit, and also by reflection, apparently coming from the virtual image of the slit. There will therefore be interference. (These two sources of waves are only approximately identical, for one is the *inverted* image of the other.)

Colors of Thin Plates. — The first interference phenomenon which was recognized as such and so explained is the production of the brilliant color effects by such thin films of transparent matter as soap bubbles, films of oil on water, layers of air between two pieces of nearly parallel glass, etc. These colors are due to the interference of the trains of waves which suffer reflection directly at one surface of the film with the waves which are refracted out from the film after one or more internal reflections. If we consider any point on the surface of a film which is receiving homogeneous light from any point source, one ray from the latter is *reflected* at the point, and other rays emerge there which have entered the film at other points and have suffered reflection at the surfaces of the film. It is evident that these rays have had paths of different lengths; and that also the ray directly reflected has suffered reflection when incident upon the surface of the film from the surrounding medium, while the emerging rays have suffered reflection when incident upon the surface of the film from its interior. Owing to this last cause there is a difference in phase introduced, in addition to that caused by the difference in path, because the reflection in the one case is from a "fast" to a "slow" medium, and is the opposite in the other. (See page 334.) This additional difference in phase is equivalent, as was shown, to half a period of the vibration. If the total effect of all the rays at the point on the surface of the film is null owing to interference, this point will appear dark; while, if the rays do not destroy each other's action, the point will be bright. It is evident also that a film of such a thickness as to cause interference for waves of a definite wave length will not, in general, cause interference for other waves; so, if white light is used, a point where there is complete interference for a definite train of waves will appear colored, owing to the fact that the other waves are not cut off; and the effect is as if one color were completely removed from the constituent colors of white light. The trains of waves which interfere at the top surface of the film are not *destroyed*, for energy cannot be annihilated; they are transmitted through the film and emerge on the lower side. Thus, when white light is incident upon a transparent film, some is reflected at the upper surface, the rest enters the film; of this a certain amount is reflected once or more times, and is finally refracted out through the upper surface, while the rest is either directly, or after two or more reflections, refracted out through the lower surface. By far the greater amount of the light is transmitted, owing to the poor reflecting power of the film. There will then be colors visible if one looks at either of the two surfaces of the film; but those seen by looking back at the second surface are much weaker than those at the other, owing to the presence of so much white light. If the film does not absorb the waves, the combined effects on its two sides are exactly equivalent to the incident waves; that is, they are "complementary."

We shall now consider in detail the case of a thin film with parallel faces, and we shall suppose that the film has a greater index of refraction than the surrounding medium, *e.g.* a film of water in air. Let the rays come from a homogeneous point source at such a distance compared with the area of the film that they may be regarded as parallel. Let $\overline{OC}$ in the cut be an incident ray; it undergoes reflection at C and gives rise to a ray $\overline{CD}$ in general. Other rays emerge by refraction at C; one of these is due to the incident ray $\overline{PA}$. This is refracted into the ray $\overline{AB}$, then reflected into $\overline{BC}$, and finally refracted out. We can calculate the difference in optical path of these rays from the source to C. Draw

$\overline{AE}$ perpendicular to the incident rays; then the phase at A and E is the same because they are on the same wave front; draw $\overline{FC}$ perpendicular to $\overline{AB}$, it represents the refracted *wave front;* and therefore the phase at C and F is the same. The difference in optical path of the two rays meeting at C is, then, $n(\overline{FB} + \overline{BC})$, where n is the index of refraction of the film with reference to the surrounding medium. Calling the thickness of the film h, and the angle of refraction into the film r, it is seen from geometry that $\overline{FB} + \overline{BC} = 2\,h \cos r$. Therefore the difference in optical path is $2\,nh \cos r$; and the corresponding *time* lag is $\frac{2\,nh \cos r}{v}$, where v is the velocity of the waves in the outer medium. There is also the additional difference in phase of half a period, $\frac{T}{2}$, due to the difference in the character of the reflection at B and at C. So the total difference in phase, expressed in terms of time, between the two rays $\overline{OCD}$ and $\overline{PABCD}$ is

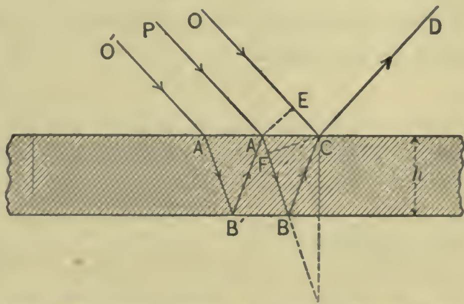


FIG. 269. — Colors of thin plates. (The transmitted rays are omitted.)

$$\frac{2\,nh \cos r}{v} + \frac{T}{2}.$$

This may be expressed differently, by substituting for v its value $\frac{l}{T}$, where l is the wave length in the outside medium; viz.,

$$T\left(\frac{2\,nh \cos r}{l} + \frac{1}{2}\right).$$

Another ray which will have a component emerging at C is $\overline{O'A'}$, incident at A', etc.; and it is not difficult to prove that the resultant *emerging* ray at C has the same amplitude as the *reflected* ray, but differs from it in phase by the amount given above. If this difference equals an odd number of half periods, the two rays interfere completely. This condition is, then, that

$$\frac{2\,nh\cos r}{l} + \frac{1}{2} = \frac{(2\,m+1)}{2},$$

where m is any whole number 0, 1, 2, 3, etc. This reduces to

$$2\,nh\cos r = ml.$$

Therefore if $\quad h = 0, \text{ or } \dfrac{l}{2\,n\cos r} \text{ or } \dfrac{l}{n\cos r}, \text{ etc.,}$

the interference is complete.

If, then, the film is wedge shaped, there will be interference at points on the surface of the film corresponding to these critical thicknesses; and alternately dark and light bands across the film will be seen.

The light used to illuminate films generally comes from an extended source, like a flame, which is not far from the film. In this case any point C on the surface of the film receives two rays from a point O of the source, viz., $\overline{OCD}$ and $\overline{OABCD'}$; and in general the directions of these rays after leaving C are not the same. Similarly, C receives rays from other points of the source, which also have different directions after leaving it. So this point may be regarded as itself a point source; and, if the eye is focused on the surface of the film, the question as to whether C appears dark or not depends upon how exactly these pairs of rays from each point of the source neutralize each other. If the incidence is practically normal, the value of r in the formula is so small that $\cos r = 1$; and the condition for complete interference is that $2\,nh = ml$, which is the same for all these

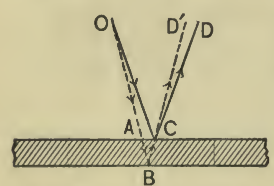


FIG. 270.—Colors of thin plates: localization of colors in the surface.

pairs of rays. If the thickness of the film at the point C satisfies this condition, and if the eye is focused on the film, C will be a dark point. So, in general, when an extended source of light is used and the incidence is practically normal, the dark or light bands are to be seen by looking at the surface of the film.

Another illustration of these formulæ is given when the convex surface of a plano-convex glass lens is pressed closely against a glass plate. The film of air between the two pieces of glass may be regarded as made up of a great number of concentric rings, each ring having the same thickness at all points. The film is then like that of a wedge one of whose surfaces is curved; and there is, of course, symmetry around the centre, or point of contact. Therefore if homogeneous light is incident normally upon this film, there will be a central dark spot surrounded by alternately bright and dark rings. The dark rings are given by $h = 0, \frac{l}{2n}, \frac{2l}{2n}$, etc., where h is the thickness of the film. These are known as "Newton's rings." h can be expressed in terms of the radius of the spherical surface of the lens and of the radius of the dark ring; and a method is thus offered for the measurement of the wave length of light; or, if this is known, for the measurement of the radius of the spherical surface of the lens.

FIG. 271.—Apparatus for Newton's rings.

If in any of the above experiments white light is used, there are no dark bands or rings — all are colored; and the details can in each case be deduced from the general principles given on page 524.

Interference over Long Paths. — In the cases so far treated, the interfering rays have been considered to differ in path only by a small amount, but there are many interesting and important phenomena in which this condition is not fulfilled. Thus let O be a homogeneous point source of light; P be a plate of transparent material; C be a converging lens; and D be a screen placed in the *principal* focus of the lens. Draw from C, the centre of the lens, a line $\overline{CE}$ at random. Then, all rays falling upon the lens parallel to this line will be brought to a focus at E. The point source O is emitting rays in all directions; one of them, $\overline{OP_1}$, is parallel to $\overline{CE}$. This ray after incidence upon the plate gives rise after successive reflections and refractions to a series of rays $\overline{Q_1R_1}$, $\overline{Q_2R_2}$, etc., all parallel to each other and to the original ray. Therefore they all unite at E. The difference in path of two consecutive rays is $2h \cos r$; and, if this equals an odd number of half wave lengths, there is complete interference at E, and also, by symmetry, for all points in a circle drawn around D with $\overline{ED}$ as a radius. If this difference equals a whole number of wave lengths, the ring is bright. Thus, corresponding to the rays from O in all directions there will be a series of circular rings, alternately dark and light, around D. Similarly, if instead of having a point source, an extended one is used, each point will give rise to identically the same series of circular rings around D; and so the effect is more intense.

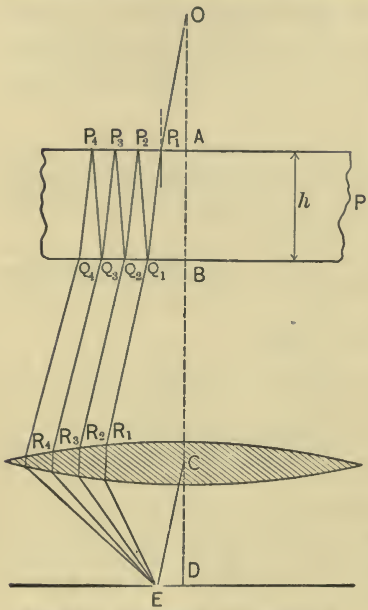


Fig. 272.—Interference over long paths.

If the source emits two trains of waves of different wave length, there will be two sets of concentric rings around D; at certain points a ring of one set may coincide with one of the other, and at others it may fall between two rings of the other set. A connection may be established between the radius of any one ring, say the tenth, the focal length of the lens, the thickness of the transparent plate, and the wave length of the light; and it is not difficult to see how a method can be devised for measuring the relation in wave length of the two trains of waves emitted by the source.

The transparent plate is in general a layer of air included between two plane parallel glass plates. If one of these is kept fixed and the other is moved, h may be varied at will. Professor Michelson of Chicago has obtained interference fringes in this manner, using radiations from mercury vapor, when h was so great that the difference in path between two interfering rays amounted to 540,000 wave lengths; Fabry and Perot of Marseilles have more recently obtained interference over a distance equivalent to 790,000 wave lengths; and Lummer has succeeded in using even greater distances. This proves that this light must be extremely homogeneous, otherwise there would be overlapping series of rings.

Stationary Waves. — Since light is a wave phenomenon, it must be possible to produce stationary light waves exactly as was done with waves in cords and in air. (See page 349.) This was shown experimentally by Wiener, in 1890, by allowing homogeneous light to fall perpendicularly upon a polished metal mirror. If *M* is the mirror, and plane waves are incident normally upon it, there will be nodal planes at the mirror and along A_1, A_2, A_3, etc., at intervals of half a wave length; and there will be loops along planes halfway between these. Wiener placed a *thin* photographic plate inclined to the mirror, and showed on development of it that light had been received along lines through B_1, B_2, B_3, etc., where the loops were, but that there had been no light at all along lines through A_1, A_2, A_3, etc., where the nodes were.

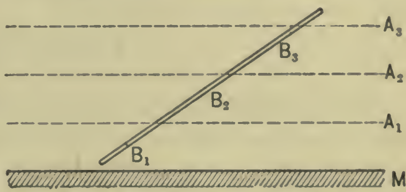


FIG. 273. — Wiener's experiment with stationary light-waves: A_1, A_2, A_3, etc., are nodal planes.

CHAPTER XXXII

DIFFRACTION

The general phenomena of diffraction have been illustrated in previous chapters in giving the explanation of the rectilinear propagation of light and the fringes observed near the shadows of opaque obstacles, in the description of the passage of waves through small openings, and in the discussion of the resolving power of lenses and prisms. Many other illustrations may be found in larger text-books; but only one additional phenomenon will be discussed here.

The Diffraction Grating. — If a great number of fine scratches are made on the surface of a thin glass plate by means of a diamond point, the lines so drawn being parallel and at equal intervals apart, the effect, so far as transmission is concerned, is exactly as if a great number of fine rectangular openings, identically alike and evenly spaced, were made in an opaque screen. A piece of glass so "ruled" is called a "plane transmission grating." Let us consider homogeneous plane waves incident normally upon such a grating. Each point of a slit, or opening between two scratches, serves as a source of waves, and sends out rays in all directions. Let

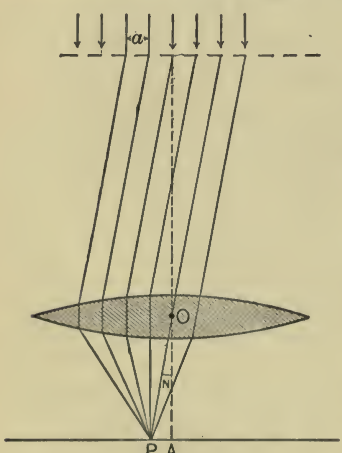


Fig. 274. — Transmission diffraction grating.

a converging lens be placed with its axis perpendicular to the plane of the grating, and let a screen be placed in its focal plane. In the cut, draw an arbitrary line $\overline{OP}$ from the centre of the lens to the screen; all the rays from the various openings of the grating which are parallel to this line will be brought to a focus at P. The difference in path of any two such rays from the corresponding edges of two consecutive openings, or from two corresponding points in two consecutive openings, may be deduced at once. Call the distance from the edge of one opening to the corresponding edge of the next, *i.e.* the "grating space," a; and the angle (POA), N. Then, the difference in path referred to is $a \sin N$; and, if this is a whole number of wave lengths, parallel rays from corresponding points in all the openings will coincide in phase at P; and it will be a bright point. (There is an exception to this, when rays from some of the points in any one opening interfere with those from other points in the same opening; but this case need not be discussed here.) There will therefore be a line of light through P parallel to the openings in the grating. The condition, then, that P should be bright is:

$$a \sin N = ml,$$

where $m = 0, 1, 2, 3$, etc., and l is the wave length. Consequently, there is a series of bright lines determined by

$$a \sin N_0 = 0,\ a \sin N_1 = l,\ a \sin N_2 = 2\,l,\ a \sin N_3 = 3\,l, \text{ etc.};$$

that is, by

$$\sin N_0 = 0,\ \sin N_1 = \frac{l}{a},\ \sin N_2 = \frac{2\,l}{a},\ \sin N_3 = \frac{3\,l}{a}, \text{ etc.}$$

The light along any line through a point P defined by these relations is bright; then for neighboring portions of the screen as one takes points farther and farther away from P the light fades gradually away and vanishes. It rises to a maximum in another line defined by the next value of N, etc.

There is a maximum for the direction $\sin N_0 = 0$, or $N_0 = 0$; that is along a line through the point A in the cut where the axis of the lens meets the screen; this is called the "central image"; the next maximum, at the angle N_1, is called the "first spectrum," etc. It is evident that there are maxima also on the other side of A, corresponding to negative values of N. The number m is said to give the "order" of the spectrum.

If white light is used, instead of homogeneous light, each constituent train of waves has its own series of spectra: a central image, and spectra of different orders on its two sides. The spectra of the different colors overlap; and the spectrum of any one order is not pure unless the individual

Fig. 275. — Photograph of spectra produced by a grating, showing the different orders.

spectra formed for any one train of waves are extremely narrow. We can easily determine how wide any one spectrum "line" is by calculating the position of the point next it on either side where the intensity is zero. This condition involves complete interference at that point of all the rays reaching it from the grating. Let P be a point where there is a maximum, and let P_1 be the nearest minimum on the side toward A. Draw the line $\overline{OP_1}$, and call the angle (P_1OA), N_1; then all rays parallel to $\overline{OP_1}$ are brought to a focus at P_1. Let us suppose that there is an *even* number of openings in the grating; if there is an odd number, we may consider the last one by itself, and its effect in comparison with that of the others may be neglected. If the number of openings is $2n$, the condition that P should be the position of the mth spectrum may be expressed by saying that the difference in path of the two rays reaching it from corresponding points in the first opening and the middle one, equals nml. For this difference in path is $na \sin N$, and it has been shown that the condition for a maximum is that $a \sin N = ml$. If P_1 is to be the nearest minimum, the difference in path between two rays from these same points must differ from this value for P by half a wave length; for, if this is true, the rays from the first and nth, the second and the $(n+1)$th, etc., will interfere completely. The condition for a minimum at P_1 is, then,

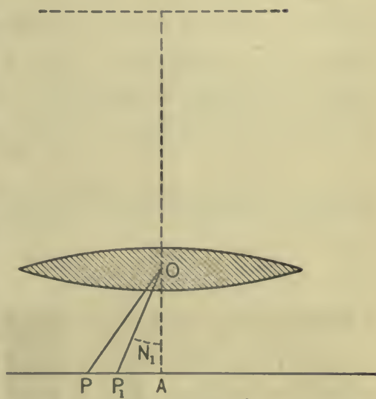


Fig. 276. — Diagram illustrating the resolving power of a grating.

$$na \sin N_1 = nml - \frac{l}{2}.$$

(Similarly, for a minimum point on the other side of P, the value would be $nml + \frac{l}{2}$.) The condition for a maximum at P is,

$$na \sin N = nml.$$

Hence,
$$\sin N - \sin N_1 = \frac{l}{2na}.$$

The entire number of openings in the grating is $2n$, and each has the grating space a; so $2na$ is the width of the grating; and the formula shows that in order for N to be nearly equal to N_1, that is, for P_1 to be very close to P, the width of the grating must be large. Under these conditions the spectrum "lines" are narrow.

The distance from one spectrum line produced by a homogeneous train of waves to the next one is determined by giving m two consecutive values in the formula $a \sin N = ml$. Thus, the fourth spectrum is at an angle N_4 whose sine satisfies the equation

$$\sin N_4 = \frac{4l}{a};$$

and the corresponding formula for the fifth spectrum is

$$\sin N_5 = \frac{5l}{a}.$$

Thus, $\sin N_5 - \sin N_4 = \frac{l}{a}$. This relation is general; and it is seen, therefore, that in order to have consécutive spectra far apart the grating space a must be small.

If white light is used, or waves from some complex source, the central image will receive light of all wave lengths; and, in addition, a series of spectra will be produced on both sides of this. In order to have these spectra *long*, *i.e.* the dispersion great, the grating space must be small; and to have the spectra *pure*, *i.e.* the "lines" narrow, the grating must be wide.

This condition for "purity" may be expressed differently. If there are two trains of waves of wave lengths l and $l + \Delta l$, which differ only slightly, their spectral images will, as a rule, overlap; but, if Δl is so large that the maximum of the waves of length $l + \Delta l$ coincides with the minimum of the other waves, the two images or "lines" may be seen distinct from each other. The condition for a maximum of the former waves is that N should be such that

$$a \sin N = m(l + \Delta l);$$

and we have just seen that the condition for the first minimum of waves of length l is that the angle N_1 should be such that

$$na \sin N_1 = nml + \frac{l}{2},$$

or that

$$a \sin N_1 = ml + \frac{l}{2n},$$

where $2n$ is the number of grating spaces. If these two positions coincide, that is, if $N = N_1$

$$m(l + \Delta l) = ml + \frac{l}{2n},$$

or

$$\frac{l}{\Delta l} = 2nm.$$

This is the spectroscopic "resolving power" of the grating; and it is seen that, in order for Δl to be small, the number of grating spaces must be large and a high order of spectrum should be used.

It should be noted that, even if the spectra formed of the components of white light are pure, the ones of *different order* are superimposed on each other. Thus, waves of wave length l have a spectrum line in the first order, at the same point where waves of wave length $\frac{l}{2}$ have a line in the second order, and where waves of wave length $\frac{l}{3}$ have a line in the third order, etc. It should be noted further that the above formulæ are all independent of the material of the grating.

If the plane waves are incident upon the grating at an angle I, different from 0°, the necessary modification in the formulæ is deduced easily. Let O_1 and O_2 be the edges of two consecutive grating openings; let $\overline{P_1O_1}$ and $\overline{P_2O_2}$ be two of the incident rays; and let $\overline{O_1Q_1}$ and $\overline{O_2Q_2}$ be the two rays from O_1 and O_2 which are diffracted in the direction defined by the angle N. Draw $\overline{O_2A}$ perpendicular to $\overline{P_1O_1}$, and $\overline{O_2B}$ perpendicular to $\overline{O_1Q_1}$. At the points A and O_2 the rays are in the same phase; therefore the difference in path of these rays when they are united by the converging lens is $\overline{AO_1} + \overline{O_1B}$. If the grating space $\overline{O_1O_2}$ is called a, as before, this difference in path may be written $a(\sin I + \sin N)$; and the condition for a spectrum line is, then, that

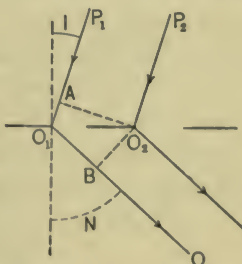


FIG. 277. — Oblique incidence upon a grating.

$$a(\sin I + \sin N) = ml.$$

All the other formulæ are modified in a similar manner.

If we consider the spectrum lines formed on the screen in the immediate neighborhood of the axis of the lens, which is supposed to be perpendicular to the surface of the grating, we may replace sin N by N, and the formula becomes

$$a(\sin I + N) = ml.$$

If white light is incident, there will be a continuous series of colored spectra. Let N_1 correspond to the waves of wave length l_1, and N_2 to those of wave length l_2; that is,

$$a(\sin I + N_1) = ml_1,$$

$$a(\sin I + N_2) = ml_2,$$

or,

$$a(N_2 - N_1) = m(l_2 - l_1).$$

This means that *lengths* along the screen, $N_2 - N_1$, are proportional to differences in the wave lengths of the corresponding waves. A spectrum where there is this simple relation between the wave lengths of the waves and their distribution is called a "normal" spectrum. It is obvious that such spectra offer great advantages when one is interested in comparing the wave lengths of different waves.

Gratings are also made by ruling lines with a diamond point upon a plane polished metallic surface; these are called "plane reflecting gratings." Those portions of the surface which are not scratched, if sufficiently narrow, diffract the light in all directions, as do also the grooves made by the diamond point. Two parallel rays in any direction coming from two different grooves have a difference in path which depends upon the grating space, the angle of incidence, and the angle of diffraction. The general formula for a bright line is, as before,

$$a(\sin I + \sin N) = ml.$$

(The *intensity* of the light diffracted in different directions will depend obviously on the shape of the groove and on the reflecting power of the material of which the grating is made.)

The method ordinarily used for observing diffraction spectra is not to focus them upon a screen, but to place the grating on the table of a spectrometer, illuminate it by means of a collimator, and turn the telescope until the spectrum lines in turn come upon the cross hairs of the eyepiece. By means of the scale divisions upon the apparatus, the angles of incidence and diffraction may be read most accurately. The grating space may also be measured by ordinary laboratory methods, using a comparator. Therefore, in this manner the wave length of the waves may be determined with great accuracy.

The values of the wave lengths corresponding to the prominent Fraunhofer lines are given in the following table:

B	687.0186 $\mu\mu$	E_2	526.9723 $\mu\mu$
C	656.3054	F	486.1527
D_1	589.6357	G	434.0634
D_2	589.0186	H	396.8625
E_1	527.0495	K	393.3825

(The unit ordinarily adopted for expressing wave lengths is one tenth of a thousandth of a micron, that is, $\frac{1}{100000000}$ cm. This is called an "Ångström unit" because it was adopted by the great Swedish physicist for the expression of his results. Thus, the wave length of D_1 is 5896.357 Ångström units.)

Concave Gratings. — The use of a plane grating, that is, one ruled on a plane surface, involves the use also of a collimator and telescope for ordinary purposes; but, if the lines are ruled on a concave spherical metallic surface, this is not so. If we imagine a plane through the centre of the sphere and bisecting the rulings on the surface, and if a point source of light is in this plane, the grating will of itself produce spectrum lines without the need of lenses. These lines are short and are parallel to the grating rulings. They intersect the plane referred to, which passes through the point source and the centre of the spherical surface, in points which in

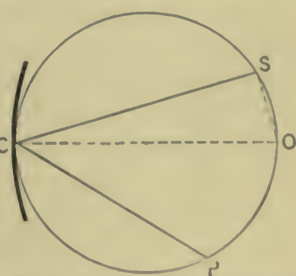


FIG. 278. — Diagram illustrating the principle of a concave grating. S is the source; C is the grating with centre at O; the image is on the circumference of the circle, *e.g.* at P.

certain cases satisfy a simple geometrical condition. If C is the centre of the grating, that is, the middle point of its ruled surface, and if O is the centre of curvature of this surface, draw a circle with $\overline{OC}$ as a diameter which is tangent to the grating surface at C. Then, if the point source S is at any point of this circle, the spectral images will be formed at points P on this same circle. If the source is emitting white light, there will be a central bright image formed by ordinary reflection, and on each side of this there will be series of spectra, all on this circle. Those spectra formed in the immediate neighborhood of O, the centre of curvature, are normal (see page 536), because $\overline{CO}$ is perpendicular to the grating, and so N in the formula is small.

There is a simple kinematic method of maintaining this normal condition for a grating, and yet varying the waves which are brought to a focus at O. If O is joined to S by a straight line, the triangle (OSC) is a right-angled one, having $\overline{OC}$ as a hypotenuse. Therefore, if two rigid beams, $\overline{SB}$ and $\overline{SA}$, are set up at right angles to each other, and are furnished with tracks along which small carriages may run, and if a beam of fixed length equal to $\overline{CO}$ is pivoted at each end to such carriages, then the points at the ends of this movable beam and the one at the intersection of the two fixed beams are always on the circumference of a circle whose diameter equals $\overline{OC}$ however the cross beam is moved, its ends always being on the two fixed ones. In practice, then, the slit or source of light is put at S, the intersection of the fixed beams; the concave grating is placed at one of the ends, C, of the movable beam and so turned that its centre of curvature comes at O, the other end of this beam; the observations of the spectra are

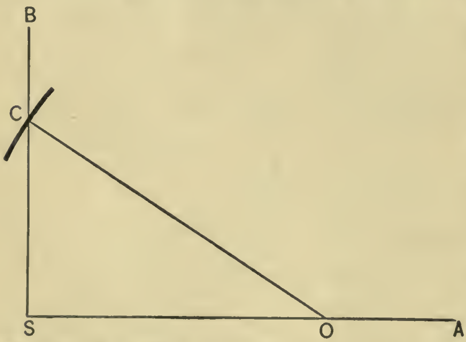


FIG. 279. — Rowland's arrangement of the concave grating.

then made, either visually or by photographic means, at O. For any definite position of the cross beam, certain waves in overlapping spectra are in focus at O; but, as the beam is moved, these change, owing to the change in the angle of incidence upon the grating, (SCO).

(When diffraction takes place through two rows of rectangular openings, which are superimposed at right angles to each other, spectra are formed, as shown in the cut, in two directions. This is illustrated by looking through an umbrella at an arc light.)

Fig. 280. — Diffraction pattern through rectangular meshes.

Historical. — Diffraction gratings were invented and first constructed in 1821 by Fraunhofer, after whom were named the dark lines of the solar spectrum. He used both transmission and reflecting plane gratings, and measured the wave lengths emitted by many sources of light. The concave grating and its peculiar mounting were invented by the late Professor Rowland; and all the gratings in use at the present time for scientific purposes throughout the world, whether plane or concave, have been made by ruling machines which he constructed for the purpose. In each of these there is a long cylindrical screw, to one end of which is attached a large toothed wheel, and which carries a long nut; the screw is supported in fixed collars near its ends, so, as it is turned by means of the toothed wheel, the nut can be made to advance; this nut pushes forward a platform, which moves on horizontal "ways"; a diamond is made automatically to move at right angles to these ways, with a to-and-fro motion. The metallic or glass surface to be ruled is attached firmly to this platform; and, as the diamond point is drawn across it, a line is ruled; then the diamond is raised and pushed back to the other side; in the meantime, by a partial turn of the screw, the platform is carried forward a short distance and the diamond point is lowered; then another line is ruled; etc. The screw is turned by means of pawls, which are worked by levers, and which push the toothed wheel around through the distance of one or more teeth at a time. Thus, if the pitch of the screw is $\frac{1}{20}$ in., and if there are 1000 teeth in the wheel, an advance of one tooth produces a motion forward of the nut of $\frac{1}{20000}$ in.; this, then, is the grating space. Nearly all of the small gratings made by Rowland's machines have 14,438 rulings to the inch, while the larger ones have 15,020. A full description of Rowland's machines is given in his *Physical Papers*, page 691.

Other forms of gratings have been made in recent years which give greater resolving power; but none are so generally useful as the ordinary concave grating.

CHAPTER XXXIII

DOUBLE REFRACTION

General Phenomena. — It was observed by Bartholinus as early as 1669 that, when a luminous point was viewed through Iceland spar, two images were seen. With any ordinary transparent substance, such as glass or water, only one image is observed; so the phenomenon is called "double refraction." It was later discovered that many substances had this property. If a doubly refracting substance is made in the form of a plate, any ray OA incident upon it from a point source O will therefore give rise to two refracted rays $\overline{AB}_1$ and $\overline{AB}_2$; and these will emerge as two parallel rays coming apparently from two virtual images O_1 and O_2, which are not in general in the line OPQ, drawn from O perpendicular to the surface; nor do, in general, the refracted rays lie in the plane of incidence.

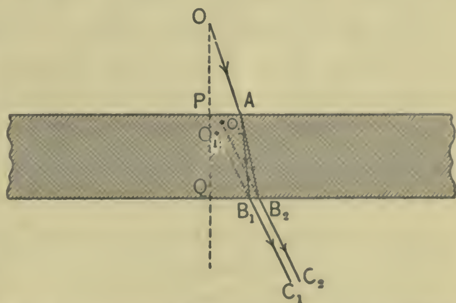


FIG. 281. — Double refraction by a plate. The two images, O_1 and O_2, are not, in general, on the line $\overline{OP}$.

Since refraction in all cases is due to the fact that the velocity of waves is not the same in all media, the explanation of double refraction in any medium is that a centre of disturbance in such a medium produces waves whose wave front at any instant consists of two parts or "sheets"; so that, if a straight line is drawn out from this centre, there will be two points on it, at different distances from the centre, which mark the advance of the disturbance in that direction at any instant. In other words, waves spread out from the centre in such a manner that there are two distinct disturbances advancing with different velocities along any direction. In some doubly refracting substances there is, however, one direction in which these two disturbances have the same velocity, while in all others there are two such directions. This direction in a given body is not a *fixed line* in the substance; for the above statements are true for all lines in the body which have this *direction;* it is called the direction of the "optic axis." Those doubly refracting substances which have only one such axis are called "uniaxal"; the others, which have two axes, are called "biaxal." All crystals which belong to the cubical system, so called, are single refracting; those that belong to the pyramidal or second system are doubly refracting and uniaxal; the other crystals are doubly refracting and biaxal. Any ordinary isotropic transparent substance, such as glass, becomes doubly refracting if it is strained in one direction by pressure, by unequal annealing, etc.

Uniaxal Substances. — In the case of uniaxal doubly refracting bodies, it is found that one of the rays obeys both of the ordinary laws of refraction, while the other in general obeys neither of them. The former is called the "ordinary ray"; the latter, the "extraordinary." In such a substance, then, a centre of disturbance gives rise to a spherical wave front, which accounts for the ordinary ray, and also to another wave front which advances with a different velocity, and which cannot be a sphere; otherwise the extraordinary ray would obey the laws of ordinary refraction, but would have a different index of refraction from the ordinary ray. If a plate of Iceland spar is held between the eye and a bright object, two images of it are seen; and if the plate is turned around an axis perpendicular to its faces, one image will revolve around the other. Huygens suggested that this second wave front was the surface of an ellipsoid; and by his work and that of later investigators, this idea has been confirmed. Since the velocity of both disturbances is the same along the optic axis, these two surfaces, which make up at any instant the wave surface produced by a point source, the sphere and the ellipsoid, must be tangent to each other at the extremities of a diameter having the direction of the axis; and, since the phenomena are symmetrical around this axis, the ellipsoid must be one of revolution around this diameter. The ellipsoid may lie inside the sphere or outside. Plane sections of the two types of wave surface through the axis are shown in the cut.

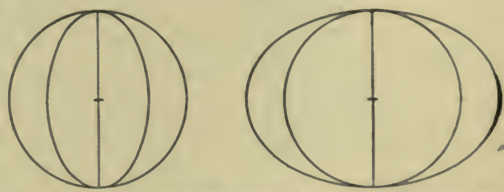


FIG. 282. — Huygens's wave-surfaces for a uniaxal doubly refracting substance.

It is a simple matter of geometry to draw the refracted rays, provided the wave surface is known. Thus, let a uniaxal substance be cut with a plane face making the angle N with the axis, and let plane waves be incident upon this surface in such a direction that the plane of incidence is one that includes the normal to the surface at any point and the direction of the axis at that point — such a plane is called a "principal section" of the crystal. If the incident wave front at any instant is represented in Fig. 283 by $\overline{OO_1}$, waves will spread out from O in the doubly refracting substance and will advance, as shown, for a definite distance, while the disturbance from O_1 reaches A. Following the method of Huygens, as explained in Chapter XX, the two wave fronts in the lower medium produced by the incident plane waves are obtained by drawing through a line perpendicular to the sheet of paper at A two planes which are tangent to the sphere and the ellipsoid of the wave surface around O. The tangent plane to the sphere touches it at C, and $\overline{OC}$, the radius, is perpendicular to it. The tangent plane to the ellipsoid touches it at D, but the perpendicular from O upon the plane is $\overline{OE}$, not $\overline{OD}$. (If the plane of incidence were not a principal section of the crystal, the point of tangency of the ellipsoid would not have been necessarily in the plane of the paper.) The point D marks that point on the wave front $\overline{AED}$ which the disturbance from O has reached at the instant of time considered. Therefore the line $\overline{OD}$ is the *ray;* while the line $\overline{OE}$ is the direction of advance of the plane wave front $\overline{AD}$, it is called the "wave normal." It is seen that the "ordinary ray," $\overline{OC}$, obeys both the laws of refraction; while the "extraordinary" one, $\overline{OD}$, does not, as a rule, obey either. The fact that the extraordinary ray has a direction different from the wave normal is an illustration of what was noted on page 386, that the ordinary case of the ray and the wave-normal being coincident was due to the velocity of the waves in the medium being the same in all directions.

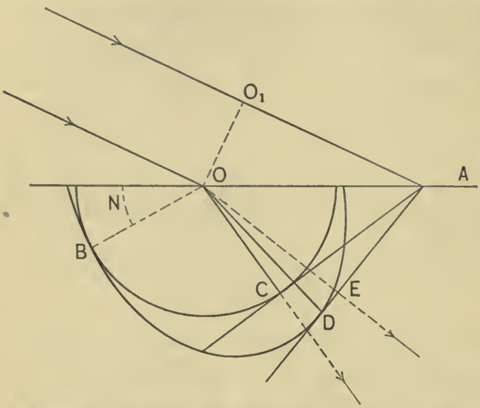


FIG. 283. — Refraction of Iceland-spar. $\overline{OB}$ is the axis; $\overline{OC}$ and $\overline{OE}$ are the normals to the refracted wave fronts; $\overline{OC}$ and $\overline{OD}$ are the ordinary and extraordinary rays.

Separation of the Ordinary and Extraordinary Rays. — It is often important to separate the ordinary and extraordinary rays so that each may be used by itself. There are several methods for doing this:

1. *Prism.* — If the uniaxal substance is made in the form of a prism and placed on a spectrometer table, the two beams of light produced by refraction of the beam from the collimator will have different directions, and may be observed separately. If a glass prism of suitable angle and material is combined with a prism of Iceland spar, the ordinary ray may be made to emerge parallel to its original direction, while the extraordinary ray will be deviated and the dispersion of white light produced by one prism may be largely neutralized. Such a compound prism is called "achromatic."

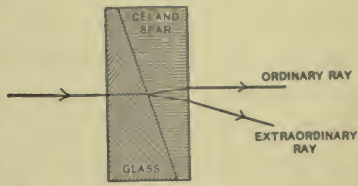


FIG. 284. — An achromatic Iceland spar prism.

2. *Absorption.* — In some uniaxal substances one of the rays is absorbed much more easily than the other. Thus, a *thin* plate of tourmaline transmits both rays; but a thicker one absorbs the *ordinary* and transmits only the *extraordinary*.

3. *Total reflection.* — Since the two rays have different velocities in the crystal, it is evidently possible to select some isotropic substance, the velocity of waves in which is intermediate between these, and to have the ray incident upon a surface separating the latter substance from the doubly refracting one at such an angle that one of the rays will suffer *total reflection* and the other will not. Thus, in Iceland spar the extraordinary ray is refracted less than the ordinary; and Canada balsam, a transparent cement, has an intermediate velocity; so if a piece of Iceland spar is cut into two pieces by an oblique section and these are cemented together by Canada balsam, we have a means of

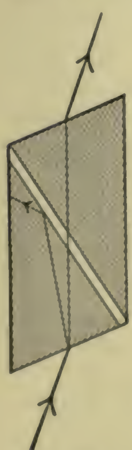


FIG. 285. A Nicol's prism.

separating the two rays. If light from any ordinary source, such as a flame, is incident upon such a compound plate, it will, on entering it, be broken up into two beams, *ordinary* and *extraordinary*, which will have different directions; the former will suffer total reflection at the section of balsam, the latter will be transmitted and will emerge at the end of the plate. The sides of the plate are generally painted black; so that the ordinary rays are all absorbed.

This piece of apparatus was invented by William Nicol of Edinburgh, and is called a "Nicol's prism," or a "nicol."

Biaxal Crystals. — In the case of biaxal crystals, neither of the rays, as a rule, obeys the ordinary laws of refraction. The form of the wave surface which was proposed by Fresnel is the true one so far as is known; but it has not been verified in all particulars. Its properties will be found discussed in any advanced book, such as Preston, *Theory of Light*, or Drude, *Optics*.

CHAPTER XXXIV

POLARIZATION

Huygens's Experiment. — In his investigation of the doubly refracting properties of Iceland spar, Huygens noticed a remarkable fact concerning the two rays transmitted by a plate made of it. They both appeared to be like ordinary beams of light; they could be reflected and refracted; they affected the sense of sight; etc.; yet they were different in one respect, as was shown by an ingenious experiment. The simplest form of this is a slight modification of Huygens's original one.

Two identical prisms are cut out of a piece of Iceland spar, so that the optic axis in each makes the same angle with the normal to the surface. These are made "achromatic" (see page 545), and are mounted in tubes so that each can be turned around a line perpendicular to its faces. (This line may be called the "axis of figure.") If light from a small source falls upon one of these prisms, two pencils emerge: one, the ordinary; the other, the extraordinary. If now the second prism is placed parallel to the first, these two pencils emerging from the latter will fall upon the former; and each will give rise to two pencils, one ordinary, the other extraordinary; thus, four pencils in all will emerge, two ordinary and two extraordinary. This is true in general; but Huygens observed that as the tube containing the second prism was turned around its axis of figure, there were four positions, 90° apart, during a complete revolution of the tube, in which only two pencils emerged; and in one of these positions the two emerging pencils coincide in direction. For an intermediate position between any two of these there are, as said, four pencils. They all appear equally bright for a position halfway between any two consecutive ones of these four positions; but, as the tube is turned, two of these grow feeble and vanish, while the other two grow more intense; then, as the tube is turned farther, these two grow feebler and finally vanish, while two others appear and grow more intense. A rotation of 90° is required to turn from one of these positions into the other. It is thus evident that the two pencils which emerge from the first prism are not like the incident light. Further, they are not like each other. For, calling the two pencils emerging from the first prism O and E; and the two pencils produced in the second prism by the former pencil, OO and OE; and those produced by the latter, EO and EE, we may state the above facts as follows: in general, OO, OE, EO, and EE are present; as the second prism is turned, a position is reached for which only OO and EE appear; as the rotation is continued for 90°, these disappear, and OE and EO only are present; etc. Thus for one position of the second prism, one of the two pencils incident upon it gives rise to an ordinary pencil, while the other produces an extraordinary one; and after a rotation of 90° this condition is reversed. Huygens noticed that when the principal sections of the two prisms were parallel, OO and EE were transmitted; but when these planes were perpendicular to each other, OE and EO were transmitted.

The explanation of all these phenomena is simple if we consider the two rays O and E as *plane polarized with their planes* of vibration at right angles to each other. (See page 313.) The waves are then to be thought of as *transverse*, and the vibrations of any one beam are all in parallel straight lines in a plane at right angles to the direction of propagation of the waves. The first prism breaks up the incident beam of light into two, whose vibrations are in directions perpendicular to each other. The experiment shows that these two directions are fixed in the prism; so that, as it is turned on its axis of figure, the directions of the vibrations of the emerging beams turn also. In other words, at any point in the surface of the prism there are only two directions in which vibrations are possible; and these are fixed in it with reference to the principal section at that point. Vibrations along one line make up the ordinary ray; those along the other the extraordinary. Thus, let C be any point of the prism of Iceland spar; $\overline{PCP}$ be the position of the principal section through it; and $\overline{CA}$ and $\overline{CB}$ be the two directions of possible vibrations, $\overline{CA}$ that of the ordinary, $\overline{CB}$ that of the extraordinary ray. When the two prisms are so placed that their principal sections are parallel, the possible directions of vibrations in the two are parallel; so that, calling the direction of the vibrations of the ordinary ray in the second one $\overline{C_1A_1}$, and that of the vibrations of the extraordinary $\overline{C_1B_1}$, $\overline{CA}$ and $\overline{C_1A_1}$ are parallel, and also $\overline{CB}$ and $\overline{C_1B_1}$. It is thus apparent why OO and EE are transmitted. Again, when the second prism is rotated through 90°, $\overline{CA}$ and $\overline{C_1B_1}$ are parallel, and $\overline{CB}$ and $\overline{C_1A_1}$; so OE and EO are transmitted. If the second prism is turned so that the principal planes are inclined to each other, let the line $\overline{C_1A_1}$ make the angle N with $\overline{CA}$; then a vibration along $\overline{CA}$ having an amplitude $\overline{CP}$ will be resolved on entering the second prism into two vibrations; one along $\overline{C_1A_1}$ having

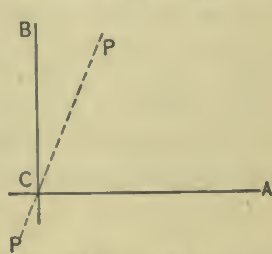


FIG. 286. — Analysis of a vibration by a doubly refracting plate, in which CA and CB are the two possible directions of vibration.

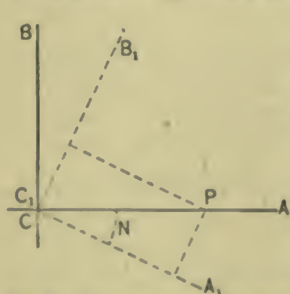


FIG. 287. — Diagram illustrating Huygens's experiment.

an amplitude $\overline{CP}$ cos N, the other along $\overline{C_1B_1}$ with an amplitude $\overline{CP}$ sin N. That is, if the amplitude of O, $\overline{OP}$, is called A, we may write $O = A$, $OO = A \cos N$, $OE = A \sin N$. Similarly, if there is a vibration along $\overline{CB}$ of amplitude A, we have $E = A$, $EO = A \sin N$, $EE = A \cos N$. We must consider the light received from ordinary sources, such as the sun, flames, etc., as being made up of vibrations in all directions in the wave front, because it is not polarized in any way, and when these vibrations are analyzed by the first prism into two sets, along $\overline{CA}$ and $\overline{CB}$, their intensities are equal; that is, their amplitudes are the same. Therefore the four pencils of light transmitted by the second prism have amplitudes given as above:

$$OO = A \cos N;$$

$$OE = A \sin N;$$

$$EO = A \sin N;$$

$$EE = A \cos N.$$

For $\quad N = 0,\ OO = A,\ OE = 0,\ EO = 0,\ EE = A.$

As N increases, OO decreases, OE and EO increase, EE decreases.

For $\quad N = 90°,\ OO = 0,\ OE = A,\ EO = A,\ EE = 0.$

In this manner the phenomena observed by Huygens are all explained; but the hypothesis on which it is based, viz., that *ether waves are transverse*, was not advanced until the early part of the nineteenth century, when it was proposed independently by Young and Fresnel. The credit of explaining the various phenomena of polarization and of defending this hypothesis must be given the latter. Huygens recognized that the only way possible to account for his observations was to assume a two-sided character for the ether waves; but the only waves known to him were the longitudinal air waves which produce sounds, and the idea of transverse waves does not seem to have occurred to him.

Phase Differences. — These emerging rays have different phases, partly because O and E are not necessarily in the same phase at any instant, and also because they take different times to traverse the plates, owing to their different velocities. Thus, if the incident waves are homogeneous, and if V_1 is the velocity of the ordinary waves and V_2 that of the extraordinary ones, and if h is the thickness of a plate, the difference in phase introduced by the plate, expressed in terms of time, between the two emerging rays is $h\left(\frac{1}{v_1}-\frac{1}{v_2}\right)$. The period of the rays is the same, since the light is assumed to be homogeneous; and since an angle 2π corresponds to a time equal to a period T, the difference of phase of the two rays expressed as an angle is $\frac{2\pi}{T}h\left(\frac{1}{v_1}-\frac{1}{v_2}\right)$. But calling the wave lengths of the two rays l_1 and l_2, $v_1=\frac{l_1}{T}$ and $v_2=\frac{l_2}{T}$; so this quantity may be written $2\pi h\left(\frac{1}{l_1}-\frac{1}{l_2}\right)$. (The velocity v_2 depends, of course, upon the *direction* of the waves in the plate, and so is not a constant.) These formulæ will be applied later to explain the colors observed in certain polarization phenomena.

Polarization by Reflection. — From these experiments of Huygens, and from the explanation of them just given, nothing can be said in regard to the connection between the position of the principal section at a point and that of the directions of possible vibrations; but a discovery of Malus, which will be described immediately, led to the proof that one of these directions of vibration lies in the principal section. (Referring to the cut on page 549, this means that $\overline{PCP}$ coincides with either $\overline{CA}$ or $\overline{CB}$. More recent experiments have shown that it coincides with $\overline{CB}$.)

Malus discovered by chance that the light reflected from a glass surface was plane polarized more or less completely, and he observed that, if sunlight (or light from any ordinary source) is incident upon such a mirror at a definite angle, called the "polarizing angle," practically *all* the reflected light is plane polarized, while the transmitted light contains waves that are also plane polarized, but with their vibrations at right angles to the former. Malus made his discovery in 1808 when looking through a plate of Iceland spar at the image of the sun reflected from the window panes of the Luxembourg Palace in Paris; he noticed that the two transmitted pencils were of unequal intensity and that, for certain positions of the crystal, one image vanished while on rotation through 90° the other one disappeared. This showed that the incident light on the crystal was plane polarized. If one looks through such a plate of Iceland spar at light reflected from a plane glass mirror at the polarizing angle, and slowly rotates the plate, it is observed that, when its principal section coincides with the plane of incidence, only the ordinary rays are transmitted; while, if the principal section is at right angles to the plane of incidence, only the extraordinary rays are transmitted. For positions of the plate between these two, both rays are transmitted, but with different intensities except for the position halfway between. The direction of the vibration of the rays reflected from the plane mirror must by symmetry be either in the plane of incidence or at right angles to it, *i.e.* parallel to the plane of the mirror; and so the fact just described proves the statement made above in regard to the connection between the principal section of the Iceland spar plate and the possible directions of vibrations. (For many reasons it is believed that the vibrations in the rays polarized by reflection from a glass plate are parallel to its plane. This is in accord with the statement that the vibrations of the extraordinary rays are in the principal section. Therefore the vibrations transmitted by a Nicol's prism are in the principal section.)

If the light is not reflected at the polarizing angle, only a portion of the reflected light is plane polarized; the rest is like the incident light, made up of rays whose vibrations are in all directions in the wave front. At the polarizing angle more of the *transmitted* light is plane polarized than for any other angle; but, as said above, these vibrations are at right angles to those of the reflected light, as may be shown by viewing it through an Iceland spar plate.

Brewster's Law. — It is found by experiment that light reflected from a plane mirror of any transparent isotropic substance which shows ordinary normal dispersion, such as all kinds of glass, water, etc., may be plane polarized for certain definite polarizing angles. (This is not true of reflection from metallic mirrors or from substances showing anomalous dispersion.) A connection between the polarizing angle of any substance and its index of refraction was established by Brewster. He found that at the polarizing angle the reflected and refracted rays were perpendicular to each other. If, in the cut, MM is the plane surface of a transparent substance, whose index of refraction with reference to the surrounding medium is n, and if BA, AD, and AC are the incident, the refracted, and the reflected rays respectively, experiments show, as just said, that if the angle (BAP) is the polarizing angle, (CAD) is a right angle. Therefore the angle of refraction (DAP') equals (CAM); so, calling the angle of incidence N, $(DAP') = \frac{\pi}{2} - N$. The index of refraction n satisfies the equation $n = \frac{\sin(BAP)}{\sin(DAP')}$; and hence, at the polarizing angle, $n = \frac{\sin N}{\cos N} = \tan N$; or, the angle of polarization of any substance is such that its tangent equals the index of refraction. This is known as "Brewster's

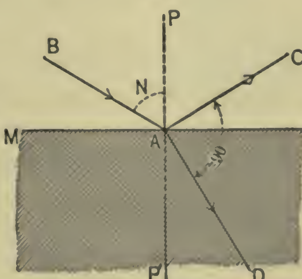


FIG. 288. — Brewster's Law in regard to the polarizing angle.

Law." (More careful experiments have shown that there is no angle of incidence for which *all* the reflected light is plane polarized; in reality there is always a small amount not so polarized, but at the polarizing angle this is small.). The value of the polarizing angle of pure water is 53° 11′; of crown glass, about 57°; etc.

Plane of Polarization. — The light which is plane polarized by reflection from a plane transparent surface is by definition said to be "polarized in the plane of incidence"; or its "plane of polarization" is said to be that of the plane of incidence. (This definition is entirely independent of any conception of the *directions of the vibrations* of the rays in this reflected beam; but, accepting the statements made above in regard to these directions, it is seen that in a plane polarized beam the direction of vibration is in the wave front and at right angles to the plane of polarization.)

Pile of Plates. — A means is obviously offered of securing plane polarized waves by reflecting ordinary sunlight or light from a flame, etc., from a glass or water surface at the polarizing angle. In general, only a small quantity of light is reflected owing to the poor reflecting power of glass or water; but the effect can be increased greatly in the case of glass by using several thin plates placed one on top of the other, thus forming a "pile of plates." When light falls upon such a pile at the polarizing angle, the reflected light is plane polarized, but part of the refracted light is not; this falls at the polarizing angle upon the surface where the top plate meets the next one, is partially reflected, and is refracted out so as to coincide in direction with the beam reflected from the plate, being polarized also like it. The light which enters the second plate is still not completely polarized, and when this meets the third plate, reflection again occurs; etc. Thus both the light which is reflected and that which is transmitted by a pile of plates is practically completely plane polarized, but in planes at right angles to each other.

If plane waves plane polarized in the plane of incidence are incident upon a glass surface at the polarizing angle, they will be entirely reflected; while, if they are polarized in a plane at right angles to this, they will be entirely transmitted. These and all the facts in regard to polarization by reflection may be shown best by the use of a piece of apparatus invented by Nörrenberg, which is shown in the cut. It consists essentially, as is seen, of two plane mirrors which may be turned about axes lying in their planes and also in such a manner that their normals may lie in different planes.

FIG. 289.—Nörrenberg's apparatus.

Plane Polarization. — We have, therefore, two general methods for the production of plane polarized light; one is to use a crystal of Iceland spar, or other uniaxal crystal, and get rid of one of the rays by the means described on page 544; the other is to use a pile of plates at the polarizing angle. In practice a Nicol's prism is almost invariably used.

Similarly, either one of these methods may be used to test whether a certain beam of light is plane polarized; thus, if such a beam is incident upon a Nicol's prism, it will be entirely extinguished for some position of the latter, as it is slowly turned on its axis of figure; and when it is extinguished, it is known that its vibrations are at right angles to the principal section of the prism.

In this manner it may be shown that, if one looks at the blue sky in a direction at right angles to a line joining the eye to the sun, the reflected sunlight — that which is scattered by the fine particles (see page 430) — is plane polarized in the plane including the two directions just mentioned. We have thus another means of securing plane polarized light. (It is possible to predict from theory what the direction of the vibration is in this case; and, by comparing this direction with the observed position of the plane of polarization, it is proved that the vibrations are at right angles to it. (See page 554.)

Interference of Plane Polarized Waves. — Fresnel and Arago performed by means of two piles of plates a most ingenious experiment to determine whether the vibrations in the waves transmitted by them were *exactly* in planes at right angles to the direction of propagation of the light. If such is the case, by using independently the two piles with their planes of incidence at right angles to each other, two beams of light may be secured in which the vibrations are in the wave fronts but are perpendicular to each other. Two such trains of waves as this cannot "interfere"; because, in order to have one train interfere with another, the vibrations of both must be in the same straight line. The experiment referred to consisted in modifying Young's original interference one by introducing a pile of plates in front of each of the two slits. We shall quote from their own description, following Crew's translation in his *Memoirs on the Wave Theory of Light:*

"It has been known for a long time that if one cuts two very narrow slits close together in a thin screen and illuminates them by a single luminous point, there will be produced behind the screen a series of bright bands resulting from the meeting of the rays passing through the right-hand slit with those passing through the left. In order to polarize at right angles the rays passing through these two apertures, . . . we selected fifteen plates as clear as possible and superposed them. This pile was next cut in two by use of a sharp tool. So that now we had two piles of plates of *almost exactly* the same thickness, at least in those parts bordering on the line of bisection; and this would be true even if the component plates had been perceptibly wedge shaped. The light transmitted by these plates was almost completely polarized when the angle of incidence was about thirty degrees. And it was exactly at this angle of incidence that the plates were inclined when they were placed in front of the slits in the copper screen.

"When the two planes of incidence were parallel, *i.e.* when the plates were inclined in the same direction; — up and down, for instance, — one could very distinctly see the interference bands produced by the two polarized pencils. In fact, they behave exactly as two rays of ordinary light. But if one of the piles be rotated about the incident ray until the two planes of incidence are at right angles to each other, the first pile, say, remaining inclined up and down while the second is inclined from right to left, then the two emergent pencils will be polarized at right angles to each other and will not, on meeting, produce any interference bands. . . . We must conclude that rays of light polarized at right angles do not affect one another."

Fresnel and Arago performed another most interesting experiment which is also described in the memoir from which we have just quoted. Its object was to determine the conditions under which interference may occur when two plane polarized waves with their vibrations parallel are brought together. It incidentally explains the production of the beautiful colors which are seen when a thin plate made of a crystal or of any doubly refracting substance is interposed between two nicols whose principal sections are not parallel and which are placed in line so that one may look through them both at a source of white light. In any such plate, whether uniaxal or biaxal, there are only two possible directions which the vibrations of the transmitted waves may have. Let these two directions at a point O of the plate be $\overline{OA}$ and $\overline{OB}$; and let the vibrations of the waves incident

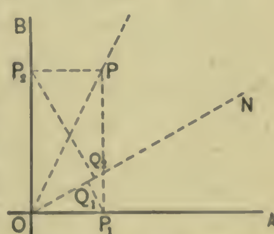


FIG. 290. — Diagram illustrating the experiment of Arago and Fresnel.

upon it from the first nicol have the direction and the amplitude $\overline{OP}$. This is resolved into two, a vibration along $\overline{OA}$ of amplitude $\overline{OP}_1$, and one along $\overline{OB}$ of amplitude $\overline{OP}_2$, where these lines are the components of $\overline{OP}$ in these directions; and these two trains of waves are transmitted by the plate. Let us consider the waves as homogeneous, so that they all have the same period. As they enter the plate they are in the same phase because they are components of the same vibration; but they will emerge with a difference in phase, expressed in time by $h\left(\frac{1}{v_1}-\frac{1}{v_2}\right)$, where h is the thickness of the plate and v_1 and v_2 are the velocities of the two waves. They will then fall upon the second nicol whose principal section may have the direction $\overline{ON}$. The waves whose vibrations have the direction and amplitude $\overline{OP}_1$ will be resolved into two, but the only one transmitted is that whose vibrations have the direction and amplitude $\overline{OQ}_1$, where this is the projection of $\overline{OP}_1$ on $\overline{ON}$. Similarly, the other train of waves whose vibrations are represented by $\overline{OP}_2$ will be resolved into two, but the only one transmitted is that whose vibrations are represented by $\overline{OQ}_2$, the projection of $\overline{OP}_2$ on $\overline{ON}$. Therefore, two trains of waves emerge from the second nicol, which are plane polarized with their vibrations in the same direction, and which have a difference in phase of $h\left(\frac{1}{v_1}-\frac{1}{v_2}\right)$. This expression may be simplified if the values of v_1 and v_2 in terms of the period and wave length are substituted, viz., $v_1=\frac{l_1}{T}$, $v_2=\frac{l_2}{T}$; hence the difference in phase may be written $Th\left(\frac{1}{l_1}-\frac{1}{l_2}\right)$. Expressed in terms of an angle this is $2\,\pi h\left(\frac{1}{l_1}-\frac{1}{l_2}\right)$.

If N_1 is the angle between $\overline{OP}$ and $\overline{OA}$,

$$\overline{OP}_1 = \overline{OP}\cos N_1,$$

$$\overline{OP}_2 = \overline{OP}\sin N_1.$$

And if N_2 is the angle between $\overline{ON}$ and $\overline{OA}$,

$$\overline{OQ}_1 = \overline{OP}_1 \cos N_2 = \overline{OP} \cos N_1 \cos N_2,$$

$$\overline{OQ}_2 = \overline{OP}_2 \sin N_2 = \overline{OP} \sin N_1 \sin N_2.$$

So if the original vibration as it leaves the first nicol is $A \cos nt$, and if this difference of phase introduced by the plate is d, the resulting vibrations as they leave the second nicol may be written

$$A \cos N_1 \cos N_2 \cos nt,$$

$$A \sin N_1 \sin N_2 \cos (nt - d).$$

If $\overline{ON}$ is perpendicular to $\overline{OP}$, the nicols are said to be "crossed," and no light is transmitted if there is no plate interposed between them. If this is the case, in the above formula $N_2 = -(90° - N_1)$; so $\cos N_2 = \sin N_1$, and $\sin N_2 = -\cos N_1$; and the two vibrations are $A \cos N_1 \sin N_1 \cos nt$ and $-A \cos N_1 \sin N_2 \cos (nt - d)$. This last vibration may be written

$$A \cos N_1 \sin N_2 \cos (nt - d - \pi).$$

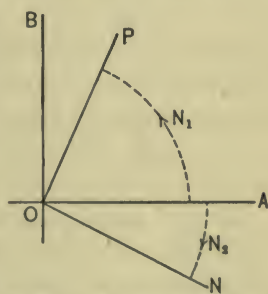


FIG. 291. — Special case of crossed nicols.

So the difference in phase is $d + \pi$; and the amplitudes are equal. If this quantity $d + \pi$ is equivalent to an odd number of half periods, *i.e.* if $d + \pi = (2m + 1)\pi$, or $d = 2m\pi$, the interference will be complete; and these homogeneous waves will be extinguished. (They suffer total reflection at the surface where the two halves of the last nicol are cemented together.) Other cases, in which the two nicols are not crossed, may be found discussed in advanced text-books.

So, if white light is used, certain waves will be absent in the transmitted light, and it will be colored. Which particular waves are absent depends upon the thickness of the double refracting plate, as is evident from the formula.

The first nicol, which polarizes the incident light, is called the "polarizer"; the second one is called the "analyser."

If there is no such plate between the crossed nicols, no light at all is transmitted; but, if a plate is introduced, certain waves appear, as just explained. The phenomenon is called "depolarization"; and the experiment serves as an extremely delicate test of the double refraction of a substance.

It at first sight appears as if, in the above experiment, the first nicol might be removed so that the light would fall directly upon the double refracting plate and then upon the second nicol, and there might still be interference; for the plate would break up the light into two beams and introduce a difference of phase between them before they were combined again by the second nicol. But the relations between the amplitudes and the phases in this case are not definite, because the light incident upon the plate is not polarized, but consists of vibrations in all directions; and so there is no permanent interference. This fact is the fundamental one established by the experiments of Fresnel and Arago.

Circular and Elliptical Polarization.—The fact that a plate of a doubly refracting substance breaks up an incident beam of plane polarized light into two such beams, polarized at right angles to each other and with a difference in phase between them which varies directly with the thickness of the plate, renders it possible to secure a circular or an elliptical vibration. A beam of light whose vibrations are of this character is said to be "circularly" or "elliptically" polarized.

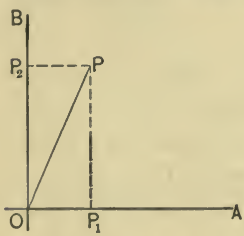


FIG. 292.—Formation of circularly and elliptically polarized light.

Thus let $\overline{OA}$ and $\overline{OB}$ be the directions of possible vibrations in the plate, and let $\overline{OP}$ be the direction of the principal section of the nicol through which the light is incident upon the plate.

If the amplitude of this plane polarized light is $\overline{OP}$, those of the two beams transmitted by the plate are $\overline{OP}_1$ and $\overline{OP}_2$. Whatever the difference of phase between them introduced by the plate, these two vibrations will combine to form an elliptical one. (See page 324.) If this difference in phase is equivalent to a quarter of a period, or to any odd number of quarter periods, the vibration is an ellipse whose axes coincide in direction with the lines $\overline{OA}$ and $\overline{OB}$. If, in addition to this condition for the thickness of the plate being satisfied, the incident vibration $\overline{OP}$ bisects the angle between $\overline{OA}$ and $\overline{OB}$, the *amplitudes* of the two transmitted beams will be the same; and they will combine to form a circular vibration. Such a plate is called a "quarter wave plate"; and obviously plates of different thicknesses must be used for waves of different wave length or color.

Fresnel's Rhomb. — If plane polarized light is *totally* reflected from the surface of a transparent substance such as glass or water, it becomes, in general, elliptically polarized; for, if the light is plane polarized in such a manner that the direction of the vibration is neither parallel to the surface nor in the plane of incidence, it is resolved by reflection into two plane polarized beams, one with its vibrations parallel to the surface, the other with its vibration in the plane of incidence. Their amplitudes are different, unless the direction of vibration in the incident beam bisects the angle between two lines in the wave front, one parallel to the surface, the other in the plane of incidence. A change of phase is introduced by the reflection, which is not the same for the two beams; and the difference for the two depends upon the angle of incidence and the material at whose surface the reflection takes place. Fresnel made a rhomb of such a particular kind of glass

FIG. 298. — Fresnel's rhomb.

and with such angles that when light was incident perpendicularly upon one of its end faces, it would suffer total reflection twice and emerge perpendicular to the opposite face with a difference of phase equivalent to a quarter of a period between the two plane polarized beams. By this means it is possible to obtain light circularly polarized, or elliptically polarized with one axis in the plane of incidence and the other at right angles to it.

Detection of circularly or elliptically Polarized Light. — If circularly polarized light is incident upon a Fresnel's rhomb or upon a quarter wave plate, it will emerge plane polarized, for the effect of these pieces of apparatus is to introduce a difference of phase of a quarter of a period between the two component plane polarized waves into which the incident waves are resolved. The existence of this plane polarized light may be detected by a nicol.

If elliptically polarized light is passed through a Fresnel's rhomb or a quarter wave plate, it will, in general, emerge elliptically polarized; but, if the plate is turned in its own plane, or the rhomb is turned around an axis perpendicular to the planes of its end faces, there will be four positions in one complete revolution for which this light will be plane polarized. This may be detected by a nicol. Other methods are described in advanced text-books.

Ordinary Light. — The light which we receive from ordinary sources, such as diffused sunlight, etc., is not polarized in any manner; yet it can be transformed into plane, circularly, or elliptically polarized light by methods which have been discussed. When ordinary light is passed through a doubly refracting substance, both the transmitted plane polarized beams are of equal intensity, and there is no permanent phase relation between them. This shows that we must consider ordinary light as due to transverse waves in which the vibration at any instant may be rectilinear or circular, etc., but in which the vibration is continually changing its form. We may regard it, then, as due to an ever changing mixture of transverse rectilinear vibrations.

Rotation of the Plane of Polarization. — There are certain substances which have a most remarkable property in regard to plane polarized light. If one of them is made in the form of a plate and a beam of homogeneous light plane polarized in a particular direction is transmitted through it, the emerging light is plane polarized, but its plane of polarization has been rotated through a certain angle. Such substances are said to be "optically active." Thus, if the incident light is produced by the use of a nicol, and a second nicol is "crossed" with it, no light passes before the "active" substance is introduced between them; but after this is done, the second nicol must be turned on its axis of figure through a definite angle before the light is again extinguished. This angle varies directly with the thickness of the substance, and is different for waves of different wave lengths, being much greater for the short waves than for the long ones. This last phenomenon is called "rotatory dispersion."

There are two classes of these substances; one is made up of bodies which are naturally active, while the other contains bodies which are active only when they are under the influence of a magnetic force. The former phenomenon was discovered by Biot; the latter, by Faraday. A kinematic explanation of this rotation was given by Fresnel; but it is not necessary to state it here. It may be found in any advanced treatise.

a. Naturally active bodies. — Examples of these bodies are quartz when cut at right angles to its optic axis, an aqueous solution of certain tartaric acids, of many of the sugars, etc. In all these cases, if the plane polarized light is made to pass through a plate and then by means of a mirror is reflected back again, the plane of polarization is rotated first in one direction and then in the opposite, so it emerges the second time polarized exactly as it was on incidence;

it is as if one screwed a screw into a board and then unscrewed it.

In certain bodies the plane of polarization is rotated in a right-handed direction, while in others it is turned in the opposite sense. Thus, if the light is emerging in a direction perpendicularly *up* from the paper, and if $\overline{AB}$ is the direction of the principal section of the second nicol, in the experiment described above, when it is so placed as to extinguish the light before the active substance is introduced; and, if after this takes place, the nicol must be turned in the direction shown by the arrow in order to extinguish the light again, the rotation is said to be "right-handed." If, on the other hand, the rotation of the nicol must be in the opposite direction, it is called "left-handed." There are two varieties of quartz, left-handed and right-handed; two varieties of active tartaric acids, etc.

Fig. 294. — Rotation of the plane of polarization.

It was discovered by Pasteur that all optically active substances were made up entirely or in part of certain crystals which had a "hemihedral" form; of which there are for any substance two possible states. These two are symmetrical with reference to a plane, like the two hands of any individual. Thus right-handed quartz has imbedded in it minute hemihedral crystals of one form; while left-handed quartz has hemihedral crystals of the symmetrical form, etc. Crystals of tartaric acid are hemihedral; and when dissolved in water the molecules retain their asymmetric character. These facts are the basis of what is called "stereochemistry," which is a branch of chemistry dealing with conceptions of the atomic arrangement in certain organic molecules. (See Richardson, *Foundations of Stereochemistry*, New York.)

b. Magnetically active substances. — When any transparent substance, such as glass, is placed in an intense magnetic field, it acquires the power of rotating the plane of polarization; but this rotation is different from that just described, because, if the rotated light is reflected back on its path, the plane is rotated still farther, in the same direction as before; it is not turned back into its previous position.

When the subject of electricity is discussed, it will be shown that if an electric current is passed through a wire wound in a helix, there is a strong magnetic field inside it, and that the direction of this field is reversed if that of the current is reversed. It is found by experiment that, if a piece of transparent matter is introduced in this helix, the direction of the rotation of the plane of polarization is that of the electric current in the helix.

Metallic Reflection. — When plane polarized light falls upon a polished metal surface, it is reflected according to the ordinary laws; but the light is elliptically polarized, unless the incidence is normal. This is owing to the fact that the incident light is broken up into two plane polarized beams which have a difference in phase.

If the metal surface is magnetic, that is, if it is made of iron, steel, nickel, etc., the character of the reflected light — the shape of the resulting ellipse — depends upon whether the metal is in its natural condition or is magnetized; if the latter is the case, the effect of the reflection also varies with the direction and the intensity of the magnetization. This is known as the "Kerr effect," and will be found fully described in any advanced text-book.

CHAPTER XXXV

VELOCITY OF LIGHT

THE first experiments to determine whether, like sound, light traveled with a measurable velocity were performed by Galileo. They consisted in having one observer flash a light which was seen by a second one at a considerable distance, who then flashed another light as quickly as possible, and this was seen and noted by the first observer. In any case there would necessarily be an interval between the instant when the first observer flashed his light and when he saw that flashed by the second one, owing to the time required to perform the manipulation, but, if time were required for the passage of the light across the space between the two observers, this interval would vary directly with the distance between the observers. No such effect was observed. The interval of time between the events referred to was apparently independent of the distance apart of the observers, and was conditioned only by their quickness of motion and perception. It was therefore concluded that light traveled with an infinite velocity.

Method of Roemer. — This opinion was maintained by every one until the year 1676, when Roemer, a Danish astronomer, then living in Paris, made certain observations on the eclipses of one of the satellites of Jupiter by the planet, which he interpreted as proving that the velocity of light was finite, but very great. As a satellite revolves around its planet — *e.g.* the moon around the earth — its motion is periodic, or may be assumed to be so to a very high degree of accuracy. So the interval of time which elapses between two consecutive instants of disappearance of a satellite of Jupiter behind the edge of the planet, when viewed from a point in space fixed with reference to the planet, must be a constant quantity. Roemer observed that this interval of time, as noted here on the earth, was not, however, the same when the earth in its motion around the sun was moving away from Jupiter, as when it was approaching it; it was longer in the former case than in the latter. Further, if this period of revolution of the satellite was noted when the earth was nearest Jupiter, and if a calculation was made, based on this, of the instant at which at the end of half a year, when the earth was farthest from Jupiter, an eclipse should take place, there was found to be a difference of 996 sec. between this calculated instant and the observed one. Roemer saw that these facts could all be explained if the assumption were made that it takes time for the propagation of light across space. As the earth is receding from Jupiter, light has to travel a greater distance at the instant of the second eclipse than at that of the first one; and so the apparent period of the satellite is greater than it would be if the earth were at rest with reference to Jupiter. Just the reverse is true when the earth is approaching the planet. So, when the earth is farthest from Jupiter, light has to travel an additional distance equal to the diameter of the earth's orbit; and the interval of 996 sec. observed, as stated above, between the calculated instant of an eclipse and the recorded one is the time required for light to pass over this distance; and so the velocity of light may be determined.

More recent observations have given 1002 sec. as the interval of time measured by Roemer; and the diameter of the earth's orbit may be taken as 2993×10^6 Km., as will be shown immediately; so the velocity of light in space given by this method is this quantity divided by 1003, or 2.984×10^{10} cm. per second.

The angle subtended at the sun by the radius of the earth, N in the diagram, is called the "solar parallax"; and it is a quantity whose value may be determined by astronomical observations with a high degree of accuracy. The accepted value of this constant at the present time is 8.79 sec. of arc. The radius of the earth is known, its value being 6378 Km. So the distance from the earth to the sun may be calculated. Referring to the cut, $\overline{AE} = 6.378 \times 10^8$ cm., and $N = 8.79$ sec. of arc. But 2π in angular measure $= 360$ degrees of arc $= 360 \times 60 \times 60$ sec.; so 1 sec. $= \frac{2\pi}{360 \times 60 \times 60}$, and 8.79 sec. $= \frac{2\pi \times 8.79}{360 \times 60 \times 60}$. Therefore, since N is so small, we may write $\overline{AE} = \overline{ES} \times N$, or

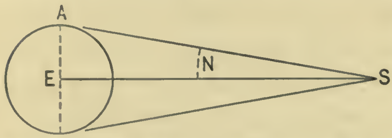


FIG. 295. — Solar parallax: S is the sun; E is the earth.

$$\overline{ES} = \frac{\overline{AE}}{N} = \frac{6.378 \times 10^8 \times 360 \times 60 \times 60}{2\pi \times 8.79} = 1.4966 \times 10^{13} \text{ cm.}$$

The *diameter* of the earth's orbit is, then, twice this, or 2.993×10^{13} cm.

Method of Bradley. — Another method by which the velocity of light could be determined was discovered by Bradley, the great English astronomer, in 1727. He had observed that if a fixed star was observed through a telescope, this instrument had to be pointed in slightly different directions at different times of the year; so that, if the image of the star were kept on the cross hairs of the telescope during the whole year, the instrument had to be kept in continual motion in such a manner that its end described a small curve. The explanation given by Bradley was extremely simple. Consider a long tube closed at its two ends by caps in which there are two openings directly opposite each other. A particle entering at one opening, with a motion parallel to the axis of the tube, will escape through the opening at the other end, *if the tube is at rest.* But, if the tube is moving at right

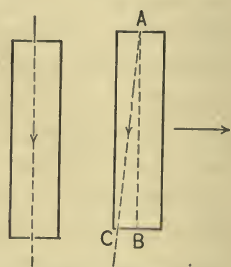


FIG. 296. — Diagram representing stellar aberration.

angles to its length, the opening in the opposite end from the one at which the particle enters must be displaced in a direction opposite to the motion of the tube if the particle is to escape. The line of motion of the particle, with reference to the tube, makes an angle with the axis of the tube whose tangent equals the ratio of the velocity of the tube to that of the particle; for, referring to the cut, if V_1 is the velocity of the tube and V_2 that of the particle, and if t is the time taken for the particle to pass through the tube, $\overline{CB} = V_1 t$ and $\overline{AB} = V_2 t$; so $\tan(BAC) = \frac{CB}{AB} = \frac{V_1}{V_2}$. Similarly, if we consider light from a fixed star entering a telescope, if the instrument is at rest, the light will emerge directly from its farther end; but, if the telescope is moving at right angles to the direction from which the light is coming, it must be inclined forward in order to see the star, and the reason for this, according to Bradley, is because the "path of the light" is along the line $\overline{AC}$. The angle through which the telescope has to be turned when it is pointed approximately perpendicular to the path of the earth in its orbit, is called the "constant of aberration." (Its value according to recent astronomical observations is slightly less than 20.5 sec. of arc.) If this angle is determined by observations on the fixed stars, the velocity of light may be calculated, assuming that in the above formula V_1 is the velocity of the earth in its orbit and V_2 that of light, because the velocity of the earth in its orbit is known. The value thus deduced is 2.982×10^{10} cm. per second.

This explanation of stellar aberration is insufficient to account for all the facts. In the above formula V_2 would be the velocity of light *inside the telescope;* therefore, if the telescope tube is filled with water, V_2 is diminished in the ratio of the index of refraction of water to air, and so the aberration angle should be increased. This experiment was actually performed by Sir George Airy;

and no change in the angle was observed. When we consider light as due to waves in the ether, and if we assume that the ether as a medium does not move as the earth travels through it in its orbital motion, the same formula, as given above, may be deduced for the aberration angle, only in it V_2 *is the velocity of light in the pure ether*, not when it is inside matter.

Method of Fizeau. — The first method for measuring the velocity of light directly here on the earth, without making use of any astronomical data, was devised and applied by Fizeau in 1849. The principle is extremely simple. A source of light is placed in such a manner that it shines

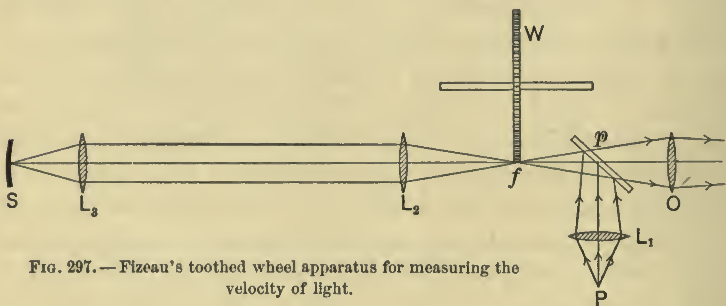


FIG. 297. — Fizeau's toothed wheel apparatus for measuring the velocity of light.

through the space between two teeth of a cogwheel, which may be driven at a high speed. At some distance on the other side of this wheel is a mirror which reflects the light it receives back toward the source of light. If then the cogwheel is rotating so rapidly that the light which passes between any two teeth travels to the mirror and is reflected back at such an interval of time that the wheel has turned through a distance which brings a tooth to where an opening was, no light passes back through the wheel. If the speed of the wheel is increased, an opening will come where the previous one was, and the return light will pass through; if the speed of the wheel is increased still more, the light will be again shut off; etc. If the speed of the wheel, and the distance over which the light travels are measured, the velocity of light can be calculated.

The actual arrangement of the apparatus is shown in the cut. W is the revolving toothed wheel; P is the source of light; p is a plate of glass which partially reflects the light from P into the direction of the wheel; L_1 is a converging lens which focuses the light from P at f, a point at the edge of the wheel; L_2 is a converging lens so placed that f is its principal focus; L_3 is another lens placed at a considerable distance from L_2, but parallel to it, with its principal focus at the middle point of the concave mirror S, whose centre of curvature is at the centre of the lens L_3. The waves from P are thus converged at f, whence they diverge and are made plane by L_2; they are converged again by L_3, and are reflected back on their path, until they reach the glass plate p by which they are in part transmitted; so an observer with an eye-piece O focused on f can determine when the reflected light passes through an opening and when it is shut off.

If the number of revolutions of the wheel per second is n, its angular velocity is $2\pi n$. If there are N teeth in the wheel, the angular distance from one edge of one tooth to the corresponding edge of the next one is $\frac{2\pi}{N}$; and so the angular distance from the middle point of one opening to the middle point of a tooth is $\frac{\pi}{N}$. The time required to turn through this angle is $\frac{\pi}{N} \cdot \frac{1}{2\pi n} = \frac{1}{2nN}$. If the distance from the wheel to the distant mirror is D (or what is practically the same, the distance from L_2 to L_3), the path of the light is $2D$. So, if the angular speed $2\pi n$ is that which corresponds to the first obscuration of the light, it has traversed a distance $2D$ in the time $\frac{1}{2nN}$; and the velocity of light is, then, $2D$ divided by $\frac{1}{2nN}$, or $4nND$.

In one of Fizeau's original experiments, the distance from L_2 to L_3 was 8.633 Km. or 8.633×10^5 cm.; the toothed wheel had 720 teeth; and it was found that the first obscuration of the reflected light occurred when the wheel was making 12.6 revolutions per second. Therefore, in the above formula $n = 12.6$, $N = 720$, $D = 8.633 \times 10^{10}$, and so the velocity of light is determined to be 3.13×10^{10} cm. per second. The great experimental difficulty is to maintain a uniform speed of the wheel and to measure it accurately. This method has been used in more recent years by Cornu and Perrotin. In the work of the latter, the distance apart of the lenses L_2 and L_3 was about 12 Km. The results of these experiments is to give 2.99820×10^{10} cm. per second as the velocity of light in air.

The object of using a concave mirror at S with its centre of curvature at the centre of L_3, instead of a plane mirror, is apparent if it is remembered that the source of light at P is not a point, but is extended. So the waves from a point near P are converged at f, and are made plane by the lens L_2; but their line of propagation is inclined slightly to that of the waves from P, and they are converged by the lens L_3 upon a point of the mirror S, a short distance away from its middle point. If the mirror were plane, these waves would then be reflected off one side, and so would not return through the lens L_3; but, since the mirror is concave, with its centre of curvature at the centre of L_3, these incident waves are reflected back through the lens and finally reach the wheel. Therefore the reflected light is brighter than it would otherwise be.

Method of Foucault. — Another method was suggested by Arago, but was first put in practical use by Foucault in 1850, and is always called by his name. (Fizeau also made some valuable suggestions in regard to it.) It consists of making use of a rotating plane mirror in the following manner: Referring to Fig. 298, there is a source of light at P; a plate of glass at p; a revolving mirror at m, whose axis of rotation is perpendicular to the plane of the paper; a converging lens L, which focuses upon a concave mirror S the light from P reflected at m; the centre of curvature of this mirror S is at m. For a suitable position of the revolving mirror the light from the source P will be focused at the middle point of the concave mirror, and will be reflected back on its path until it reaches the glass plate, when it will be in part reflected and will form an image at a point P'. The revolving mirror is, of necessity, small, almost linear; and so, in order to collect more light from the source P, the mirror S is concave. (Another method, adopted later, was to place the lens L at a distance from m equal to its focal length; and in this case the mirror S may be plane.) If, then, a reflected image of P is formed at P' when the revolving mirror is at rest, a displaced image P'' will be formed when the mirror is turning rapidly, because in the time taken for the waves to pass from m to S and back again, the mirror m will have turned a short distance, and so a reflected ray coming from S will have an angle of incidence upon m different from that which it would have if the mirror had not moved. This ray, after reflection from m, will have a different path from the incident ray.

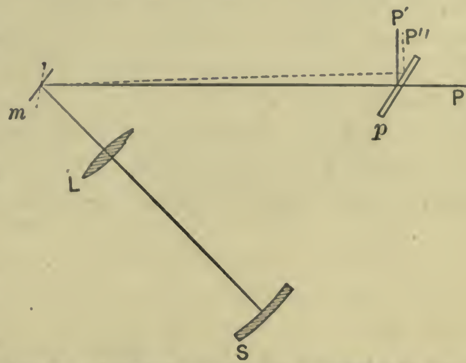


FIG. 298. — Foucault's revolving mirror apparatus for measuring the velocity of light.

If the mirror is making n revolutions per second, its angular velocity is $2\pi n$, and that of the reflected ray is $4\pi n$. (See page 445.) If the distance from m to S is D, if that from m to P is r, and if v is the velocity of light, $\frac{2D}{v}$ is the time required for the light to pass from m to S and back again; in this time the mirror will have turned through an angle $\frac{2D}{v} \cdot 2\pi n$, and the reflected ray through $\frac{2D}{v} \cdot 4\pi n$; therefore the image of P will be displaced by a distance r times this. So, calling $\overline{P'P''}$, x,

$$x = \frac{8\pi Dnr}{v}, \text{ or } v = \frac{8\pi Dnr}{x}.$$

Foucault increased the effective distance D by having the light reflected several times back and forth between five mirrors before it was finally returned to the revolving mirror; but in no case did he obtain a very large displacement $\overline{P'P''}$. Michelson, however, by changing the arrangement of the apparatus was able to increase D to 600 m.; and even when the mirror was turning at the moderate speed of 200 revolutions per second, he obtained a displacement of 13 cm. This method was improved still more by Newcomb, who

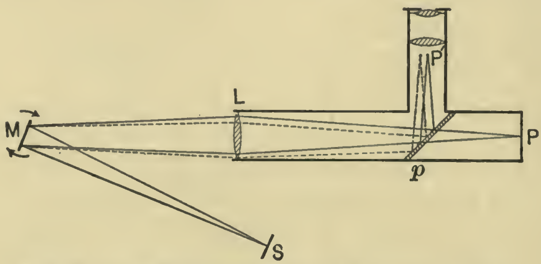


FIG. 298 *a*.—Michelson's modification of Foucault's apparatus.

operated over a distance of 3721 m. The final result obtained for the velocity of light in air by this method was 2.999778×10^{10} cm. per second.

The mean of all the best values for the velocity of light in the ether is 2.999880×10^{10} cm. per second, with a probable error of about 20 Km. It should be noted that the values of this velocity obtained directly by the methods of Fizeau and Foucault are for the velocity in air; and, since the index of refraction of air is 1.00029, the velocity in the pure ether is greater in this ratio. The figure given above for the final value is corrected in this manner so as to apply to the pure ether.

Velocity of Waves of Different Periods. — In the pure ether of interstellar space all ether waves, of whatever period, travel with the same velocity so far as is known, as is shown by the fact that the color of any one of Jupiter's satellites is the same to our eyes when we observe it as it goes into eclipse and as it emerges. If the short waves traveled faster than the longer ones, the satellite would appear red as it disappeared and blue as it reappeared; and the converse would be true if the long waves traveled more rapidly.

In ordinary transparent matter, however, not alone do all waves travel more slowly than in the pure ether, but the waves of different periods have different velocities. It is this which explains refraction and dispersion. Foucault showed by direct experiment that the velocity of light was less in water than in air, by placing a long tube of water in his apparatus immediately in front of the concave mirror. (Actually he used two concave mirrors, one for the light passing in air, the other for light passing in water, and thus obtained two displaced images.) Michelson showed the same for water and for carbon bisulphide; and he also proved directly that in these substances red light travels more rapidly than blue.

CHAPTER XXXVI

RADIATION AND ABSORPTION SPECTRA

Discovery by Newton of Nature of White Light. In the year 1672 Newton made the interesting discovery that when sunlight was admitted into a darkened room through a small opening and was allowed to traverse a glass prism, the transmitted light was no longer white, but consisted of beams of different colors, each color having a different refrangibility and therefore a direction differing slightly from that of its neighbors. He recognized as distinct colors violet, indigo, blue, green, yellow, orange, and red; but all other intermediate shades were present also. He performed further the reverse experiment of combining these colors by means of a second prism, and produced white light again. He also showed that it was impossible by means of a second prism to break up any of these spectrum colors into parts. These observations prove that white light is due to a combination of simple elementary causes; and we know from Young's experiments that these are trains of waves of definite wave lengths, each train being characteristic of a definite color if it is perceived by the eye.

These experiments of Newton form the basis of our explanation of the color of natural objects and of the science of spectrum analysis. Various bodies in the universe are emitting light (or, more generally, all bodies are emitting ether waves); all bodies reflect light (or ether waves) to a greater or less extent; so, if we look at any object, the light (or ether waves) which we receive is due to various causes. We can analyze this radiation into its component parts by means of suitable dispersive apparatus, and can then detect these separate trains of waves by proper means. We shall consider in this chapter (1) methods of producing ether waves, especially those which appeal to our sense of sight; (2) different forms of dispersive apparatus and different modes of recognizing trains of waves of different wave length; (3) the results of the examination of the radiations from different sources.

Sources and Cause of Radiation

Radiation owing to Temperature. — All substances in the universe are, so far as known to us, emitting ether waves, owing to the vibrations of certain parts inside their molecules. If the body when placed in a darkened room can be seen by the eye, it is said to give off light. The ordinary method of making a body luminous is to raise its temperature; thus a body may be exposed to a hot flame, such as one from a Bunsen burner, or it may be placed in the poles of an electric arc light. (See page 665.) At such high temperatures, many bodies are vaporized, and their vapors are then at this temperature. The laws of radiation due to this cause have been discussed in Chapter XIV.

Electro-luminescence. — Again, if an electric spark is made to pass between two metal points, they are vaporized and the vapors are luminous; not, however, owing entirely to the temperature being raised. (The same statement is true of the luminosity of the vapors in the electric arc; it is only in part due to the temperature of the vapors.) Similarly, if a gas or vapor is inclosed in a hollow vessel, such as a glass bulb, and an electric discharge through it is produced by any means, it becomes luminous. These cases of luminosity are said to be due to "electro-luminescence."

Chemical Luminescence. — In certain chemical reactions light is emitted; for instance, when a piece of decayed wood slowly oxidizes, or when phosphorus is oxidized. These are illustrations of "chemical luminescence."

Fluorescence and Phosphorescence. — There are many bodies which emit waves as a result of their having absorbed other ether waves, quite apart from any radiation due to temperature alone. Some bodies emit these waves only while they are absorbing the other waves; while others continue to emit them even after the absorption ceases. All bodies of this kind are called "fluorescent," and the phenomenon itself is called "fluorescence"; while the second division of these bodies, as just described, are called "phosphorescent," and the phenomenon is called "phosphorescence." If a beam of light is passed into a fluorescent substance, certain trains of waves are absorbed and others are transmitted; the energy of these absorbed waves is not spent in producing heat effects, but in emitting other ether waves, which proceed out in all directions. So this fluorescent light may best be seen by looking at the substance from one side.

This phenomenon was first observed by Herschel and Brewster, but was first thoroughly investigated by the late Sir George Stokes. He showed that in all cases observed by him, the fluorescent light was of a wave length longer than that of the waves whose absorption caused the fluorescence. This relation is not, however, true in all cases. Some common illustrations of fluorescence are the colors seen in certain forms of fluor spar (whence the name of the phenomenon); the color of canary glass — which is ordinary glass containing traces of certain salts of uranium; the color of a decoction of the bark of chestnut trees; the color of the surface layers of kerosene oil; etc.

Phosphorescence is exhibited by the sulphides of barium, calcium, strontium, etc., and by a great many ordinary substances to a certain extent. Sometimes the light is emitted for only a minute fraction of a second; but in other cases it continues for hours.

Conclusion. — In many cases it is impossible to say exactly what is the cause of the luminosity; and in nearly all there are several phenomena involved. We can, however, divide all cases of radiation into two classes: in one, the substance that is radiating does not change so long as its temperature is maintained constant; in the other, the substance does change even if its temperature is kept unchanged. In the first class of bodies, the radiation is a purely temperature effect; and to them Balfour Stewart's or Kirchhoff's law — as it is more often called (see page 301) — and the other laws of radiation may be applied. This is not true of the bodies of the second class, in which molecular changes are going on.

Spectroscopes

Different Forms. — In order to study the radiation of any body, some method of dispersing it into a pure spectrum, and some instrument which is sensitive to the various radiations, must be used. As we have seen, there are three ways in which dispersion may be secured: by the use of a prism, a grating, or some interference apparatus. Further, a slit (or small source of light) and a converging lens must be used. Thus we have prism, grating, and interference spectroscopes. The conditions as to the purity of the spectra formed by these instruments have been discussed in previous chapters.

Precautions. — There is one obvious precaution which must be taken with these instruments; allowance must be made for absorption of the ether waves produced in the apparatus itself. Thus, a glass prism can be used to study the spectra of visible sources of light, but not of those which emit only very short or very long waves, because glass absorbs them. A quartz prism is ordinarily used for examining the spectra produced by very short waves; a prism of rock salt, of fluorite, or of silvite, those produced by long waves. Again, a reflecting grating does not reflect all waves equally; and air absorbs certain long waves, certain visible ones, and all the extremely short ones. These absorption phenomena must be observed by preliminary investigations.

Receiving Instruments. — In order to detect the radiations various means must be adopted, as has been already explained. For a limited range of wave lengths the eye may be used; for these and for shorter ones, photographic methods may be applied; for longer waves some thermometric device is ordinarily made use of, such as a bolometer, a thermopyle, a radiometer, etc. These instruments do not necessarily measure the *intensity* of the radiation, but by suitable calibration methods many of them may be used for this purpose. There are, of course, many other methods of observation.

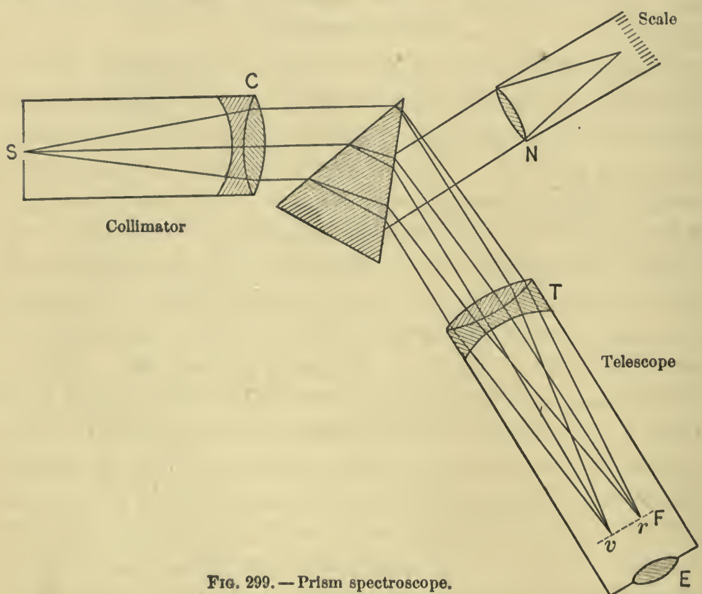


Fig. 299. — Prism spectroscope.

Scale Attachments. — Many spectroscopes are provided with divided scales so placed as to coincide with the spectra formed; and in this manner the position of any definite train of waves may be recorded and thus described. Since by means of a grating the wave lengths of any radiation may be measured, it is a simple matter by using, in a preliminary experiment, certain radiations whose wave lengths are known, to calibrate this scale; and then the instrument may be used to measure the wave lengths of other radiations.

Different Kinds of Spectra

Continuous and Discontinuous Spectra. — A distinction has been made already between continuous and discontinuous spectra. In the former all waves within certain limits are present; so there are no gaps in them. In the latter only certain isolated trains of waves are emitted, thus forming separate "lines." Investigations show that all solids and liquids — with possibly a few exceptions — emit continuous spectra; while all gases and vapors emit discontinuous ones. (This is obviously what one would expect to be the case from the kinetic theory of different forms of matter.)

Gaseous Spectra. — The spectrum of a gas depends, naturally, upon the manner in which it is rendered luminous. So we have "flame spectra," "arc spectra," "spark spectra," "fluorescent spectra," etc. If, however, a gas is made luminous in any definite manner, the waves it emits are definite and characteristic of the gas. Thus, different gases may be identified by their spectra; and in many cases the discovery of new lines in the spectrum of a gas that was supposed to be pure has led to the identification of new elements.

Absorption Spectra. — If the radiation from a solid or liquid falls upon any body, certain waves are absorbed; and so only a portion of the incident waves are transmitted. The spectrum of this transmitted radiation is called the "absorption spectrum" of the body which produces the absorption. This absorption takes place in many ways, as has been already stated. In all, the absorption is due, in the main, to the resonance of the minute parts of the molecule or of the molecules themselves; and in the greater number of bodies the energy absorbed is distributed among the molecules of the body, and is manifest by heat effects.

This is called "body absorption." In other substances the energy absorbed in the interior is spent in emitting other waves of longer wave length, thus producing fluorescence. In certain bodies, the absorption takes place in a very thin surface layer; but the larger portion of the energy incident upon the surface is reflected directly. This is the case with the metals and a few other bodies, and is therefore called "metallic absorption."

The law of Kirchhoff in regard to the equality of radiating and absorbing powers may be applied to a substance which exhibits body absorption only. Thus, if a substance absorbs certain trains of waves of definite wave lengths, it has the power of emitting them if rendered luminous by *means of temperature alone* (and, also, often if other means are used), and the intensities of absorption and of radiation are the same if the temperature of the substance is the same in the two cases. While if the temperature of either condition is decreased, so is the intensity of the effect. Thus, if a white-hot solid is placed behind a quantity of *cooler* gas or vapor, the absorption spectrum is a continuous one from which certain isolated waves are absent; *and these are identical with those which the luminous gas would emit.* The gas absorbs certain waves and transmits the others; it also radiates waves of the same wave length as those which it absorbs; but the intensity of these radiations is so much less than that of those which are transmitted, that the spectrum is practically as if the gas did not radiate.

If the gas or vapor is at a higher temperature than the white-hot solid, the spectrum will be that of the luminous gas with a continuous weak background; *i.e.* it is a bright-line spectrum. If the gas or vapor and the solid are at the same temperature, the spectrum will be continuous.

Solar and Stellar Spectra. — These facts are illustrated in the spectra of the sun and of the stars. The solar spectrum and certain stellar spectra are observed to be absorption

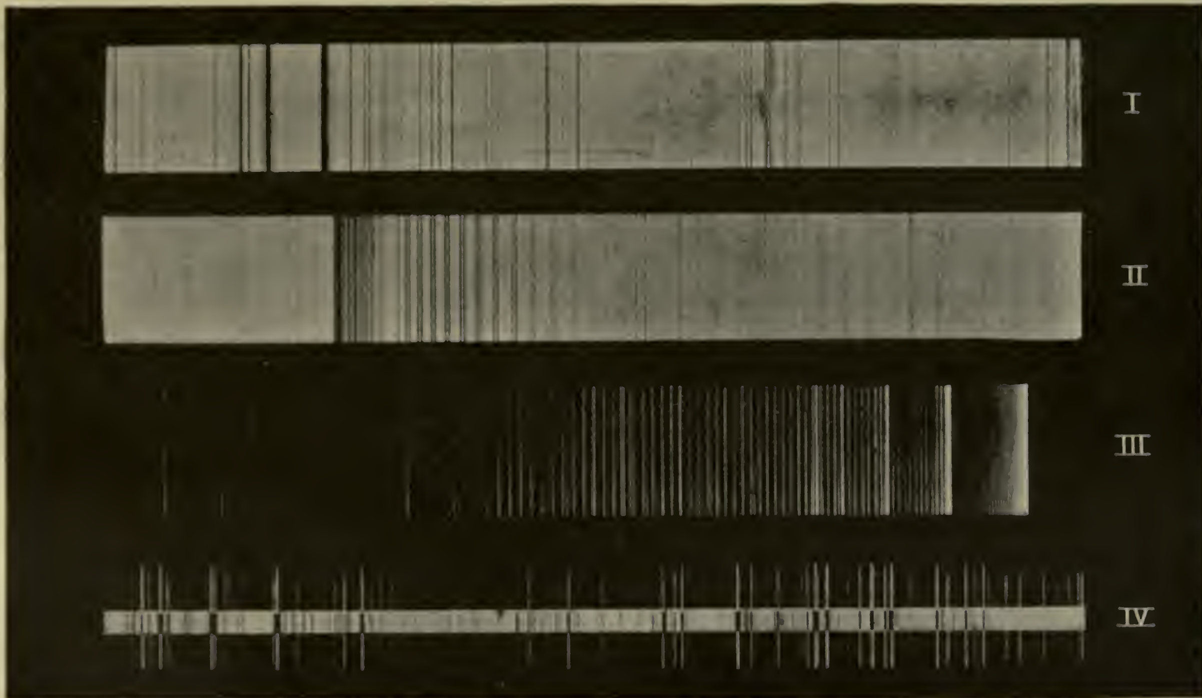


Fig. 300. — (I) portion of solar spectrum, showing absorption lines due to vapors on the sun; (II) portion of solar spectrum, showing absorption lines due to oxygen in the air; (III) portion of the emission spectrum of carbon; (IV) portions of the solar spectrum and the emission spectrum of iron vapor.

ones; while other stars produce bright-line emission spectra. The explanation of the latter is obvious: the stars producing them are surrounded by a luminous gaseous atmosphere, which is hotter than the interior portions. Similarly, in the case of the solar spectrum and other absorption stellar spectra, the explanation is that the interior portions are solid or liquid, and are at a higher temperature than the atmosphere of gases and vapors outside. These vapors are naturally those formed by the evaporation of the interior substances.

The absorption spectrum in the case of the sun consists of the Fraunhofer lines; and they can be identified almost completely with the emission spectra of the vapors of certain substances here on the earth; and thus the constitution of the sun is known. The following are a few of the substances which are in this manner known to be in the sun: calcium, iron, hydrogen, oxygen, sodium, nickel, magnesium, cobalt, silicon, aluminium, carbon, copper, zinc, cadmium, silver, tin, lead, etc.

Some of the absorption lines in the solar spectrum are due to absorption by the atmosphere around the earth. Thus, certain groups of lines known as the "A," "B," "a" and "δ" "bands" are due to absorption by the oxygen in the air, while numerous other lines are due to the presence of water vapor. A method for distinguishing between solar and terrestrial lines will be described presently.

Similarly, by a study of the spectra of the stars, either emission or absorption, a great deal may be learned in regard to their constitution, and also motion, as will be shown immediately. The study of these and similar phenomena forms the science of Astrophysics. An excellent book to consult on this subject is Miss Clerke's *Problems in Astrophysics*, New York, 1903.

In speaking of wave motion a certain general property, known as Doppler's principle, was described (see page 345).

It states that when a source of waves is approaching a point in space, the wave number at this point is increased, while the converse is true if the source is receding. In the case of ether waves that are being dispersed by a glass or quartz prism or grating, this would be shown by a change in their refrangibility — an increase if the source is approaching the earth, a decrease if it is receding from it. Therefore, if a star is emitting certain trains of waves, their corresponding spectrum lines will all be shifted sidewise by an amount depending upon the velocity of the star in the line of sight. If these lines, then, all apparently agree exactly with lines observed here on the earth in the laboratory for any known vapor, except that they are all slightly displaced, the obvious explanation is that the star is moving in the line of sight; and its velocity in this direction may be deduced from the amount of the observed shift. (This statement is not absolutely correct, for shifts of the lines may sometimes be due to anomalous dispersion or to abnormal pressures in the atmospheres of the stars.) Similarly, if the image of the sun is focused by a lens upon the slit of a spectroscope, and it is so arranged that first one edge and then the other of the sun's image is on the slit, the lines in the solar spectrum that are due to *solar* absorption will be shifted slightly, owing to the fact that one edge of the sun is receding from the earth while the other is approaching it, because of the rotation of the sun. But those lines in the spectrum due to absorption in the earth's atmosphere will not be so displaced.

The student should consult Ames, *Prismatic and Diffraction Spectra*, for Fraunhofer's original memoirs, and Brace, *The Laws of Radiation and Absorption*, for the memoirs of Kirchhoff and Bunsen.

CHAPTER XXXVII

EXPLANATION OF COLOR

General Discussion. — The color of an object that is self-luminous depends upon the character of the light that it emits. If it radiates all the visible waves with suitable intensities, it will produce in a normal eye the sensation that we call "white." (The case of defective eyes will be considered in the next chapter.) If the intensity of certain waves is abnormally great, the light appears colored, as is shown when "red fire," a "sodium flame," etc., are used.

The color of most objects, however, is due to the fact that they are illuminated and either reflect or transmit light to the eye of an observer. It is obvious that the color of the object will depend fundamentally upon that of the illuminating light; but we are so accustomed to viewing objects in the white light produced by diffused sunlight, that in describing the color of any object it is always assumed that white light is used with which to illuminate it.

When we consider the color of an illuminated body, it is evident that it may be due to any one of several causes. It has been explained in the previous chapter that absorption of light may take place in many different ways, and corresponding to each of these there will be certain color phenomena. Again, we have seen how colors may be produced by any dispersive action, such as that of a prism, a grating, or an interference mechanism. These various cases will now be discussed briefly.

Absorption Colors

Body Absorption. — The most familiar kind of absorption is that shown when the incident light is absorbed in the interior of the body and the energy of the absorbed waves is spent in producing heat effects. The light that is transmitted appears colored, therefore, owing to the disappearance of certain trains of waves. If a single train of waves of wave length l is absorbed, which corresponds, therefore, to a definite color, the transmitted light will include all the other trains of waves, which will combine in the eye to produce a definite color, called the "complementary color" of that of the waves which were absorbed. If the body absorbs two trains of waves, it may happen that the intensity of the absorption is not the same for both trains; that is, it may require a greater thickness of the body to extinguish one color than is required for the other; and it is thus apparent how such a body may appear of a different color as its thickness is varied.

It is evident that if the absorbing substance is *transparent* for those waves which it does not absorb, it cannot itself be seen when viewed from the same end as the incident light, or from one side; but if owing to any cause the body *diffuses* the light which it does not absorb, then it will appear of the same color when viewed from any direction. Thus a tank containing colored water will appear practically black, except when the transmitted light is viewed, if there are no minute solid particles in suspension; but if these are introduced, it appears colored from all points of view.

If two portions of matter having body absorption are so placed that the incident light falls upon one, and the transmitted light is then incident upon the other, the color of the emerging light is that due to the waves which are left after two absorptions. The color of all leaves and flowers, of various cloths, of paints, of bricks, etc., is due to body absorption. If two paints are mixed, their color is, as just explained, that due to the absorption by both the paints; there is a double *subtraction*, as it were, from the incident light.

The nature of the light that gives an object its color may be determined in two ways: one is to illuminate the object with white light and analyze by a spectroscope that light which is diffused; the other is to form a continuous spectrum on a white wall and move the object along this; if it appears black for any position, it means that the color corresponding to this position is absorbed, but if it transmits and so diffuses any particular color of the spectrum, it will in the corresponding position appear of this color.

Fluorescence and Phosphorescence. — If the energy of the waves absorbed in the interior of a body is spent in producing other waves, which are therefore radiated in all directions, the phenomenon is called, as has been said, fluorescence. The fluorescent light consists in general of waves whose wave length is longer than that of the waves whose absorption produces the fluorescence. In this case the color of the transmitted and the diffused light is not the same. It is evident that if the fluorescent body is thick, the waves which cause the fluorescence may be entirely absorbed in that portion of the body which is first traversed by the light; so that the fluorescence will occur in this portion only. In some cases this color is confined to almost the surface layers.

If the emission of light continues after the incident light is intercepted, the phenomenon is called, as has been said, phosphorescence. Evidently there is some molecular transformation involved in this.

Surface Color. — When polished metals and many of the aniline dyes in the solid form (*e.g.* a dried drop of red ink on paper) are viewed in white light, they have a peculiar appearance which is called "metallic lustre." This is due to the fact that they reflect certain waves much more intensely than others, or, in other words, they have "selective" reflection. This process does not take place in the interior of the substance, as in the case of a colored liquid, but at the surface.

If a substance showing this metallic lustre, or surface color, is made in a film which is sufficiently thin, it will transmit certain waves. But the color by reflected light is not the same as by transmitted; in some cases they are approximately complementary.

These substances which exhibit surface color have anomalous dispersion and change plane polarized light into elliptically polarized light by reflection.

Scattering by Fine Particles. — If the light traverses a region where there are numerous minute particles, it may happen that they are of such a size as to scatter certain trains of waves, and to let pass unaffected all trains of longer wave length. The light so scattered is plane polarized if it is viewed at right angles to the incident beam. This scattering is the explanation of the blue color of the sky, as has been already said, and of the color of sunset clouds, at least in part. The phenomenon also plays a most important part in determining how much radiation (visible and invisible) reaches the earth from the sun.

Dispersion

It is not necessary to say anything here in regard to the dispersive action of prisms, gratings, etc., but a few illustrations may be given of colors due to it. Prismatic dispersion is shown by rainbows, halos around the sun and moon, dewdrops, diamonds when suitably cut, etc. Diffraction colors are seen when looking at mother-of-pearl, at certain fine feathers, at coronæ (the colored rings around the moon), through fine-meshed cloth at a bright light, etc. Interference colors are shown by soap bubbles and other thin films of transparent matter, by opals, etc.

CHAPTER XXXVIII

THE EYE AND COLOR SENSATION

A TEXT-BOOK of Physics is not the proper place for a detailed description of the structure of the human eye or of the various theories which have been advanced to account for the sensation of color. Some treatise on Physiology or on Physiological Optics should be consulted. It is simply necessary to state here a few facts which are of physical importance.

The Eye. — From an optical standpoint the eye consists of a converging lens which is provided in front with a diaphragm of adjustable diameter, the "iris," and whose focal length can be changed at will to a certain degree. (This power of accommodation is greatly decreased as one grows old.) This lens exhibits both spherical and chromatic aberration, but not to a noticeable degree in general. The medium on one side the lens is the air, but on the other is a liquid filling the cavity of the eye.

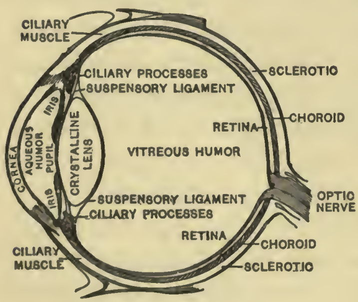


FIG. 301.—The human eye.

At the rear of this is the "retina," upon which a normal eye forms an image of the object viewed. A "near-sighted" eye has its focus in front of the retina; while a "far-sighted" one has its back of it. In the former case, the image may be formed on the retina if a diverging lens is used in front of the eye; in the latter, if a converging one is substituted.

Perception of Color. — The retina consists of a structure of minute parts which are intimately connected with the endings of the optic nerve. The exact mode of excitation of these nerve endings by the incident ether waves is not known. Certain portions of the retina, viz., those remote from its centre, play no part in color sensation ; for, when waves of all wave lengths lying within the limits of the visible spectrum are incident upon them, one is conscious of a sensation of gray only. This is true of all portions of the retina if the light is faint, with the exception of a small area, called the "yellow spot," which gives color sensations only. This spot is slightly off the axis of the eye considered as a lens. There is also a minute area — called the "blind point" — near the centre of the retina, where the optic nerve enters, at which no sensation of light is produced. Over the other central portions of the retina, light of all different colors may be perceived.

Addition of Colors. — It has been known since the experiments of Newton that, in order to produce the sensation "white," it was not necessary to have all the trains of waves in a continuous spectrum from violet to red. Corresponding to any color there is another such that if these two sensations are produced *simultaneously* in the eye, *white* is perceived. These two colors are called, as has been said, complementary. One way of producing these simultaneous sensations is to paint different sectors of a circular piece of cardboard with the two colors, and then to rotate it rapidly while it is illuminated with white light. Thus at consecutive minute intervals of time if one looks at the rotating disk, the eye receives first one impression and then the other ; but since, if impressions reach the eye at intervals faster than about thirty or forty a second, a continuous effect is produced, the eye in this case receives two simultaneous impressions. This is what may be called the *addition* of colors ; and it is evident that the mixing of paints, or the subtraction of color, has no connection with it.

Similarly, a great variety of choices of *three* colors may be made which when added in suitable intensities will produce white light. Taking any three such colors and adding them in different intensities, any other color which is desired may be produced. This proves, then, that in order to account for the perception of colors of all kinds, it is simply necessary to assume that in the eye there are three sets of nerves corresponding respectively to these three colors. (Some writers claim that there are four such sets of nerves, while Hering has an entirely different theory. The statements made here are in accord with all ordinary facts, and embody a portion of the theory of color sensation that was advanced by Thomas Young and supported and extended by Helmholtz.)

The Young-Helmholtz Theory of Color Sensation. — The choice of these three colors is to a certain extent arbitrary; but it is limited by an investigation of different cases of color blindness. Many people are afflicted with an inability to recognize certain objects as colored which appear so to the normal eye; to them they appear gray. Other colored objects appear to them to have colors different from those which would be ascribed to them by an observer whose eyes are normal. The simplest explanation is that in these cases one or more of the sets of nerves referred to above do not respond to stimuli. By an examination of a great many cases of color blindness the conclusion has been reached that the three fundamental color sensations are red (about wave length 671 $\mu\mu$), green (505 $\mu\mu$), and blue (470 $\mu\mu$).

The general explanation of color sensation may then be explained by assuming that there are these three sets of nerves which when excited produce these sensations respectively, red, green, and blue, and that when any train of waves of a definite wave length reaches the retina it stimulates all three sets of nerves, but to different degrees.

Curves can be drawn as shown in the cut which give the results of experiments on adding these three colors with varying intensities. Thus, considering the vertical line through F, the three curves are so drawn that if red light of wave length 671 $\mu\mu$ and of an intensity proportional to the length of the line $\overline{F1}$, is added to green light of wave length 505 $\mu\mu$ and of intensity proportional to the length $\overline{F2}$, and to blue light of wave length 470 $\mu\mu$ and of intensity proportional to the length $\overline{F3}$, the color perceived is blue corresponding to a wave length 486 $\mu\mu$. Consequently, if we can assume that when waves of this wave length fall upon the retina they stimulate the "red," "green," and "blue" sets of nerves to degrees which are proportional to $\overline{F1}$, $\overline{F2}$, and $\overline{F3}$, the phenomena of color sensation have been explained.

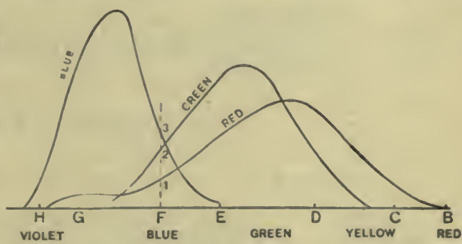


FIG. 302. — Curves of color sensation.

BOOKS OF REFERENCE

EDSER. Light for Advanced Students. London. 1902.

An excellent text-book, containing descriptions of all the fundamental phenomena and references to all the recent investigations.

PRESTON. The Theory of Light. London. 2d edition. 1895.

The recognized book of reference for all the elementary phenomena of Light.

DRUDE. Optics. (Translation.) New York. 1902.

The best modern text-book on Light, giving the accepted theories and explanations of all optical phenomena.

LUMMER. Photographic Optics. (Translation.) London. 1900.

A text-book on geometrical Optics, with special reference to lenses.

MAGNETISM

CHAPTER XXXIX

PERMANENT AND INDUCED MAGNETIZATION

Magnets. — A body which has the property of attracting pieces of iron is called a "magnet"; that is, if a magnet is brought near a piece of iron there is a force between them which is shown by their approaching each other if either (or both) is free to move. Such bodies occur in nature, for one of the forms of iron ore which is not uncommon, a mixture of FeO and Fe_2O_3, is magnetic. It is, moreover, a simple matter to make any piece of iron or of ordinary steel into a magnet. There are two general methods for doing this: one depends upon a property of an electric current, the other, upon what is called magnetic induction. If an electric current is made to traverse a wire which is wound in the form of a spiral spring, or helix, the apparatus is called a "solenoid"; and experiments show that, if a piece of iron or steel is placed inside this solenoid, it becomes a magnet. Or, if a piece of iron or steel is brought near, not necessarily in contact with, a magnet, it is made a magnet also. If a piece of iron is magnetized in this manner, and the magnetizing agency is removed, the iron will lose its magnetism very easily, if it is jarred or subjected to an increase in temperature; but this is not true of the piece of steel — it remains a magnet under all ordinary conditions. All magnets in ordinary use are made therefore of steel, some kinds of which are much better than others. Much progress in this respect has been made in recent years.

Magnets are usually made in the form of bars, rods, or elongated lozenge-shaped "needles." Sometimes the bars or rods are bent into the shape of a U or of a horseshoe; and in this form they are called horseshoe magnets.

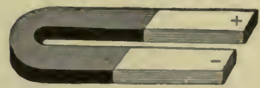
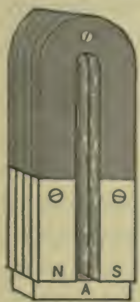
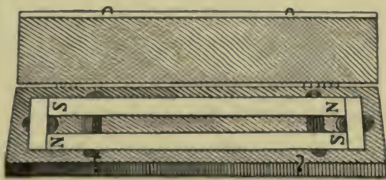


FIG. 303. — Horseshoe magnet.

Experiments show that long magnets are more permanent than short ones; and that they remain magnetized longer if their ends are joined by a piece of soft iron. Thus an iron bar, called the "armature," is always placed across the ends of a horseshoe magnet when it is not in use.



Horseshoe magnet with soft iron armature.



Bar magnets with armatures.

FIG. 304.

Magnetic and Diamagnetic Bodies. — It is found by experiment that a magnet can attract other kinds of matter than iron; such as many forms of steel, nickel, cobalt, manganese, etc. These bodies are called "magnetic," and any one of them can be made a magnet by the methods described above for iron or steel. Again, there are many other substances which are repelled by a magnet; such as bismuth, antimony, and zinc. These bodies are called "diamagnetic." Faraday made the most important observation that the question as to whether a body is attracted or repelled by a magnet depends fundamentally upon the material medium in which the magnet and the body are immersed. In the above definitions of magnetic and diamagnetic bodies this medium is assumed to be the ordinary atmosphere. Faraday showed that, while in one medium a body might be attracted by a magnet, in another it might be repelled. Thus the importance of the medium in the consideration of magnetic phenomena is shown.

Poles. — If an iron or steel rod or "needle" is magnetized by means of a long solenoid, and if it is then removed and suspended by a fine thread or on a vertical pivot, so that it is free to rotate in a horizontal plane, it will turn and after a number of vibrations gradually come to rest in a direction which is approximately (or exactly) north and south. (This fact in regard to a magnetized bar or needle has been known for many centuries, and has been made use of by mariners and travelers.) The end which points toward the north is called the "north pole" of the magnet; the other, the "south pole." The direction in which it points is called "magnetic north and south."

Fig. 305. — Pivoted magnetic needle.

If a magnet is suspended as just described, and another is brought near it, it may be shown that there is a force of attraction between a north and a south pole, but one of repulsion between two north poles or between two south poles. "Unlike poles attract, like ones repel." It is easily proved, further, that the greater the distance apart of the magnets, the less is the force. In order, then, to explain the reason why a magnet when freely suspended points in a north and south direction, all that it is necessary to assume is that the earth itself has the properties of a magnet. The particular magnetic properties that experiments show it to have will be described later in Chapter XLI.

Magnetism a Molecular Property. — If a magnet is broken up into smaller pieces, each fragment, however minute, is found to be a magnet, with a north and a south pole. This leads one to believe that magnetism is a molecular property of all magnetic substances; and all observations are in support of this idea. Every property of a body except its mass and weight is changed when it is magnetized; and conversely, any change that is known to affect the molecules of a body will affect the magnetism of a magnet.

Thus, when an iron rod is magnetized, its length, its volume, its elasticity, etc., are all changed; and when a magnet is hammered or twisted or heated, its magnetism is altered, as is shown by a change in the force which it exerts upon another magnet or upon a piece of iron at a fixed distance from it.

Induction. — We make the assumption, then, that each molecule of a magnetic substance, *e.g.* of a piece of iron, or of nickel, etc., is a magnet; in other words, that each molecule of any one magnetic substance has a certain mass and other mechanical properties and is at the same time a magnet. When the substance is in its natural condition, we can assume that these molecular magnets are not arranged in any order, but are distributed at random; so that, as far as *external* actions are concerned, each tiny magnet is neutralized by those around it. But if a magnet is brought near such a piece of magnetic substance, each of the latter's molecular magnets is acted upon by a force due to the magnet; and the molecules are all turned, more or less completely, in an orderly and regular direction. Thus, if the magnetic substance is a rod or bar, and the magnet is in this form also and is brought near one end of the former, so that its north pole is nearest it, the molecules will turn so that their *south* poles are toward the north pole of the magnet. Therefore the molecular magnets no longer neutralize each other; they now have an external action, and, in fact, the bar which they constitute is now a magnet with its south pole toward the north pole of the magnetizing magnet. The change produced in the arrangement of the molecular magnets by the magnet is

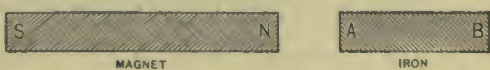


Fig. 806. — Magnetic induction: the end *A* of the iron bar becomes a south pole.

roughly indicated in the accompanying cut, where each molecule is represented by an elongated rectangle whose ends are shaded differently. (Of course, we do not assume that a molecule has actually the shape of a rectangle.)

This explains, then, not alone why a piece of magnetic substance is magnetized by the magnet, but also why the two attract each other. The phenomenon is called "magnetic

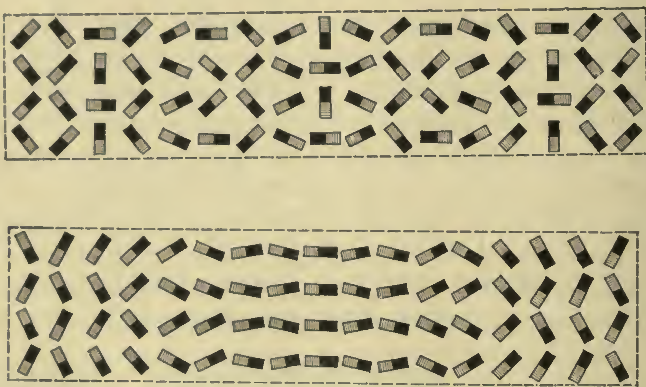


FIG. 307. — Arrangement of molecules in an unmagnetized and a magnetized iron bar.

induction"; or the former is said to be magnetized by "induction." Strictly speaking, these names apply to the phenomenon only so long as the magnetizing magnet is kept in its position near the magnetic substance; when the two are separated, the latter remains a magnet, although a weaker one, for a greater or less time, as described above; but its magnetism is now spoken of not as induced, but as "intrinsic" or "permanent." Similarly, when a rod is magnetized by the action of a solenoid, the magnetism is said to be induced; etc.

The reason why a long thin magnet is more permanent than a short one is clear, because in the latter the two ends are closer together and the molecular magnets at one end may disturb the direction of those at the other, and so produce demagnetization. The action of the armature of a horseshoe magnet is also easily explained: it keeps the molecular magnets at the ends from changing their positions.

It is evident that, when a bar or rod is magnetized by the action of a magnet at one end, the molecular magnets in the former will be arranged in an orderly manner at the end near the magnet; but at its other end

FIG. 308. — Magnetization of an iron bar by induction.

these minute magnets will not be so systematically distributed. The magnetizing of the bar or rod may be made more complete if two magnets are used, one at each end of the rod, and turned in opposite directions as shown in the cut.

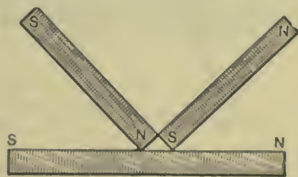


FIG. 309. — Process of magnetizing an iron bar by means of two magnets.

The action may be made still more complete if the two magnets are placed on top of the rod at its middle point — opposite poles being in contact — and are slightly inclined to it, as shown in the cut (and if then the two magnets are drawn off the bar lengthwise in opposite directions). The process should be repeated several times. If during any of these processes the rod is hammered or jarred, the magnetization is increased.

Experience shows that it is impossible to magnetize a bar more than up to a certain degree; it is then said to be "saturated." This fact and all those just described in regard to methods of magnetization are explained easily if it is assumed that the molecules of a magnetic substance are themselves magnets. One can, in fact, make a model of a magnetic substance by placing on a board a great number of small magnets, all pivoted so as to be free to turn around vertical axes; and the entire process of magnetization can be imitated by placing this board inside a large solenoid or bringing magnets near it.

Temperature Effects. — If a "permanent" magnet has its temperature raised sufficiently high, it loses its magnetism, and becomes again simply a magnetic substance. If, however, when magnetism is being induced in a magnetic substance, its temperature is raised, — not too high, — the induced magnetization is increased. These facts are at once explained on the molecular theory of magnetism. (Iron loses its magnetic properties at about 785° C.; nickel, at about 300° C.)

The Poles of a Magnet are of Equal Strength. — Several interesting facts are learned when one studies the force which a bar magnet exerts, either on another magnet or on a piece of magnetic substance. One way of doing this qualitatively is to place the magnet on a horizontal table, and to move near it a small pivoted magnetic needle — such as an ordinary pocket compass. If the latter is *short*, it may be pushed up close to the magnet; and it is observed that when this is done, the small needle does not always point toward or away from one of two *points*, one in each end of the magnet, as it would if there were two *centres of force* in the magnet. On the contrary, the needle points nearly perpendicular to the

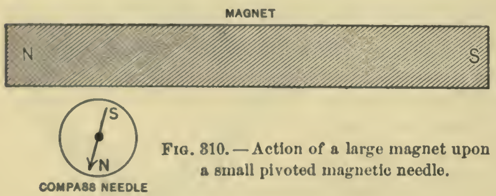


Fig. 310. — Action of a large magnet upon a small pivoted magnetic needle.

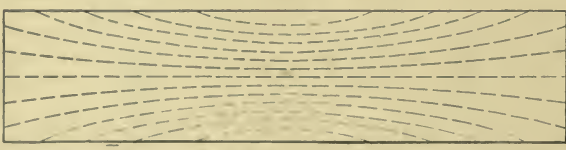


Fig. 311. — Diagram representing arrangement of molecular magnets in a bar magnet.

surface of the magnet at all points except near its middle; and this is what we would expect from our knowledge of the nature of magnetization. The molecular magnets near the centre of the magnet are all more or less parallel to its axis; but near the ends they are turned so that for one half the magnet their north poles are in the surface, while for the other half their south poles are there. We must then regard *each point* of the surface of a magnet as a centre of force; and the total force of a magnet on another magnet, or on a magnetic substance, is found by adding geometrically all the forces due to these magnetic poles in the surface. These may, however, be combined mathematically in certain cases in a manner that is instructive.

Let a bar magnet be under the action of another magnet perpendicular to it at its middle point, whose south pole is turned toward the former, but which is so far away that all these minute surface forces may be considered *parallel*. (We can, if we wish to, consider the second magnet as so long that we need not take into account the forces due to its north end.) All the forces over the north half of the magnet may be added so as to form the resultant R_n; similarly, those over the south half may be added to form R_s. If the magnet is turned so that it is no longer perpendicular to the distant one, these forces still keep their value; and their direction is toward the latter. This condition is practically the case when a magnet is under the action of the earth, and if a magnet is suspended so as to be free, — not alone to turn, but also to *move* as a whole, *e.g.* if it is floated on a cork resting on the surface of a tank of water, — it is observed that the magnet, if not originally in a north or south direction, *turns but does not move as a whole*. This proves that the two forces R_n and R_s are equal in amount but opposite in direction, forming a couple.

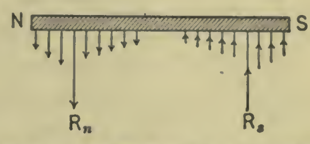


FIG. 812. — Forces acting on a bar magnet owing to a distant magnet.

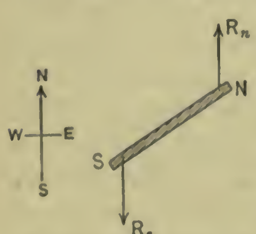


FIG. 813. — Moment acting on a magnet when placed in the earth's magnetic field.

(See page 100.) If a bar magnet could be made whose molecular magnets were all parallel to its axis, there would be magnetic forces only at its ends. So if in the actual magnet just described L is the distance apart of R_n and R_s measured along the magnet, and if this special bar magnet is made of the length L, it might replace the former so far as action at a considerable distance is concerned. The forces on the two ends of the latter magnet due to the distant one are then equal in amount but opposite in direction. We are therefore led to make the assumption that each molecular magnet has two centres of forces, a minute distance apart, which are equal in amount but opposite in kind; that is, if one centre exerts a force of attraction on one end of any distant magnet, the other exerts on it a force of repulsion. Thus, if N_1S_1 and N_2S_2 are two molecular magnets, there are four forces acting on each, as shown, and if S_2 is at the same distance from both N_1 and S_1, the two forces acting on it are equal in amount but are in nearly opposite directions. All observed facts in regard to magnetization are in accord with this assumption. We say that each molecular magnet has two poles whose "strengths" are equal but opposite; or that they have equal but opposite "quantities of magnetism" or "magnetic charges." The same statement in regard to magnetic charges must then be true of any magnet, however large or complicated, because it is made up of magnetic molecules, and therefore contains as much south as north magnetism. It is impossible to separate a north pole from an equal south pole and obtain them distinct from each other, because, when a magnetic molecule is decomposed into simpler parts, it ceases to be a magnet, and equal amounts of north and south magnetism vanish.

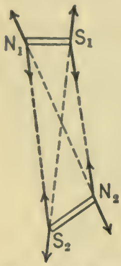


FIG. 314.—Forces between two small magnets.

Magnetic Field and Lines of Force. — When a small magnet is placed near a large one, it is acted on by certain forces;

and, in general, a region in which a small magnet experiences forces is called a "magnetic field." A simple mode of studying and describing the properties of such a field is to draw what are called "lines of force." These are continuous lines such that any one indicates by its tangent at any point the direction in which a *north* pole would move if placed there. Thus, if *P* is any point on a line of force *AB*, a north pole, if placed there, would move in the direction of the curve at *P*, while a south pole would, of course, move in the opposite direction. (If the poles had no inertia, they would actually trace out the whole curve *AB*.) Therefore, if a small *magnet* is placed with its centre at *P*, there will be a moment acting on it which will make it turn and lie tangent to the line of force at *P*.

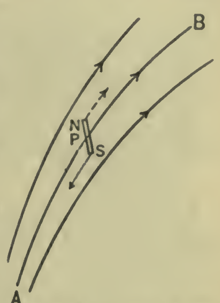


FIG. 315.—Lines of magnetic force, and the forces acting on a small magnet.

This leads at once to two modes of actually drawing the lines of force in any field. One is to move a small magnet from point to point in the field, noting at each its direction; thus, starting from any point *P*, the magnet should be moved a short distance in such a direction that in its new position it is tangent to the same curve to which it was tangent at *P*, etc.; a line can be drawn through these points. Then another starting point is chosen and another line is drawn, etc. The other method is to place a horizontal sheet of paper or a glass plate in the field of force, sprinkle fine iron filings over it, and jar the paper or plate until the filings assume definite positions. Under the action of the magnetic forces of the field each minute filing becomes magnetized, and therefore acts like a small magnet; and under the action of the couple it turns and places itself along the line of force. Thus the action is just as if one had many thousand minute magnetic needles in the field at one time. (Actually in neither of these cases do the forces acting on the small magnet form a couple, unless the magnet which causes the field of force is at a great distance; but the resultant force of translation is as a rule minute.)

Several illustrations of lines of magnetic force, as mapped by means of iron filings, are shown in the accompanying cuts; viz., around a single bar magnet, and between two magnetic poles, alike and unlike. The *direction* of a line of force is defined

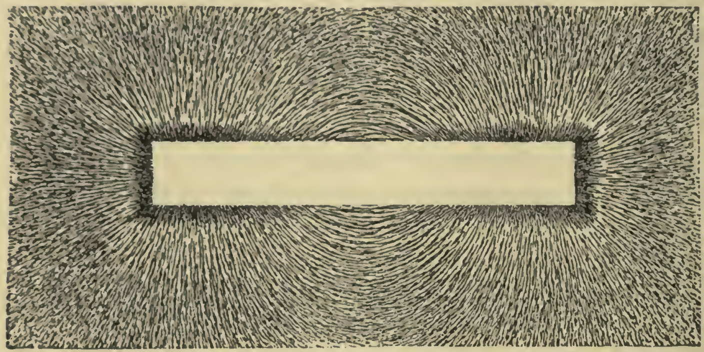


FIG. 316. — A bar magnet.

by its being that in which a *north* pole would move. It follows, therefore, that lines of force start from north poles of magnets and end on south poles. Further, two lines of force cannot cross. It is interesting to note the effect upon a field of force of the introduction of a piece of magnetic substance as shown in Figs. 317 and 317*a*. From the first of these it is apparent that the lines of force, instead of diverging away from the end of the magnet, are converged and enter the end of the iron bar, from the sides and other end of which they again emerge. The action is exactly as if it were

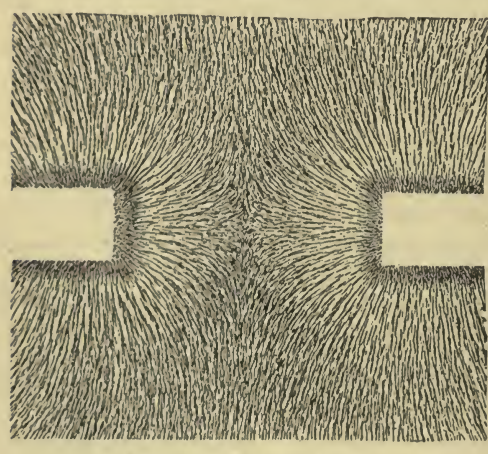


FIG. 316 *a*. — Two similar poles.

easier for the lines of force to traverse a space when it is filled with iron than when it is occupied by air; the iron is therefore said to have a greater magnetic "permeability" than air.

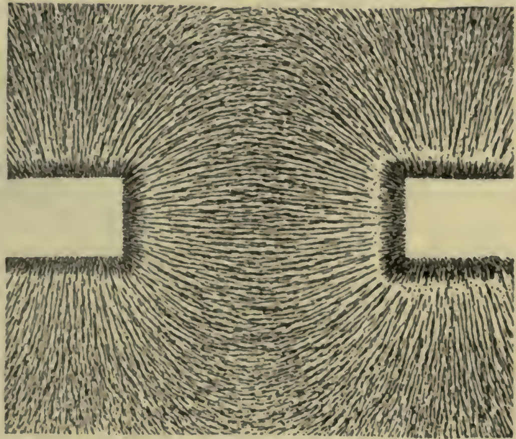


FIG. 316 *b*. — Two unlike poles.

The explanation in terms of force of the convergence of these lines toward the iron is evident. In Fig. 318, let *SN* be a magnet and *AB* be a bar of iron; under the action of the former it is magnetized with *A* as a south pole and *B* a north one. The force which would act upon a north pole placed at any point, *P*, in the field of the magnet due to the magnet alone might have the direction and the amount $\overline{PQ}_1$, while that due to the induced magnetization in *AB* might be indicated by $\overline{PQ}_2$; so the resultant force would be $\overline{PQ}_3$. Thus, in general, if a piece of iron is put in a magnetic field, the lines of force crowd into it out of the air.

FIG. 317. — Magnet and bar of iron.

If the position and shape of the lines of force in any of the illustrations given above be considered, it is seen that they may be described by saying that the lines are distributed as if lines in the same direction repelled each other sidewise, and as if all the lines were stretched so that they exerted a

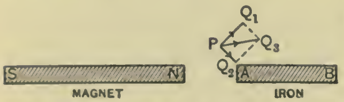


FIG. 318. — Explanation of change in direction of a line of force at *P* owing to the presence of the soft iron.

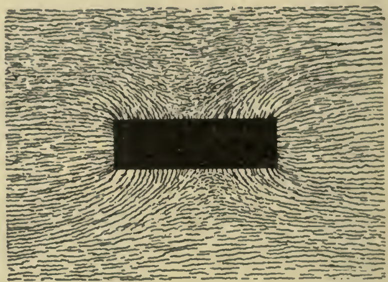


FIG. 317 *a*. — Bar of soft iron placed in a uniform magnetic field.

tension. (Attraction of iron and of unlike poles may be "explained" thus.) Lines of force do not, of course, have a physical existence; and the above statements are simply descriptions of the appearance of their geometrical curves. By means of these ideas it is often possible to give simple descriptions of complicated cases of magnetic forces.

CHAPTER XL

MAGNETIC FORCE AND INDUCTION

Quantity of Magnetic Charge; Law of Force. — If the action of several magnets upon one which is pivoted is observed, it is seen that the intensity of this action, as measured by the deflection of the magnet from a north and south line, depends upon many things. It is different if the magnets are inclined at different directions to the pivoted one; and so for purposes of comparison of different magnets we may place them, in turn, east and west (magnetically) of the pivoted magnet in a horizontal plane, with their north poles pointed toward the latter. As a result, the latter will be deflected and will come to rest, making a definite angle with its original north and south position, which will vary with the distance of the magnet from it, and also, in general, with different magnets when placed in the same position. This indicates that the magnetic forces between poles vary with their distance apart and with different magnets.

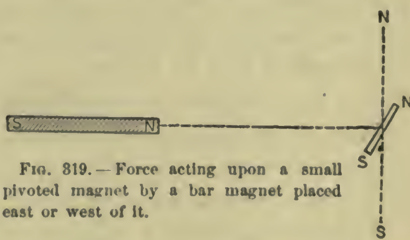


FIG. 819. — Force acting upon a small pivoted magnet by a bar magnet placed east or west of it.

Let us assume, as an ideal case, that the magnets have all their magnetic charges at their ends (see page 602), and let us assume that we can assign a numerical value to this magnetic charge, so that the forces it exerts are proportional to it. Thus, if m is the magnetic charge of the north pole, $-m$ is that of the south pole; and the forces which each exerts are proportional to m. Then, if there is another magnet which under similar conditions exerts different forces, its poles must have different magnetic charges, which may be written m_1 and $-m_1$. Therefore, if these two magnets are acting on each other, the force of the north poles on each other is proportional to the product mm_1. Experiments show that this force varies with their distance apart, being less for a great than for a small distance. Coulomb made the assumption that, so far as distance was concerned, the force varied inversely as its square. So calling this distance r, the law of action of two poles is assumed to be that the force between them is proportional to $\frac{mm_1}{r^2}$. This is known as Coulomb's Law, and it was verified by him (1785) so far as was possible with the instruments at his command. It was verified also by Gauss, and to a greater degree of accuracy; but our main reason for believing that the law is exact is that all of its consequences are found to be in accord with the varied facts of electrical engineering, into which enter so many questions connected with magnets and magnetic fields.

In order to assign a number to the magnetic charge of any magnet, it is necessary to define a unit charge; and in doing this it must be remembered that magnetic forces are different in different media. (See page 595.) Making use of the C. G. S. system of units, a "unit magnetic charge" is defined to be such a charge that, if placed at a distance of 1 cm. *in air* from another equal charge, the force between them is 1 dyne. Then, if a charge equal to m of these units is placed at a distance r cm. from a charge m_1 in air, the force between them expressed in dynes is given by the equation $f=\frac{mm_1}{r^2}$. In any other medium the force is proportional to this, and therefore, following the accepted system of symbols and writing as a factor of proportionality $\frac{1}{\mu}$, the force in any medium is $f=\frac{mm_1}{\mu r^2}$. From what has just been said, the value of μ for air is *one* on the C. G. S. system of units and using as a unit magnetic charge that defined above; naturally, if another unit charge were adopted, the value of this constant for air would be different. The factor μ is a quantity which is characteristic of any medium — it is not, however, necessarily a constant. It is called the "permeability," for reasons which will appear later.

Intensity of Magnetic Field; Magnetic Moment. — When a magnet is placed in a magnetic field, it is acted on by two forces, one at each pole, they being the resultants of all the forces acting over the surface, as explained on page 601. Let us assume the simplest case, viz., that the charges are entirely at the ends; then, if the magnet is *short*, the forces at the two ends are equal in amount, although opposite in direction, because the two ends are at almost the same point in the field. The "intensity" of the field at any point is defined to be the value of the force which would be exerted on a unit north charge if placed at that point. Therefore if the intensity of the field is R, and if m is the charge on either pole of a short magnet which is placed in the field, there is a force Rm acting on each pole. If l is the length of the magnet and if it is in such a position that it makes an angle N with the direction of the field, the perpendicular distance between the two forces is $l \sin N$; so the strength of the couple acting on the magnet is $Rml \sin N$. If the magnet is pivoted

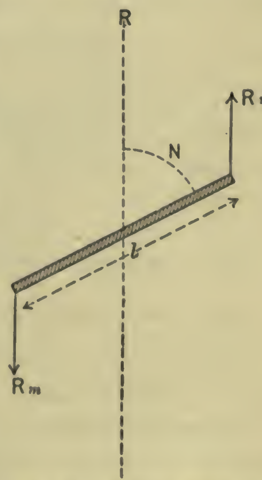


FIG. 320. — Moment acting on a bar magnet when placed in a field of intensity R.

around an axis perpendicular to a plane which includes the magnet and the line of force at its middle point, it will turn under the action of this couple toward the direction of the field ; so the moment of this couple should be written $-Rml \sin N$. The product ml is evidently a property of the magnet itself, and it has received the name "magnetic moment of the magnet."

In the general case of any kind of magnet, the magnetic moment around any axis is defined to be the maximum value of the moment of the forces acting on the magnet when it is placed in a field whose intensity is *one*, with this axis at right angles to the field.

If a magnet is broken up into parts, these are found, as a rule, to have different magnetic moments; and the value of the "intensity of magnetization" at any point of the magnet is defined to be that of the magnetic moment per unit volume around that point. Thus, if M is the magnetic moment of a portion whose volume is V, the ratio $\frac{M}{V}$ in the limit, as V is taken smaller and smaller, is the value of the intensity of magnetization. If the magnetic charges are entirely at the ends, the intensity is the same throughout the magnet; so if this is a cylinder of length l and cross section A, and if the charges on each end are m, the magnetic moment M is ml and the volume V is lA. So the intensity of magnetization $\frac{M}{V}=\frac{m}{A}$; or, it has the same value as that of the charge per unit area on the ends; this is called the "surface density" of the charge.

Magnetic Pendulum. — Measurement of RM. It has been proved in the preceding paragraph that, if placed in a field of intensity R, a magnet of magnetic moment M is acted on by a couple whose moment is $-RM \sin N$, when it makes an angle N with the direction of the field. If, then, it is pivoted so as to be free to turn, and if I is its moment of inertia about the axis of rotation (see page 88), its angular acceleration due to the magnetic field is $\frac{-RM\sin N}{I}$. Therefore the magnet will make oscillations to and fro, through the direction of the field of force. If the angle of oscillation is small, the acceleration is $-\frac{RM}{I}N$; and the vibrations are harmonic with a period $2\pi\sqrt{\frac{I}{RM}}$. (See page 91.) The moment of inertia may be calculated from a knowledge of the dimensions of the magnet, and its period of oscillation may be measured; so calling this last T, we may write $RM = \frac{4\pi^2 I}{T^2}$, and therefore the value of the product RM may be determined. (We shall show in the next paragraph how the value of the ratio $\frac{R}{M}$ may be determined; and so the values of both R and M may be obtained.)

A method is thus offered for comparing the intensities of different fields of force. Call these R_1 and R_2; and using the same magnet in the two cases, let T_1 and T_2 be its periods of oscillation in the two fields. Then M and I being constants for the magnet are the same in both experiments; and we have the two equations

$$R_1 M = \frac{4\pi^2 I}{T_1^2};\ R_2 M = \frac{4\pi^2 I}{T_2^2} \text{ or } R_1 : R_2 = \frac{1}{T_1^2} : \frac{1}{T_2^2}.$$

Measurement of $\frac{M}{R}$. — If the same bar magnet whose period has been determined when vibrating freely in a uniform field of intensity R is placed at rest at right angles to the field, it will itself produce a field of force which at some distance away is nearly uniform, and at right angles to the existing field R. So, if a *small* magnetic needle is suspended at a point some distance away in the direction of the length of the bar magnet, in such a manner as to be free to turn around an axis perpendicular both to the field of force R

and to the bar magnet, it will be under the influence of two fields of force at right angles to each other. It will therefore place itself at such an angle that the moments due to the two fields are equal but opposite, so that they neutralize each other. Call, for the time being, the intensity of the field due to the bar magnet, f. The moment acting on this magnetic needle due to the field whose intensity is R, when the

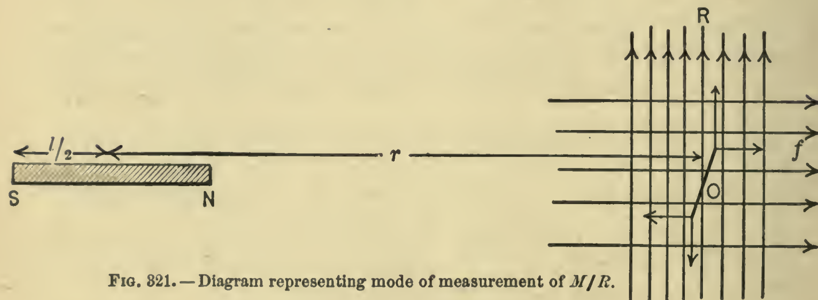


FIG. 821.—Diagram representing mode of measurement of M/R.

magnet makes with its direction the angle N, is $RM' \sin N$, if M' is the magnetic moment of the magnetic needle. But the magnet makes the angle $90° - N$ with the field of intensity f; hence the moment due to it is $fM' \cos N$. The moments are in opposite directions, and must be equal, if the needle is at rest.

Hence $$RM' \sin N = fM' \cos N,$$

or $$\tan N = \frac{f}{R}.$$

The value of f may, however, be easily expressed in terms of M, the magnetic moment of the large bar magnet, and r, the distance from the centre of this magnet to the centre of the magnetic needle. f is the intensity at the point O, the centre of the needle; *i.e.* it is the force which would act on a unit north pole if placed there. The large bar magnet, as placed in the diagram, has a pole of strength $+m$ at a distance $r - \frac{l}{2}$ away from O, and another of strength $-m$ at a distance $r + \frac{l}{2}$. Hence the force on a unit north pole at O is

$$f = \frac{m}{\left(r - \frac{l}{2}\right)^2} - \frac{m}{\left(r + \frac{l}{2}\right)^2} = \frac{2\,mlr}{\left(r^2 - \frac{l^2}{4}\right)^2}.$$

But if r is very great in comparison with l,

$$f = \frac{2\,mlr}{r^4} = \frac{2\,M}{r^3}.$$

Consequently, substituting in the formula for tan N,

$$\tan N = \frac{2\,M}{Rr^3},$$

or

$$\frac{M}{R} = \frac{r^3 \tan N}{2}.$$

Various precautions and modifications for this experiment are explained in laboratory manuals, but it is evident that both r and N can be measured; and so $\frac{M}{R}$ may be determined.

Measurement of *R* or *M*. — By a combination of the two formulæ for RM and $\frac{R}{M}$, it is seen that

$$R^2 = \frac{8\,\pi^2 I}{T^2 r^3 \tan N},$$

$$M^2 = \frac{2\,\pi^2 I r^3 \tan N}{T^2};$$

and so both R and M may be measured.

If R is known for any one field, it has been explained how its value for any other field may be determined by means of a vibrating magnet whose period can be measured.

Magnetic Tubes. — If one refers to the illustrations of lines of force given on page 604, it is evident that these lines are most crowded together at those places where the intensity of the field is the greatest, and are the farthest apart at those points where the intensity is the least. This suggests a systematic mode of drawing lines of force. We can describe a small closed curve at some point near the magnet, and can imagine lines of force drawn through each point of this curve; these lines, if continued, will of course be found to start from a north pole of a magnet and end on a south pole; so they thus form a hollow tube leading from one pole to the other, whose cross section is small near each end, but greater at a distance. If the initial small curve is taken of exactly the proper size, this tube will inclose at its two ends a *unit* magnetic charge. Such a tube is called a "unit tube"; and, if the magnet has a charge m at each end, there are m tubes leaving the north pole and returning to the south pole. It is evident that where the cross section of a tube is least, the intensity of the field is greatest; and *vice versa*. Similarly, if A is the area of any small surface in the field at right angles to the force, and if there are N tubes passing through this surface, the intensity of the field at that point is proportional to the limiting value of the ratio $\frac{N}{A}$, as A is taken smaller and smaller. In words, the intensity of the field at any point is proportional to the number of tubes per unit area at that point.

Magnetic Induction. — As was explained on page 605, and as is apparent from the cuts on that page, the effect of introducing a piece of iron or other magnetic material into a field of force in air is to cause the lines of force to change their direction and enter the iron. If the iron is in the form of a rod, and if its cross section is A, more tubes pass through it than did through the same area of air before the iron was substituted for it. If the original field of force in the air is uniform, so that the intensity is the same at all points, the lines of force are all parallel, and the tubes are all of the same cross section. If the intensity of the field is R, the number of tubes per unit area is proportional to this. If, now, a long iron rod is introduced parallel to the field, the number of tubes per unit area of its cross section is greatly increased; and it may be proved by methods of the infinitesimal calculus that the ratio of this number to the previous one equals the value of the quantity μ for iron, as defined on page 608. It is for this reason that μ is called the permeability. (For different kinds of iron, and for different conditions, μ may have values as great as 2000.) It is thus seen that for any magnetic substance μ is greater than for air. The number of tubes per unit area in the iron (or other magnetic substance), when in a field of intensity R, is, then, proportional to the product μR; and this quantity has received the name of the "magnetic induction" at the point where R is the intensity of the field and μ is the permeability. The fact that the tubes do not simply *end* on the iron rod, but must be considered as passing *through* it, may be proved by certain phenomena of electric currents which will be discussed in a later chapter, and also by the simple experiment of cutting the rod into two pieces by a transverse section and separating them slightly; the field of force in the crevasse is found to be much more intense than in the original field.

When the magnet causing the field of force is removed, the iron remains magnetized, and therefore some tubes still remain passing through it. So we are led to believe that in the case of an ordinary magnet these tubes do not *end* at its poles, but continue through it, forming endless closed rings, exactly as if one were to take a piece of rubber tubing and bring its two ends together. This conception of tubes of magnetic induction is due to Faraday.

Diamagnetic Bodies. — The case of a diamagnetic body may be treated in a similar manner. Experiments show that if such a body is made in the form of a bar, and brought near a bar magnet, it is magnetized by induction, but in a direction opposite to what it would be if it were iron. In the cut, where AB is the diamagnetic body, the end A, nearest the north pole of the magnet, is thus found to be a north pole, while B is a south one. There is repulsion, therefore, between the magnet and the diamagnetic body. If the lines of force are drawn, they are found to be as shown in the cut, showing that, instead of being crowded together in the diamagnetic body, they avoid it. There is therefore a smaller number of tubes in the bar than there was in the original field before the bar was intro-

FIG. 822.—Action of a magnet on a diamagnetic body: A becomes a north pole.

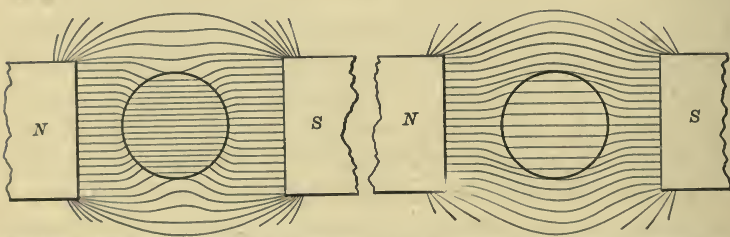


FIG. 823.—Effect of introducing in a uniform magnetic field: (1) a sphere of magnetic substance; (2) a sphere of diamagnetic substance.

duced. This proves that the value of the permeability μ for a diamagnetic substance is less than for air.

Energy Relations. — Since all motions in a system of bodies take place in such a manner that the potential energy of the system becomes less (see page 114), it must be possible to explain from this point of view the attraction and repulsion observed with magnetic and diamagnetic substances. If we consider the formula $f = \frac{mm_1}{\mu r^2}$ for the forces between magnetic charges, it is evident that there are changes in the potential energy whenever magnets are moved with reference to each other; because work is done in overcoming the forces, or by the forces. The magnets themselves are not changed; and so we are led to believe that the energy is located in the surrounding medium where the magnetic field exists. It follows from the formula that f is small if μ is large, or, in words, the forces are small if the permeability of the medium is large; and consequently in such a medium the energy per unit volume is also small, since small amounts of work are involved in any changes, other things being equal. (Magnetic forces can be felt through a vacuum, and so the energy of a magnetic field is, in the case of any material medium, both in the ether and in the matter.)

Attraction and Repulsion. — Therefore, if a piece of iron — for which μ is greater than for air — is introduced into a field of force in air near a magnet, the energy in the *space* occupied by the iron is less than when it was occupied by air; and the decrease in the energy is greater if the field of force is intense than if it is feeble. In other words, if a piece of iron is moved up gradually toward a magnet, the potential energy of the field becomes less and less; therefore, if a magnet and a piece of iron are left to themselves, there is a force of attraction between them, and they will approach each other. Similarly, a magnet will attract a piece of any magnetic substance *in air*. Conversely, and for obvious reasons, a magnet will repel a piece of any diamagnetic substance in air. In general, if μ for any substance is greater than for the surrounding medium, it is attracted by a magnet; while, if it is less than for the medium, there is repulsion. The observations described on page 595 are therefore explained.

An exactly analogous phenomenon in mechanics is afforded by the motions of a block of stone and a block of wood when immersed in a tank of water: the former will be attracted by the earth and will sink; the latter will be repelled and will rise. The explanation in both cases is that the motion takes place in such a direction as to make the potential energy of the system less. The stone sinks because it is heavier than the water; and therefore by replacing an equal volume of water closer to the earth, the potential energy of gravitation is decreased. The wood rises because it is lighter than the water; and, therefore, if it moves up and water replaces it, the potential energy is again decreased.

CHAPTER XLI

MAGNETISM OF THE EARTH

Magnetic Elements. — The fact that there is a magnetic field of force on the surface of the earth is proved by the observations on the motion of a suspended magnet, which were referred to on page 596. If a bar magnet or a magnetic needle is suspended in such a manner that it can turn freely in all directions, it will finally come to rest in a position such that its axis is inclined with reference to a horizontal plane and lies in a vertical plane which nearly, if not quite, coincides with the geographical meridian at the point of suspension. This vertical plane is called the "magnetic meridian" at the point; and the angle it makes with the geographical meridian is called the "magnetic declination" or the "variation." The angle which the axis of the needle makes with the horizontal plane is called the "magnetic inclination" or "dip." The earth's magnetic field at any point is then completely defined by its intensity, the declination and the inclination. These three quantities are called the "magnetic elements."

FIG. 324. — A magnetic needle suspended free to turn in any direction.

The dip can be measured by observing the angle which the needle makes with the horizontal plane. The variation is most easily determined by mounting the needle so that it is free to turn about a *vertical* pivot, and noting the angle it makes with a true north-and-south line, which may be located by astronomical methods. It is convenient for most purposes of measurement to consider the earth's magnetic force as resolved into two components, one horizontal, the other vertical. Thus, if $\overline{OD}$ is the direction of the field of force at O, and if $\overline{OA}$ and $\overline{OB}$ are horizontal and vertical lines through O in the same plane as $\overline{OD}$, the angle (AOD) is the dip; and calling it N and the intensity of the force R, the horizontal component is $R \cos N$, and the vertical one $R \sin N$. The former can be measured with great accuracy by the method described on pages 610–613. Therefore, since the angle of dip, N, can be measured directly, the value of R may be deduced. Further, if the ratio of the horizontal and vertical components can be measured, the dip may be calculated; for, calling these H and V, $\tan N = \frac{V}{H}$.

FIG. 325. — A dip circle.

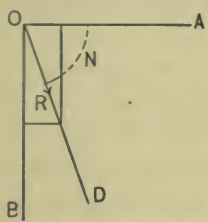


FIG. 326. — Diagram representing the horizontal and vertical components of the earth's magnetic force.

Variations in the Elements. — Observations show that the values of all three of these elements at any one point are continually changing. So far as is known, these changes are periodic, that is, for instance, the dip makes a pendulum-like oscillation during the twenty-four hours, increasing slowly, then decreasing, etc.; further, the mean value for any one day changes slightly the next day, and so on, having an oscillation whose period is a year; and, again, the mean value for any one year is not the same for the next year, but changes slightly; but the period of this change is not known, for since regular observations began to be taken — about the year 1540 — this oscillation in the mean annual value of the dip has not been completed. Similar statements may be made in regard to the other two magnetic elements; there are daily, yearly, and secular changes, so called. (There are other periodic changes than these, but they are the most important.) It often happens that there is a sudden and unexpected disturbance of the magnetic elements of a magnitude far greater than the regular changes; this constitutes a "magnetic storm." The explanation of such phenomena is not known; but observations have shown that they occur most frequently when the spots on the sun and when auroræ in our atmosphere are most numerous.

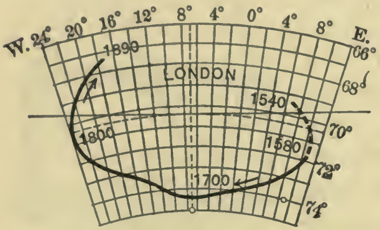


Fig. 327. — Chart showing secular change in the earth's magnetism, from observations made at London; the black line indicates both the inclination and the declination.

Magnetic Maps. — The magnetic field over the earth's surface may best be described by drawing on a map of the earth certain lines which indicate the values of the elements at any one epoch. Thus lines can be drawn such that at each point of the earth's surface, through whose position on the map any one line passes, the value of the declination (or variation) is the same. Such lines are called "isogonals," and are of the greatest possible assistance to mariners and surveyors. They are shown in the cut for the year 1900, and each one is marked with a certain number, *e.g.* 5°, which indicates the value of the variation for all points on that line. These lines run approximately north and south; and it should be observed that for two lines the variation is zero, *i.e.* at points on them a magnetic needle points true north and south: they are called "agonic" lines. One of these is approximately a great circle of the earth; the other lies in northern Asia, and is called the "Siberian oval."

Again, lines can be drawn which indicate in a similar manner the inclination or dip; they are called "isoclinals," and are approximately parallels of latitude. The line of zero dip is called an "aclinic" line, or the "magnetic equator." There are two points in the earth's surface where the dip is 90°; these are often called the "magnetic poles." The lines for the year 1900 are shown in the cut.

Other lines, giving other information, may also be drawn; but they need not be described here.

Conclusion. — The explanation of the magnetic action of the earth is not known. It has been proved, however, that it is due almost entirely to causes which are within the earth itself. Certain of the periodic changes are occasioned, however, by external causes, such as electric currents in the atmosphere.

Historical Sketch of Magnetism

The property which the lodestone possesses, of attracting iron, was known centuries before the beginning of the Christian Era, because it is mentioned by Thales, who lived from the year 640 to 546 B.C.

The Greeks and the Romans were acquainted with the fact that the intervention of other bodies, like brass, does not destroy magnetic effects. That like poles repel and unlike attract, and that a lodestone possesses the power to communicate polarity to inert iron, were known at least as early as the twelfth century.

The compass was in daily use in Europe also as early as this, but the discoveries of magnetic declination and its variation from place to place were made by Columbus in 1492.

Hartman is reputed to have discovered the dip in 1544. He obtained a value of 9° where he should have obtained

LINES OF EQUAL DECLINATION FOR THE YEAR 1900.

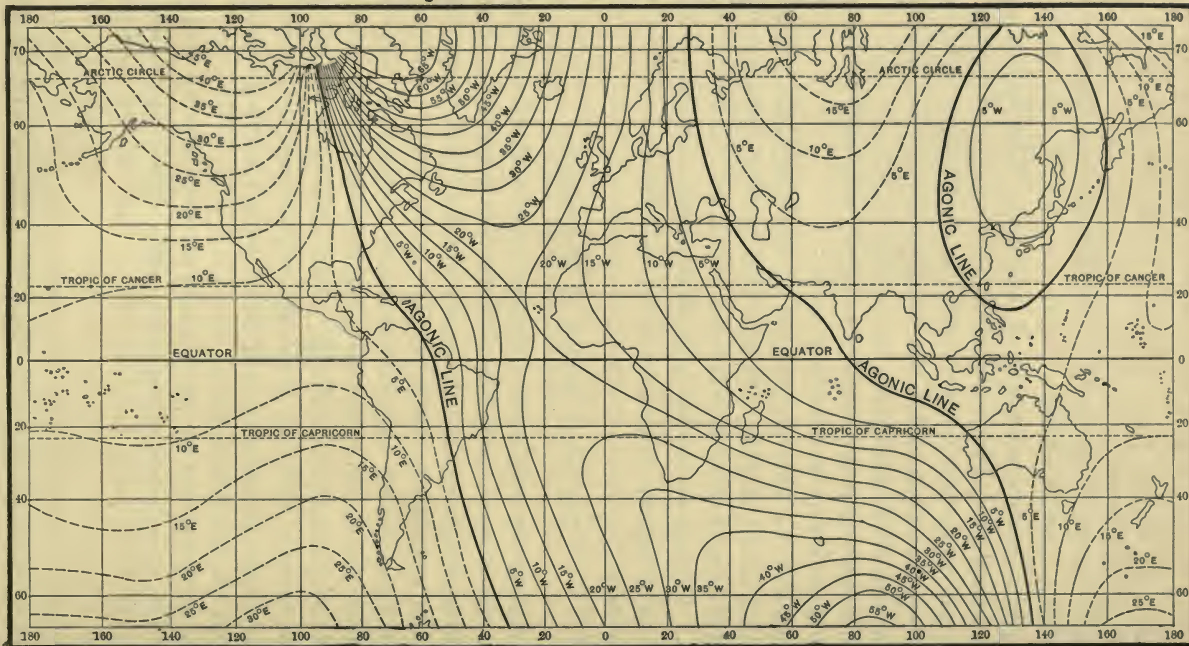


Fig. 328.

LINES OF EQUAL DIP FOR THE YEAR 1900.

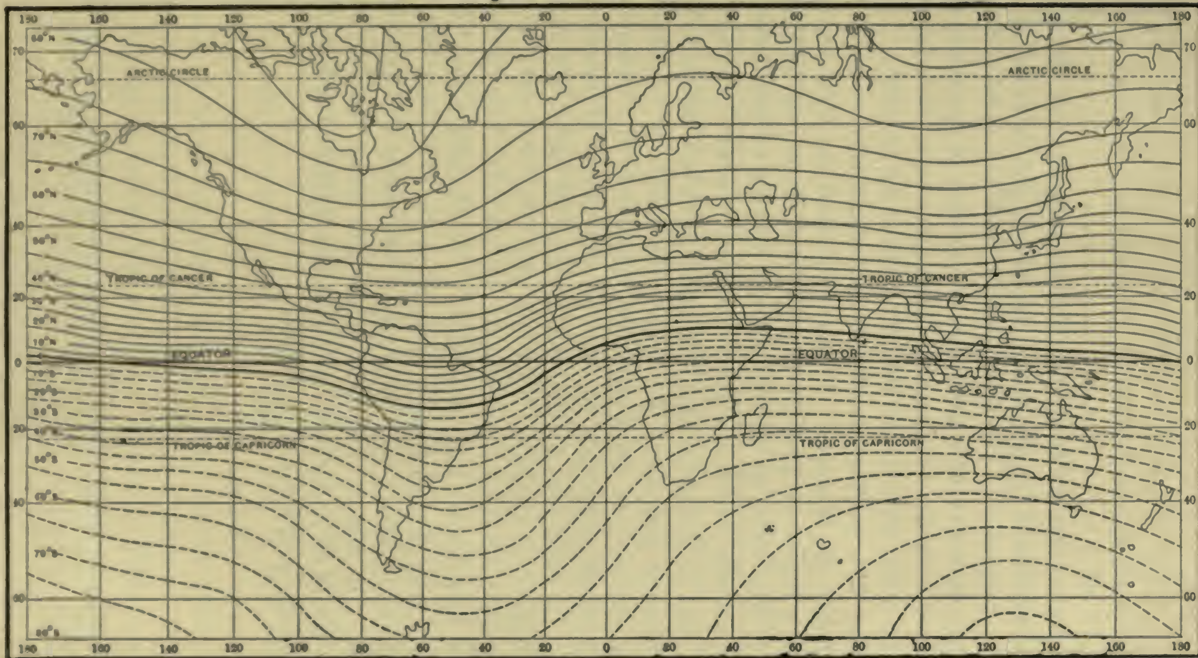


Fig. 829.

70°. This fact was not published, and Norman, in 1576, independently discovered it in London, obtaining a value of 71° 50′.

Norman was probably the first to suggest that the source of attraction is in the earth, and not in the heavens as generally supposed. He also showed that the earth's field is simply directive and produces no motion of translation, by floating a needle on water. The variations in the magnetism of the earth were discovered by Gellibrand in 1636.

The first systematic treatise on magnetism was William Gilbert's *De Magnete*. It was published in 1600, and contains a complete account of what was known as magnetism up to that time, as well as a great number of new ideas and experiments which are due to Gilbert himself. Gilbert was the first to recognize the difference between temporary and permanent magnets; to detect the effect of a change in temperature; to show that the fragments of a magnet are themselves magnets; to observe the effect of hammering, etc.; to make use of the idea of lines of force, although in an imperfect manner. The fact that iron was not the only magnetic substance was shown by Brandt, who proved in 1733 that cobalt was magnetic. The diamagnetism of bismuth was recognized by Brugmans in 1778, but the first systematic study of the subject was made by Faraday in 1845. It was he also who made the great discovery that the forces of attraction and repulsion depend fundamentally upon the surrounding medium.

ELECTROSTATICS

CHAPTER XLII

FUNDAMENTAL PHENOMENA

Introduction. — It is observed if a piece of silk is rubbed against a glass rod and is then separated from it that both now have the power of attracting small fragments of paper, of metal foil, of thread, etc., toward those portions of their surfaces which had been in contact, and that, further, if either one of them is suspended so as to be free to move, it may be attracted by the other. The silk and the glass are said to be "electrified," to have on them "electrical charges" or "charges of electricity," or, more simply, to be "charged." The same phenomena may be observed with any two portions of different kinds of matter; but with certain kinds the forces of attraction are manifested, not alone by those portions of the surface where they were in contact with the other body, but also over all their surface. This is true of metals, for instance. So, if one end of a long metal wire is charged, the forces are evident over all its length; the wire is said to "conduct" the charge, and it and similar bodies are called "conductors." By suitable means the charges on the other end of the wire may be removed; but, if the charge on the first end is continually renewed, charges will appear again at the former one, etc. This may be called, then, an electric "current," and while it is going on, many interesting phenomena occur both in the wire and outside it. We are thus led to divide the subject of Electricity into two parts: one deals with electrical phenomena when the charges are at rest, it is called "Electrostatics"; the other with the phenomena of electric currents, it is called "Electrodynamics." We shall begin with the former.

Electrical Charges. — As said above, when two portions of different kinds of matter are rubbed together and then separated, they are charged, and can produce forces which they could not when in an uncharged or neutral condition. The act of *friction* is not essential; all that is necessary is that the two pieces of matter should be brought *closely* in contact. We distinguish, too, as stated above, between "conductors" and "non-conductors." The following bodies are the commonest illustrations of conductors: all metals, either solid or liquid; water containing in solution almost any salt or acid; the human body; the earth. The following are illustrations of non-conductors: glass, silk, paper, cloths, dry wood, porcelain, rubber, sulphur. In order to produce any appreciable charge, therefore, in a conductor, it must not come in contact with the hand, but must be "insulated" by holding it in a piece of paper or cloth.

Energy of Charges. — The fact that forces are exhibited near charged bodies and that therefore work can be done by producing motion, proves that there must be energy associated with charges. This is evident also, because, as stated above, when two bodies are charged by rubbing them against each other and then separating them, one attracts the other, and this proves that in order to separate them work was required. In other words, work is necessary in order to have electrical charges. This energy which is associated with charges is not in the bodies themselves: it is in the medium which surrounds them wherever the electrical forces may be felt, that is, throughout the "electric field." This fact is proved by the phenomena of electric sparks. It is known to every one that if the electric charges are too intense, sparks take place in the medium (*e.g.* ordinary lightning), and these are due to the breaking down of the material structure of the medium. If a spark passes through a sheet of paper or a pane of glass, a hole is made in it; if the spark is in air, the molecules of its gases are broken into parts. This proves that the medium must have been greatly strained just before the sparks passed; and, if it was strained, it must have possessed potential energy. Electric forces may be shown in a vacuum; and therefore the seat of the energy of electric charges is in both the surrounding ether and the material medium immersed in it. The importance of the nature of this medium in all electrical phenomena is thus established.

Positive and Negative Charges. — If two rods of the same kind of glass are charged by means of a piece of silk, and if one is suspended horizontally in a paper sling so that it is free to turn, it may be seen, on bringing the charged portion of the other rod near it, that one *repels* the other. Whereas, if the piece of silk which was used to charge the glass rods is brought near, there is *attraction*. Similarly, if other charged bodies are brought near the suspended glass rod, some repel it and the others attract it. All those charged bodies which repel it are said to be "positively" charged; while those which attract it are said to be "negatively" charged. This amounts to a definition of positive or plus (+) and negative or minus (−) charges. Thus the experiments just described show that

FIG. 330. — A glass rod suspended in a stirrup so as to be free to turn.

glass rubbed with silk is charged positively; and that the silk is charged negatively. Similarly, in all cases, experiments show that when any two bodies are brought in contact and then separated, they are charged oppositely.

If different charged bodies are suspended in turn, it is observed that it is a general law that a positive charge attracts a negative one but repels another positive charge, and that a negative charge repels another negative one. "Like charges repel; unlike ones attract." It is found, further, that the force becomes less as the distance apart of the charges which are acting is increased.

A body which is charged positively when rubbed with some definite body may be charged negatively when rubbed with another one. And, further, the character of the charge received by a body often depends upon the condition of its surface, whether it is smooth or scratched, etc. Thus, glass is charged positively by a piece of silk, but negatively by a piece of flannel; and smooth glass may be charged positively, while, if it is rough, it may be charged negatively. By a careful study of the character of the charges produced on different bodies when rubbed with other ones, it is found that it is possible to arrange all bodies in a series, A, B, C, etc., such that if B is rubbed with A it is negatively charged, whereas if it is rubbed with C it is charged positively. Such an arrangement is called the "electrostatic series." A few of its terms are: cat's fur, flannel, glass, cotton, silk, wood, the metals, rubber, sealing wax, resin, sulphur.

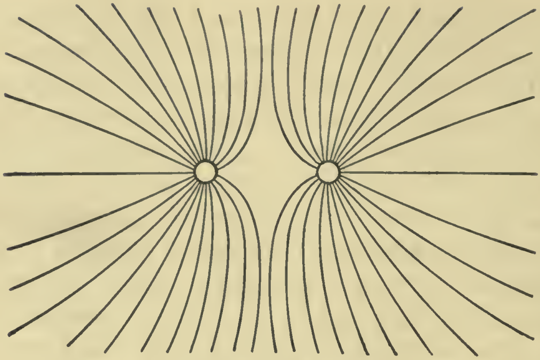
Conductors. — We say that a body is charged at any point if electrical forces are exhibited when a small piece of matter is brought near that point. If the charged body is a conductor, there are no forces shown in its interior; if it is a hollow solid, — like a hollow ball, — there are no forces in the interior region; in other words, if a conducting body is charged, the charge is entirely on its surface.

This phenomenon may be considered as due to the repulsion of a charge by a similar charge; the charges distribute themselves as far apart as possible, and, since a conductor allows charges to flow, they will all be on the surface. (This is true only after the charges have come to rest; it does not hold when there are currents.) This fact may be proved by direct experiment in many ways. Faraday made a metal box large enough to allow him to enter it and carry with him his instruments; and he showed that, however the box was charged, there were no effects inside after the charges came to rest. (Similarly, he showed that whatever electrical charges or changes he produced inside, there was no electrical force outside. The explanation of this will be given later.)

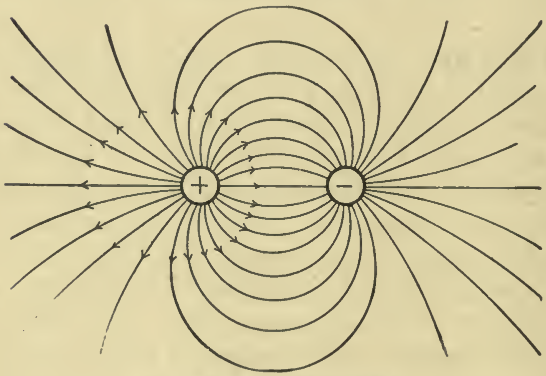
Lines of Force. — The region around charged bodies in which electrical forces may be shown is called the "electric field"; and a "line of electric force" is a line in the field such that at each of its points its tangent is the direction in which a minute body charged positively would move if left to itself. (A negatively charged body would, of course, move in the opposite direction.) If a line of force is continued, it will be found, therefore, to start from a positive charge and to end on a negative one. Two lines of force cannot cross, for that would mean that at the point of intersection a charged body would move in two directions. There are no lines of force inside a conducting body; they all end at its surface.

In Fig. 331, lines of force are drawn for several special cases. It is seen that the phenomena of attraction and repulsion and the distribution of the lines themselves may be described by saying that lines in the same direction repel each other, and that there is a tension in the lines tending to make them contract. It may not be unnecessary to state the obvious fact that these lines have no physical existence, but are merely geometrical constructions.

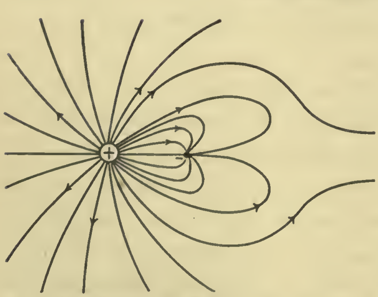
The lines of force may be mapped by a method exactly



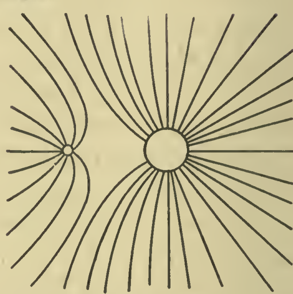
Two equal and similar charges.



Two equal but opposite charges.



Two unlike charges, the positive one four times as great as the negative one.



Two similar charges, one four times as great as the other.

FIG. 331. — Lines of electrostatic force.

similar to that described for a magnetic field. It will be shown in the next paragraph that when any piece of matter is put in an electric field it becomes electrically charged, some portions with plus electricity, others with minus. If it is an elongated body, its two ends become charged oppositely; and so, if it is short and is suspended or pivoted at its middle point, it will turn and set itself along the line of force. It plays, then, the same part in an electric field as does the short magnet in the magnetic case.

Induction. — If an uncharged body is insulated from the earth and is brought near a charged body which is also insulated, the former will exhibit electric forces. On the side nearer the charged body, it will apparently be charged with the opposite kind of electricity to that of the latter; while on its more remote side it will be charged with the same kind. Thus, if in the cut the charged body has a positive charge, the end *A* of the other body which was originally uncharged will exhibit the properties of a negative charge, while the end *B* will have those of a positive charge. If the charged body is withdrawn, these charges on the other disappear. The phenomenon is called "electrostatic induction"; and the charges are called "induced" ones.

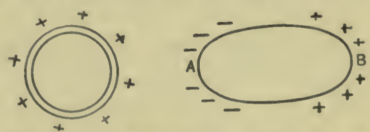


FIG. 332. — An electrified body is brought near an unelectrified one. Induction.

If the uncharged body which is brought near the charged one is a piece of a non-conductor, — *e.g.* of glass, of sulphur, etc., — its molecules are affected by the electric forces and the whole body is strained, it is said to be "polarized." The electric charges are distributed exactly as are the magnetic ones in the case of a piece of iron which is brought near a magnet. (See page 597.) There is, in fact, almost complete analogy between the case of a piece of non-conductor put in an electric field and a piece of magnetic substance put in a magnetic field. There may be one difference; after the latter is removed from the field, it remains a magnet, while in the former case the electric charges in general disappear. Cases have, however, been observed in which the electric charges remained; and so the analogy in these cases is exact.

FIG. 333. — Induced charges on a conductor.

If the uncharged body which is brought near the charged one is a conductor, it becomes charged, oppositely on its two ends or faces, as described above. There is an essential difference, however, between this case and that of a piece of non-conductor, owing to the fact that lines of force do not pass through a conductor and, therefore, end on its surface, while they can and do pass through a non-conductor. This difference will be explained more fully in the next chapter.

These induced charges on a conductor are caused by the attraction of the charge on the charged body for an unlike charge and its repulsion of a similar charge; it being borne in mind that these forces are due to the fact that when unlike charges approach each other, or when like charges recede from each other, the potential energy in the medium becomes less. Thus, if a *conductor* is joined to the earth by a con-

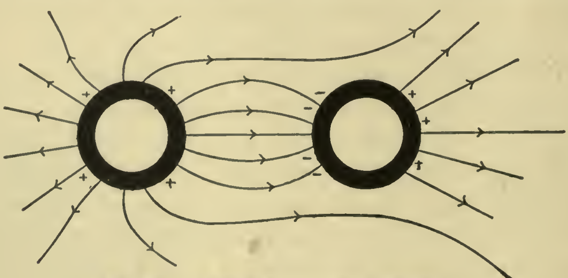


FIG. 334. — Charging a conductor by induction.

ducting wire, and if a positively charged body is brought near it, a positive charge is repelled to the earth and the conductor itself has a negative charge; if now the conducting wire is removed, the conductor retains its charge. The distribution of the lines of force is shown in the cut. This process of charging a conductor is known as "charging by induction."

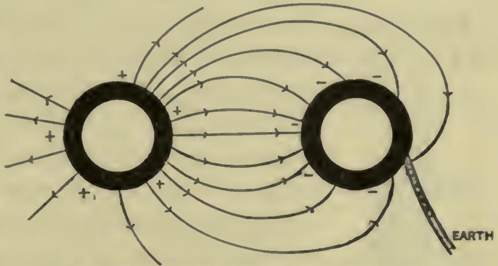


FIG. 334 *a*.—Charging a conductor by induction.

Experiments show that, if a charged body has points on its surface, the electric force in the air is greatest near them; and, in fact, if such a charged conductor is carried into a darkened room, faint sparks will be seen at the points. The charges are passing off to the particles of dust and to other small portions of matter in the surrounding air. These thus become charged with the same kind of electricity as that on the body, and are, therefore, repelled by the latter, forming a current in the air, or a wind. This is often sufficient to be felt by the hand or to blow out a candle flame. If, then, a pointed *conductor* is brought near a charged body, so that its points are toward the latter, which may be either a conductor or a non-conductor, the latter will induce charges on the former; and those on the points turned toward the charged body will escape, be drawn to the latter, and "discharge" it by neutralizing the charges on it; the other induced charges, which are like those on the body originally charged, will remain on the conductor. The final action, therefore, is as if the charge

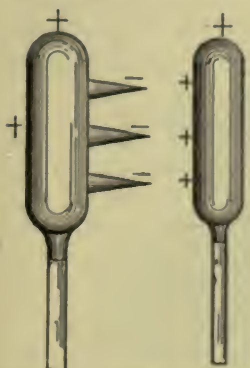


FIG. 335.—Action of points by induction.

were bodily transferred to the pointed conductor. This action of such a pointed conductor, or a "comb," is made use of in many electrical machines. (Its importance was first recognized by Benjamin Franklin (1747). It is the reason why lightning rods are always made with sharp points.)

Electroscopes. — It may be well to explain at this point one or two simple instruments which are used in the study of electric phenomena. One of the most useful of these is the "gold-leaf electroscope," which consists essentially of two vertical slender strips of thin gold foil connected at their upper ends to a metal rod which is attached to a metal plate or ball. The gold leaves are, as a rule, inclosed in a glass bottle so as to prevent any action of draughts of air. If the plate or ball is given a charge, this will spread over the leaves, and since they are now charged alike, they will repel each other, and will diverge. The angle of divergence will vary with the intensity of the force of repulsion. Further, if a charged body is simply brought near the plate (or ball), charges will be induced on the leaves and they will diverge.

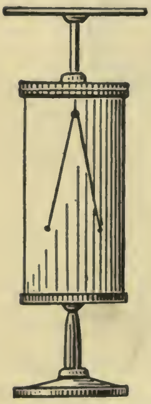


Fig. 336.—Gold-leaf electroscope.

In most gold-leaf electroscopes there are thin strips of tin foil fastened to the walls of the glass vessel and attached to the metal base of the instrument, so that if the gold leaves are diverged too far they will not communicate their charge to the non-conducting glass walls, but to the conducting strip, which will carry the charges to the outside of the instrument.

Another simple instrument is the "pith-ball electroscope." It consists of a small pith ball covered with a thin layer of metal foil and supported from a vertical metal rod by a fine wire or other conductor. If the rod is charged, it will transfer some of its charge to the pith ball, which will be repelled. The angle its supporting wire makes with the rod is a measure of the force. It is obvious that a single gold leaf could be used in place of the pith ball, or that two pith balls could be used in place of the two gold leaves in the former instrument.


FIG. 837. — Pith-ball electroscope.

Electrical Machines. — As we have seen, electrical charges may be produced by two independent methods: by friction or contact between two different bodies, and by induction on a conductor. Corresponding to these are two types of machines for producing charges continuously.

a. Friction Machine. — There are various forms of these so-called friction machines; but a description of the one shown in the cut will apply to all. There is a large glass plate pivoted on an axle, which is clasped at one point by two metal clips lined with leather; so that, as the wheel is turned, the glass becomes charged positively and the clamps negatively. The charges are removed from

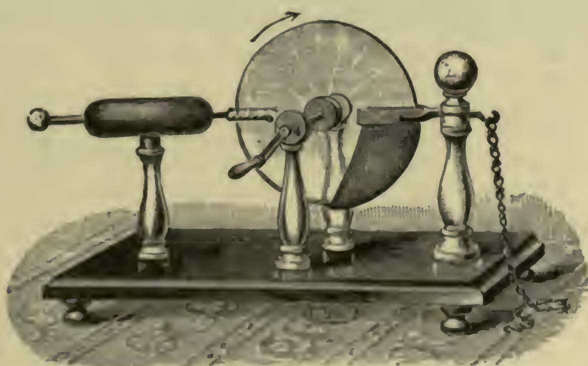


FIG. 838. — Friction electrical machine.

the latter by joining them to the earth, and from the former by the use of a pointed conductor or "comb." A positive charge is thus accumulated on the large conductor which is joined to the comb.

b. Induction Machine. — The simplest form of instrument for producing charges by induction is the so-called "electrophorus," which was invented by Volta about 1775. It consists of a thick plate *A*, of some non-conducting substance such as glass or hard rubber, which rests in a metal base *B*; and of a loose metal cap *C*, provided with an insulating handle *D*. In using the instrument, the cap is removed and the upper surface of *A* is charged by friction with a piece of flannel or cat's fur; let it be assumed that it is thus charged negatively. This charge will induce a plus charge on the upper surface of the metal base *B*, and the induced minus charge flows off to the earth. (The function of this induced charge on *B* is by its attraction for the charge on *A* to prevent the latter from escaping or leaking.) The metal cap *C* is now lowered on *A*. Actually, it touches it at the most in only a few points and so does not receive any appreciable charge from *A* directly. But the charge on *A* induces a positive charge on the lower side of *C* and a negative one on the upper side. This cover *C* is now touched with the finger or otherwise connected to the earth; so the negative charge is removed, and only the positive one remains. Connection with the earth is now broken, and if the cap is lifted by its handle, it will carry with it its positive charge. This charge may be transferred to some conductor; and the cap being discharged may be replaced on the plate *A*, which still retains its charge. So the process may be repeated indefinitely.

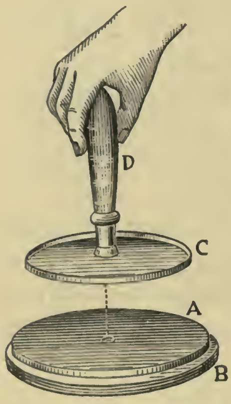


Fig. 339. — Electrophorus.

Machines have been made by which these various steps are carried out automatically. One of these is shown in the cut. The explanation of its action is simple, but is so long that it need not be given here. It may be found in almost

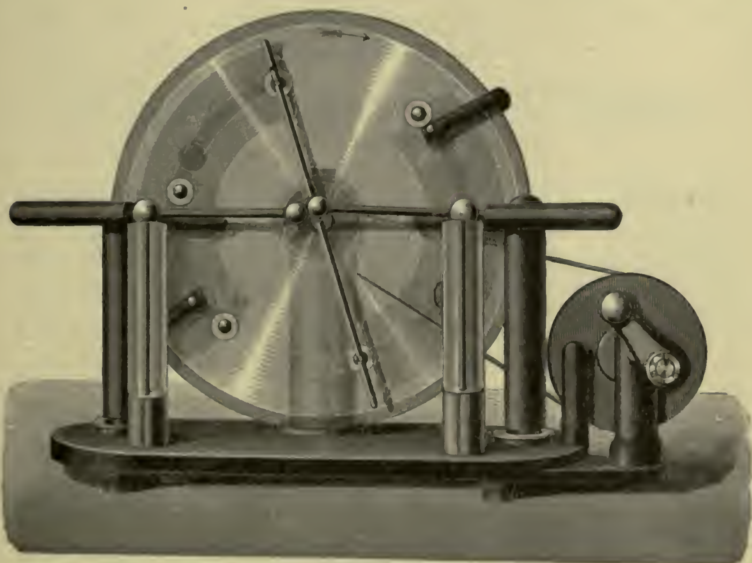


FIG. 340. — Induction electrical machine.

any special treatise on Electricity, such as Cumming, *Electricity;* Perkins, *Electricity and Magnetism;* S. P. Thompson, *Elementary Lessons in Electricity and Magnetism*, etc.

CHAPTER XLIII

ELECTRIC FORCE; MEASUREMENT OF ELECTRIC QUANTITIES

Quantities of Electricity. — The fact that an electric charge is a quantity to which a numerical value may be assigned is suggested by many experiments. If a hollow metal vessel, like a can, is placed on top of a gold-leaf electroscope, and if a charged body is lowered into it by means of a silk fibre (or other non-conductor), the leaves diverge owing to induction; but the amount of the divergence is found to depend upon the charge lowered, not upon its position inside the vessel. Further, if another similar charge is lowered into the vessel, the leaves diverge still more, but the amount of this does not change if the two charged bodies are brought in contact, or even if they touch the walls of the vessel. We are thus led to speak of the "quantity" or "amount" of charge, or of electricity. It should be noted that the last experiment described shows that the *total quantity* of electricity on two or more conductors is the same before and after they touch. Thus we speak of the "Conservation of Electricity."

Equal Quantities of + and − always produced. — If two uncharged bodies are lowered into the vessel, *e.g.* a piece of silk and a piece of glass, the leaves do not diverge, even when these two bodies are touched or rubbed together and then separated. But, if one of them is now removed, the leaves do diverge, showing that the two bodies were charged, but with exactly equal amounts of opposite kinds of electricity. Similarly, if an insulated conductor is lowered into the can in which there is already a charged body, there is no change in the divergence of the leaves, thus proving that the two induced charges are of exactly equal amounts of opposite kind. Thus it can be stated as a general law that whenever a charge of any kind is produced, an equal charge of the opposite kind also appears.

Faraday's "Ice-pail Experiment." Dielectrics. — An interesting experiment in this connection is one due to Faraday. If a charged body is lowered by means of a silk cord into the interior of a nearly closed hollow conducting vessel, which is joined by a wire to an electroscope, the leaves of the latter will diverge; but, as said above, the amount of the divergence does not change as the charged body is moved about inside the vessel, or even if the two touch. (Faraday in his original experiment used a metal "ice-pail" as the vessel.) If the charged body is a conductor, it will lose its charge when it touches the metal vessel, because the charge will all go to the outside of the hollow conductor. (See page 628.) But since the divergence of the leaves is not affected, there can have been no change in the charge on this conductor. The explanation is that when the charged body is lowered into the hollow vessel, it induces an *equal* amount of electricity on the inner wall of the vessel of a kind opposite to its own — and therefore also an equal amount of the same kind as its own on the outside of the wall of the vessel; so, when the charged conductor touches the inner wall of the vessel, equal amounts of plus and minus charges pass to the outer surface, and there is consequently no change in the external conditions. It is ob-

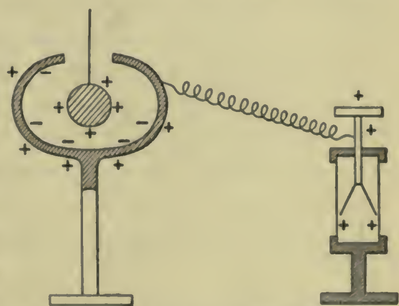


FIG. 341. — Faraday's ice-pail experiment.

served, further, that when the charged body is lowered into the hollow vessel, the divergence of the gold leaves is the same whatever non-conducting medium is used to fill the vessel or is present in it: air, oil, sulphur, etc. Thus, electrical effects are transmitted *through* these various substances; and for that reason Faraday called them "dielectrics." (Actually there is a distinction between the idea of a dielectric and that of a non-conductor or insulator, but it need not be emphasized here.)

Law of Force. — The force between two charged bodies is found to depend upon the amounts of their charges, their distance apart, and the nature of the surrounding medium or dielectric. So far as distance is concerned, Cavendish proposed the relation that the force varied inversely as the square of the distance. He showed by most ingenious mathematical reasoning that, if this were true, the charge on a spherical conductor must be entirely on its outer surface, even if there were bodies in its interior which were joined with it; and he then proved by direct experiment that the charge was entirely on the outer surface. (This was previous to 1773.) This same suggestion as to the law of force was made independently by Coulomb (1785); and he verified it by direct experiment, placing charges at different distances apart.

A unit electrical charge may be defined in a manner similar to that used for a magnetic charge. On the C. G. S. system of units a unit charge is defined to be such a one that, if it is at a distance of 1 cm. in air from an equal charge, the force is 1 dyne. This is called the "C. G. S. Electrostatic Unit." Then, if a charge whose value is e is at a distance of r cm. from a charge e_1 in air, the force in dynes between them is $\frac{ee_1}{r^2}$. But it is found that the force depends upon the surrounding medium; and this is expressed by writing the value of the force in dynes as $f = \frac{ee_1}{Kr^2}$, where K is a quantity which is characteristic (and constant) for any one dielectric. It is called the "dielectric constant." Using the system of units defined above, the value of K for air is *one*. Its value for all other dielectrics (with the exception of a few gases) is greater than for air, as may be proved by directly measuring forces in different media.

Tubes of Induction. — The "intensity" of an electric field of force at any point is defined to be the value of the force which would act on a unit positive charge if placed at that point. Exactly as in the case of a magnetic field, too, *tubes* bounded by lines of electric force can be drawn; and if they are of the proper size they will end on unit plus and minus charges. They are called "tubes of induction." Since there is no force inside a closed conductor, even if it is charged, these tubes must end on its surface, not traverse it. They do, however, pass through a dielectric, as is shown by Faraday's experiment described on page 639. In this the tubes all start from either the charge which is introduced or from the inner wall of the vessel, and end on the other charge.

We can, in fact, define a real charge as one which originates tubes of induction. Thus, in the cases of induction described on page 631 the tubes from the charged body pass into and through the dielectric body; but they end on the conducting one, and an equal number leave the other end. Thus, the charges on the latter body are *real;* while in the former case they are only *apparent*, the forces manifested being due, as may be shown by methods of the infinitesimal calculus, to the fact that the tubes are passing from one dielectric into another.

The number of unit tubes per square centimetre at right angles to the field at any point is, as in the case of magnetism, proportional to the intensity of the field at that point.

Explanation of Attraction and Repulsion. — The explanation of electric attraction and repulsion may be given in terms of energy exactly as was done for magnetic forces; the constant K taking the place of μ. Since K is less for air than for all other dielectrics, a small piece of any such dielectric is more permeable for tubes than is air, and is attracted by a charged body if air is the surrounding medium. Further, since there is no force and therefore no energy inside a conducting body, a small piece of a conductor is attracted even more than a piece of non-conductor. The action of charges on each other has already been discussed.

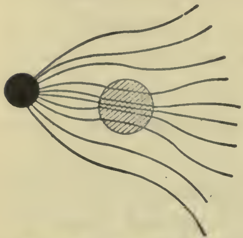


FIG. 342. — Effect of introducing a piece of glass or paper in an electrostatic field in air.

Electric Potential. — The properties of electric charges and the condition for their being in equilibrium may be expressed in a different manner. When a charge is moved in an electric field, work is done, either at the expense of the energy of the field or against the forces of the field. Thus, if a plus charge moves in the direction of the field of force (or a minus charge in the opposite direction), the field loses energy and the charged body gains kinetic energy as it moves; if a plus charge is moved against a field of force (or a minus charge in the opposite direction), work is done by some outside agency, and the energy of the field is increased.

We have a mechanical analogy in moving a body toward or away from the earth; if it is raised, work is done against gravitational forces and the potential energy is increased; if it falls, it gains kinetic energy at the expense of the potential energy. We may define the potential energy of a body of unit mass with reference to the earth when placed at any point as the "potential at that point," or we may say that the potential at a point is the work required to raise a body of unit mass from the earth to it; and we may describe the gravitational forces in terms of this quantity. The potential is evidently constant at all points of any horizontal plane; and the higher the plane is from the earth, so much the greater is the potential. Such a surface of constant potential is called an "equipotential" one. Evidently there is no change in energy as a body moves along such a surface; but the change as a body of unit mass is moved from a point P_1 in an equipotential surface, the potential of whose points is V_1, to a point P_2 in a second equipotential surface whose potential is V_2, is $V_2 - V_1$, and is entirely independent of the path of the motion. (See page 108.) The line of action of the force at any point is perpendicular to the equipotential surface through that point; because, if it were inclined to it, there would be a component force in the surface, and work would be required to move a body against it, which is contrary to the idea of an equipotential surface. The direction of the force is, obviously, from points of high to those of low potential.

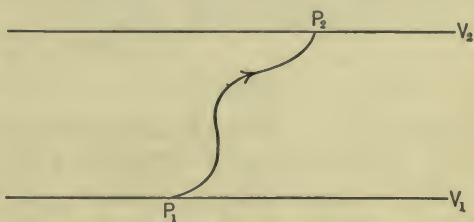


FIG. 343. — V_1 and V_2 are two horizontal planes.

Similarly, in the case of electrical phenomena, we may choose the earth as our standard body, since it is a conductor, and is so large that its electrical condition may be regarded as permanent, and may define the "electric potential" at any point in an electric field as the work required to carry a unit plus charge from the earth to that point. It is not necessary to specify any particular point on the earth, because the potentials of all points of a conducting body are the same, if the charges are at rest. If this were not so, it would require work to carry a charge from one point to another in the conductor; this would presuppose that there was an electrical force in a conductor; and, as we know, this is not true. So, since the earth may be regarded on the whole as a conductor, all points of its surface are at the same potential, whose numerical value is zero in accordance with the definition of potential given above. (This does not imply a zero *amount* of anything; for potential is not a quantity which can be measured. See page 11. We give it a number, just as we give temperature a number. 0° temperature does not mean a zero amount of anything, but indicates a temperature which serves as the starting point of a thermometer scale. So the potential of the earth is 0, because the earth is the body of reference. Actually the earth is not a good conductor, and there may be local differences of potential.) Similarly, the potential at any point far removed from the electric charges, that is, at "infinity," is zero, because no work would be required to move a unit charge from such point to one on the farther side of the earth, where by definition the potential is zero.

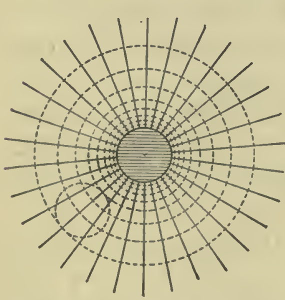


Fig. 344. — Lines of force and equipotential surfaces around a charged spherical conductor.

We can draw equipotential surfaces in the field of force; the lines of force are at right angles to them; and the direction of the lines is from high to low potential. Thus, if the field is due to a charged spherical conductor, whose complementary charge is at a great distance, everything is symmetrical with reference to its centre; the equipotential surfaces are concentric spheres, and the lines of force are portions of radii starting from the spherical conductor.

If the charge on the conductor is plus, the potential at points near it is higher than that at those more distant; if the charge is negative, just the reverse is true. Thus, if a plus charge is put at any point, the potential of all points near by is raised; while the contrary is the case if the charge is negative.

Induction. — We can thus explain the appearance of induced charges on a conductor. Let a positively charged body be brought near an insulated uncharged conductor *AB*. All points of this must be at the same potential since it is a conductor; but if the conductor were absent, the potential at a point *A* near the charged body would be higher than at a point *B* which is more remote; consequently, if the potentials at *A* and *B* are to be the same, a *negative* charge must appear at *A* so as to lower its potential, and a *positive* charge at *B* so as to raise its potential. Or, again, since when the conductor is absent, the potential at *A* is greater than at *B*, the electric force is in the direction from *A* to *B*; and, when the conductor is introduced, a plus charge moves in the direction of the force toward *B*, and a negative charge moves in the opposite direction toward *A*.

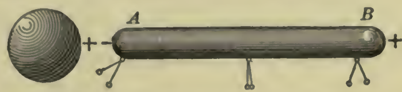


FIG. 845. — Electrostatic induction.

Further, if the conductor is joined to the earth by a wire, its potential must be zero; but under the influence of the

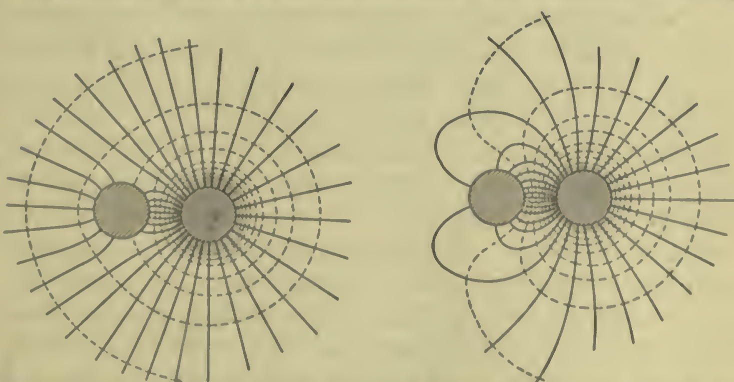


FIG. 846. — Effect upon lines of force and equipotential surfaces of introducing a spherical conductor in the field and then joining it to the earth.

charged body alone it would be some positive amount; therefore, in order to lower it to zero, a *negative* charge must appear on it. (The same explanation can be given of the inducing action of a negative charge.)

Distribution of Charges. — The fact that the potentials at all points of a conductor on which the charges are at rest are the same is a consequence, as was shown above, of the fact that there are no forces in a closed conductor. This may be expressed in a different manner: the charges on a conductor are all on the surface, and they are so distributed that the intensity at any point inside is zero, or, what is the same thing, that the potential at all points is the same. Thus, consider a closed conductor of any shape, and let P be any point in its interior and Q_1, Q_2, Q_3, etc., be any points of its surface. The charge at Q_1 is at the distance r_1 from P; that at Q_2, at the distance r_2, etc. So, if A_1, A_2, A_3, etc., are *small* areas at Q_1, Q_2, Q_3, etc.; and if d_1, d_2, d_3, etc., are the values of the *surface density* of the charges at these points, *i.e.* the charge per unit area, the intensity at P, or the force acting on a *unit* plus charge, if placed there, is the geometrical sum of $\frac{d_1A_1}{r_1^2}$, $\frac{d_2A_2}{r_2^2}$, $\frac{d_3A_3}{r_3^2}$, etc. This sum must be zero. By considerations of this kind it may be proved that the surface density at a point is greater than over a plane surface. (See page 633.)

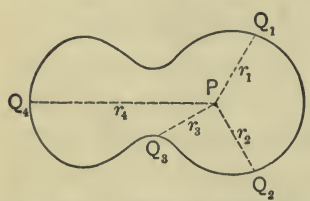


Fig. 347. — Diagram illustrating the fact that the force inside a closed conductor is zero.

Sparks. — One of the commonest phenomena associated with electric charges is that of sparks. They are occasioned, as has been explained, by the mechanical rupture of the material medium in an electric field; and they prove the existence of a great strain due to the electric forces. The intensity of the field at any point may be expressed in terms of the potential. If V and $V + \Delta V$ are the potentials at two neighboring points at a distance apart Δx, the electric force is in the direction from the second point to the first; and if R is its numerical value $R\Delta x = \Delta V$, because each member of the equation expresses the work required to move a unit charge from one point to the other. Since R is actually in the direction in which V *decreases*, the exact formula is $R\Delta x = -\Delta V$; or $R = -\frac{\Delta V}{\Delta x}$. Consequently, if the intensity is great, there must be a great fall of potential in a small distance. Thus, if the difference of potential between two conductors is high, there is danger of a spark passing between them, and a connection may be found by experiment between the potential difference and the spark length in any dielectric under definite conditions. A limit is therefore fixed by the electric properties or "strength" of the air for the value to which the potential of a conductor may be raised; for, if it is exceeded, a spark will pass to the earth or to particles of foreign matter in the air.

When a spark passes between two conductors, its path through the air is an excellent conductor; and therefore both bodies are brought to the same potential. The potential of one is raised by the passage to it of a certain amount of positive electricity, or by the withdrawal from it of a certain amount of negative electricity; and that of the other is lowered by the opposite process. (See Electric Currents, page 663.) The luminous character of a spark or discharge in any gas is due to the luminosity produced by the electrical changes which accompany the disruption of the molecules and the conduction of the current.

Capacity of a Conductor. — If we consider an insulated conductor by itself in space, it is evident that if it is charged positively, it will itself have a positive potential, and that if its charge is increased, so is its potential, because a greater amount of work would be required to bring up to the conductor a unit charge from the earth. If the charge is doubled, so is this amount of work, and therefore so is the potential, etc. We may express this fact in a formula, writing e for the charge, V for the potential of the conductor, and C as a factor of proportionality, viz., $e = CV$. This quantity C is called the "capacity" of the isolated conductor; it may be defined, as is seen from the formula, as equal in value to that charge which would raise the potential of the conductor by a unit amount. It is evident from general considerations that C must depend upon the shape and dimensions of the conductor, and upon the dielectric constant of the surrounding dielectric.

If air is the dielectric and if the conductor is charged with a quantity, e, the potential V_A is, by definition, the work required to carry to it a unit plus charge from the earth; while if the dielectric has the value K, the forces are diminished K-fold (since the electric forces vary inversely as K), and the potential of the conductor, V_K, or the work now required to bring up a unit plus charge, is less than V_A in the ratio $1 : K$; or, $V_A = KV_K$. So if C_A and C_K are the capacities in the two cases, $e = C_A V_A = C_K V_K$; and hence $C_K = KC_A$. In words, the capacity of a conductor varies directly as the dielectric constant of the surrounding medium.

The connection between the capacity of a conductor and its shape and size may be deduced in certain simple cases by means of the infinitesimal calculus. Thus, it is known for a sphere, an ellipsoid, a cylinder, etc. The capacity of a sphere of radius a in air is numerically equal to a; and, therefore, in any other medium it is Ka.

If a charge is distributed over two spherical conductors of radii r_1 and r_2, which are in contact, their potentials are the same, but their charges are different. If the dielectric is air, we have, writing e_1 and e_2 for the charges and V for the common potential, $e_1 = r_1 V$, $e_2 = r_2 V$; and the surface densities of the charges on the two spheres are $\frac{e_1}{4\pi r_1^2}$ and $\frac{e_2}{4\pi r_2^2}$, *if we assume the distribution over each to be uniform.* Calling these d_1 and

d_2, $d_1 = \frac{r_1 V}{4\pi r_1^2} = \frac{V}{4\pi r_1}$; $d_2 = \frac{r_2 V}{4\pi r_2^2} = \frac{V}{4\pi r_2}$; or $d_1 : d_2 = \frac{1}{r_1} : \frac{1}{r_2}$. This indicates that if $r_1 > r_2$, $d_1 < d_2$. So, if the *curvature* is great, the surface density is great. A point on the surface of a sphere may be compared roughly with a small sphere attached to it; and so we see why the surface density of the charge on a point is so great.

Energy of a Charge. — The energy of a charged conductor is located in the surrounding dielectric; but its numerical value can be expressed in terms of the charge, e, and the potential, V, of the conductor itself. We can imagine the conductor as being originally uncharged, and the process of charging as consisting in the bringing up to it from the earth a series of minute charges. In this manner the charge gradually increases from 0 to its final value, e; and the potential rises from 0 to V. Since the potential at an instant varies directly as the charge at that moment, the mean value of the potential during the process is $\frac{1}{2}V$; that is, the work done in charging the conductor is the same as if the whole charge, e, were brought up against this potential of $\frac{1}{2}V$, instead of the small amounts having been brought up against the continually increasing potential. By definition, the potential is the work required to move a *unit* plus charge from the earth up to the conductor, and so the work required to move the charge e is the product of e and the potential. In the present case, then, the work done is the product of e and $\frac{1}{2}V$, or $\frac{1}{2}eV$. This is the value of the energy of the field. It should be noted that in this mode of considering the charging of the conductor, an equal charge, $-e$, of the opposite kind, is left on the earth, whose potential is zero.

Similarly, the energy due to a charge $-e$ whose potential is V is $-\frac{1}{2}eV$. (This does not mean that there is such a thing as *negative* energy; for if a charge $-e$ is by itself in space, its potential V has a negative value.) So, in general, if there are two conductors with equal and opposite charges, $+e$ and $-e$, at potentials V_2 and V_1, the energy in the dielectric surrounding them is $\frac{1}{2}e(V_2 - V_1)$.

The energy of an isolated conductor can be expressed in another form, which is often useful. Calling it W, the formula is $W = \frac{1}{2} eV$; but $e = CV$, so we may write $W = \frac{1}{2}\frac{e^2}{C}$ or $W = \frac{1}{2} CV^2$. The capacity is a constant for a given conductor in a given dielectric, and is independent of the charge or its potential; so these formulæ show that the energy is independent of the sign of the charge, depending simply upon its numerical value.

Mechanical and Thermal Analogies. — An analogy may be drawn between electrostatic potential and fluid pressure, which is useful. A fluid, either gas or liquid, always flows from points of high to those of low pressure: a positive electric charge moves from points of high to those of low potential. When a gas is compressed by a pump into a vessel of any kind, the pressure continues to increase until a point is reached at which the vessel breaks or the gas leaks, and this maximum pressure does not depend upon the *size* of the vessel, but upon its strength, etc.; when the charge of a conductor is increased, the potential rises, and a condition is finally reached when a spark passes or the charge leaks off, but this maximum potential is determined by the "strength" of the surrounding dielectric, not by the size or capacity of the conductor.

Similarly, heat energy always flows from bodies at high temperature to those at low; and, if a small flame is maintained at as high a temperature as a large one, it is just as useful.

Condensers; Capacity. — Owing to the liability of a charged conductor to lose its charge if its potential is high, a method has been devised by which the conductor may keep its charge unaltered, but may have its potential lowered. When this is done, its charge may be increased before there is again danger of its escaping. The apparatus is called a "condenser." The general principle, then, is to make use of any processes which will decrease the potential of a charged conductor.

If the conductor is a plate and is charged positively, the charge will be distributed the same on its two sides, if it is isolated; but if, as shown in the second figure in the cut, another conducting plate is brought near it, minus and plus charges will be induced on this, and an additional amount of the plus charge on the first plate will be attracted around to the face opposite the second one. As a result, the potential of the first plate is lowered; because if a unit plus charge is brought up to it from the earth, less work is required than before, owing to the action of the induced *negative* charge on the second plate. The induced *plus* charge on this plate serves to keep the potential high; and if it is removed by joining the plate to the earth, the potential is lowered still more. This potential is now the work required to carry a unit plus charge across from the second plate, whose potential is zero, to the first one; and it may be decreased again if a dielectric, such as glass, is substituted for the air between the plates; for if K is increased, the force required to move a charge is decreased. Therefore, in the end practically all the charge on the first plate is on the face toward the second one, and there is an equal amount of electricity of the opposite kind on that face of the second plate which is toward the first one; and the potential of the first plate is greatly below that which it was originally.

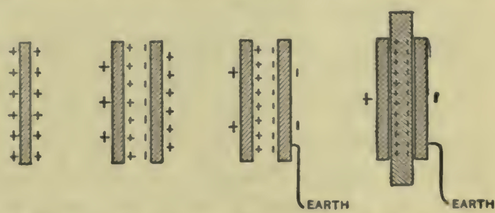


FIG. 848. — Different steps in the construction of a condenser.

If the connection with the earth is removed, and the whole apparatus is moved elsewhere, possibly near some other charged bodies, the potentials of the two plates will change, but *their difference remains constant*, because it equals the work required to move a unit positive charge from one plate to the other; and we may assume that the two plates are so close together that this work depends simply upon the charges on them.

If, then, $+e$ and $-e$ are the charges on the two plates, and V_2 and V_1 are their potentials, $V_2 - V_1$ is a constant so long as e does not change. If it varies, so does $V_2 - V_1$; and one is proportional to the other. We may therefore write $e = C(V_2 - V_1)$, where C is called the "capacity of the condenser." It is evident that this quantity is a constant for a given combination of two conductors of definite size and shape separated by a definite dielectric of a definite thickness; and it has, of course, no connection with the similar constant for a single isolated conductor.

The numerical value of the capacity may be calculated for many simple cases. A condenser consists essentially of two similarly shaped conductors placed close together and separated by a dielectric, such as glass, mica, etc. The commonest forms are those in which the conducting plates, or "armatures," are parallel plates, concentric spheres, or coaxial cylinders. A few facts in regard to the capacity of these condensers are evident from the formula of definition: $C = \frac{e}{V_2 - V_1}$. The capacity must vary directly as the dielectric constant of the two dielectrics, because for a given value of e, $V_2 - V_1$ varies inversely as K; the capacity must vary directly as the area of the armatures, because for a given value of $V_2 - V_1$, e varies directly as this area; the capacity must increase as the armatures are made to approach each other, because for a given value of e, $V_2 - V_1$, varies inversely as the distance apart of the armatures. Exact calculations by means of the calculus show that the capacity, on the C. G. S. system, of two parallel plates of area A and at a distance d apart is $C = \frac{KA}{4\pi d}$; that the capacity of two concentric spheres of radii r_1 and r_2 is $C = \frac{Kr_1r_2}{r_2 - r_1}$; and that the capacity of two coaxial cylinders of radii r_1 and r_2, *per unit length*, is $C = \dfrac{K}{2\log\dfrac{r_2}{r_1}}$.

This formula for two parallel plates holds, of course, for those portions of the plates only which are not near the edges; for it is only over

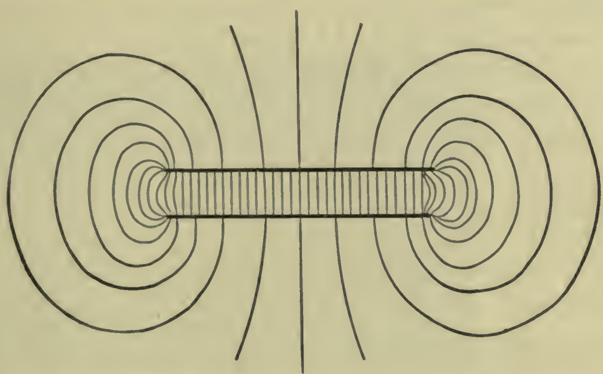


FIG. 349. — Lines of electrostatic force in the case of a plate condenser, the dielectric being air.

those portions that the conditions are uniform. This is shown in the cut, which represents the lines of force. Therefore if such a condenser is to be used for purposes of measurement, a device is employed, invented by Lord Kelvin, which consists in having a portion of one plate near its centre cut out from the rest of the plate and supported independently of it. The charges on this disc are uniformly distributed, and the formula given above applies to its capacity.

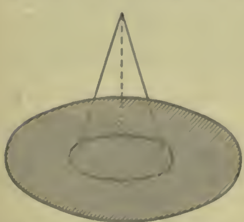


FIG. 350. — One plate of a "guard ring" condenser.

One form of condenser most often used for qualitative experiments is the so-called "Leyden jar." It consists of a glass bottle whose inner and outer surfaces, except for their

FIG. 351. — Leyden jar.

upper portions, are coated with tin foil; the bottle is closed by a wooden stopper through which passes a brass rod tipped with a knob and ending below in a chain which makes contact with the inner coating. The capacity of such a condenser may be calculated fairly closely from the formula for two parallel plates.

Discharge of Condensers. — If a conductor carrying a knob is attached to each armature of a condenser, and if these two knobs are gradually pushed toward each other by means of some insulating rod, a limiting distance will be reached at which a spark will pass between them. This distance varies with the difference of potential $V_2 - V_1$; the greater it is, the longer the sparking distance. When the spark takes place, the potentials of the two armatures become the same, and since the plates had equal and opposite charges, the final charge on both plates is zero: the condenser is said to be "discharged."

Experiments show that there are two types of discharge. In one, the charge on each plate becomes gradually less, the potentials gradually approach the same value; and in the end the charges have disappeared and the potentials have become equal. This is known as a "steady" discharge, and occurs if considerable opposition is offered to the discharge; for instance, if one of the knobs between which the discharge takes place is joined to its armature by a poor conductor, such as a wet thread. In the other type of discharge, the charge on either plate disappears rapidly, then it becomes charged again with electricity of the *opposite* kind, this disappears, it becomes charged again as it was originally, etc.; each successive charge being less in amount than the previous one. This is called an "oscillatory" discharge, for obvious reasons; and it occurs if the opposition to the discharge is small. (We have mechanical analogies of a steady discharge in a pendulum moving in a viscous liquid such as molasses, and of an oscillatory one in a pendulum vibrating in air.)

If the spark of the discharge is viewed in a revolving mirror, that of a steady discharge appears like a broad continuous band, gradually fading away, while that of an oscillatory one is seen to consist of a series of distinct sparks, showing that the electric current is intense, then vanishes, rises again, etc.

Electric Waves. — It is proved by experiment that, when an electric oscillation of this kind takes place, there are dis-

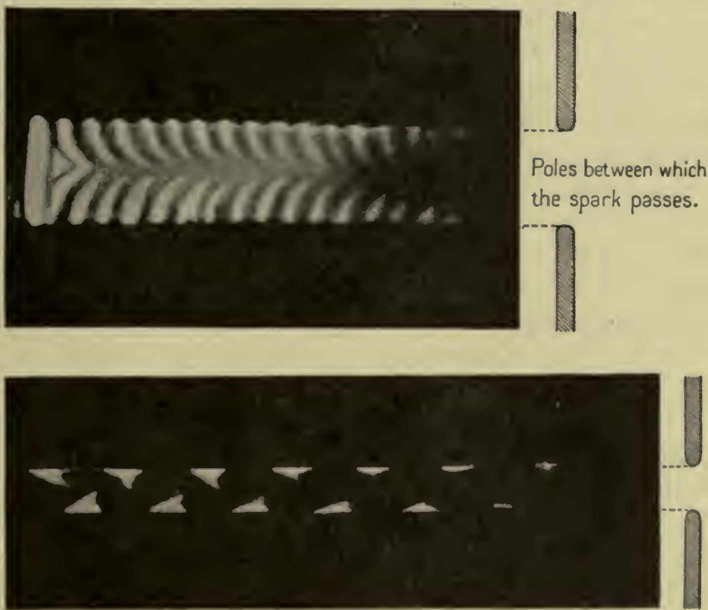


FIG. 852. — Photographs of an oscillating spark when viewed in a revolving mirror.

turbances of the nature of waves produced in the surrounding medium. Before the discharge occurs the electric intensity at any point near by has a definite value, but as the discharge goes on, the electric intensity varies in amount and direction, and this oscillation produces waves. Experiments have shown that these waves travel in air with the velocity of light and that they are transverse. They can be reflected, refracted, diffracted, polarized, etc. Their wave lengths can be measured, and waves as short as a small fraction of a centimetre have been obtained; as a rule, however, they are much longer. These waves may be detected by means which will be described later. One method may, however, be mentioned here; if they fall upon two conductors which are close together, they will — under suitable precautions — cause minute sparks to pass between them.

These "electro-magnetic" waves, so called, were first investigated by Joseph Henry, in 1842, but were rediscovered many years later by Hertz. They serve a commercial purpose in the various systems of wireless telegraphy which are now in daily use.

Since the waves travel in air with the velocity of light, it is proved that they are ether waves. One would not expect them to travel in solid dielectrics such as glass with the same velocity as does light, because their wave lengths are so different, and it has been shown in Chapter XXX that the velocity of waves varies greatly with the wave length in all solid or liquid media.

The medium, then, which serves as the means by which magnetic and electric forces are manifested, which is the "carrier" of the tubes of induction, is the luminiferous ether. This fact was first suspected by Faraday, but was proved by Maxwell by an indirect method.

Condensers (*continued*). — The energy of a charged condenser is, from what was proved above, $\frac{1}{2}e(V_2 - V_1)$. This may be written $\frac{1}{2}\frac{e^2}{C}$ or $\frac{1}{2}C(V_2 - V_1)^2$. Since the field of force is confined almost entirely to the space between the two armatures, as is shown in the cut for a parallel plate condenser, the energy is located there also.

Condensers are often joined together so as to increase their action. There are two general methods of doing this. Let the two plates of the first condenser be called P_1 and Q_1;

those of the second, P_2 and Q_2, etc.; and let them be always charged in such a manner that P_1, P_2, etc., are positive, and Q_1, Q_2, etc., are negative. Then, if P_1, P_2, etc., are connected by wires, and Q_1, Q_2, etc., are also connected, the condensers are said to be "in parallel." Whereas, if Q_1 is joined to P_2, Q_2 to P_3, etc., they are said to be "in series."

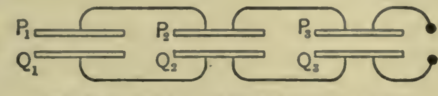


FIG. 853. — Three condensers joined in parallel.

Let the condensers all have the same capacity and all be charged alike before they are connected; then their *differences* in potential are all equal, but the potentials of any two plates, *e.g.* P_1 and P_2, need not be the same. Let V_1 and U_1 be the potentials of P_1 and Q_1, V_2 and U_2 those of P_2 and Q_2, etc. Then in all cases $V_1 - U_1 = V_2 - U_2 =$ etc.

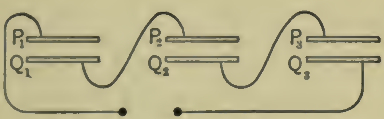
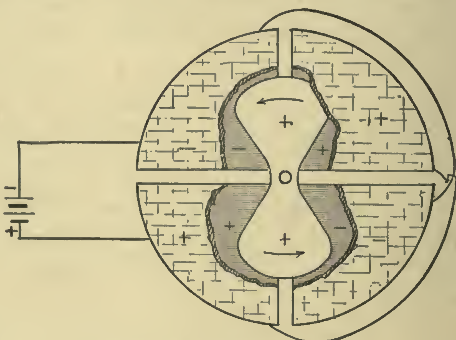


FIG. 854. — Three condensers joined in series.

If the condensers are joined in parallel, $V_1 = V_2 = V_3 =$ etc., and $U_1 = U_2 = U_3 =$ etc.; so, it is exactly as if the condenser were made up of two large plates, one, P_1, P_2, P_3, etc., the other, Q_1, Q_2, Q_3, etc. The difference of potential is $V_1 - U_1$, and the total charge on either "plate" is ne, where n is the number of condensers connected and e is the charge on each plate; so the capacity is increased n times. Thus, joining in parallel gives an increased *quantity*, but does not change the difference in potential.

If the condensers are joined in series, $U_1 = V_2$, $U_2 = V_3$, etc.; so, if there are n condensers, $V_1 - U_n = n(V_1 - U_1)$. Thus if P_1 and Q_n are connected so as to discharge the condenser, the difference of potential is increased n-fold; but the *quantity* of electricity discharged is the same as for a single condenser. Since the distance between two conductors at which a spark will take place is increased if their difference of potential is increased, joining condensers in series increases their sparking distance. (When two or more condensers are joined in series, the minus charge on Q_1 does not combine with the equal plus charge on P_2, etc., until P_1 is joined to Q_n. Before this, the minus charge on Q_1 is held in place by the attraction of the plus charge on P_1, etc.)

Electrometers. — Before we can explain how the various electric quantities are measured, it is necessary to describe an instrument which enables us to measure differences in potential.



Principle of quadrant electrometer.

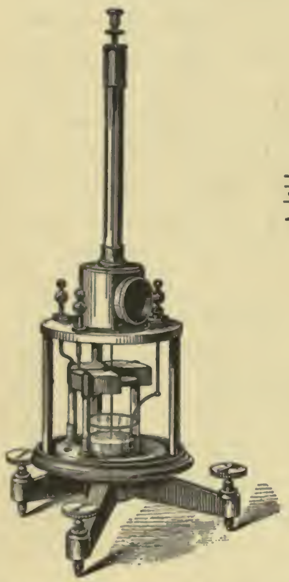


Fig. 855. — Thomson's quadrant electrometer; one of the quadrants is removed so as to show the "needle."

Such an instrument is called an "electrometer." There are many forms which may be used to measure the *ratio* of two differences in potential; so that, if one is known, the other may be calculated. The best of these is the "quadrant electrometer," which was invented by Lord Kelvin, then William Thomson. It consists, as shown in the drawing, of a cylindrical metal box which is divided by two transverse cuts into four "quadrants," and of a horizontal metal "needle" shaped like a solid figure eight, which is suspended by a fibre. The pairs of diagonally opposite quadrants are connected by wires, and the needle is raised to a high potential by some electrical machine. If the difference of potential of two plates of a condenser is to be measured, each is joined to a pair of quadrants; and the needle, which takes a symmetrical position with reference to the quadrants when they are not at different potentials, will now move so as to enter one pair, until it is brought to rest by the torsion in the fibre. The needle forms with the two plates of a quadrant a condenser, and the motion takes place in such a direction as to make as small as possible the energy of the condensers it makes with the four quadrants. It may be proved by methods of the infinitesimal calculus that the angle through which the needle turns varies directly as the difference of potential of its two sets of quadrants. Thus, two differences of potential may be compared by measuring the corresponding deflections of the needle.

In order, however, to *measure* any one difference of potential, a different instrument must be used. This is the "absolute electrometer," which was also invented by Lord Kelvin. As shown in the cut, it consists of a parallel plate condenser, with a disc cut out of the upper plate as described on page 653. In practice this is suspended from one arm of a balance. The two plates of the condenser are joined to the two conductors whose difference of potential is desired; the plates are thus charged with opposite kinds of electricity, and the force of attraction on the movable disc may be measured by putting weights in the balance pan. If the area of this disc is A, the distance apart of the plates, d, the difference of potential, $V_2 - V_1$, the dielectric con-

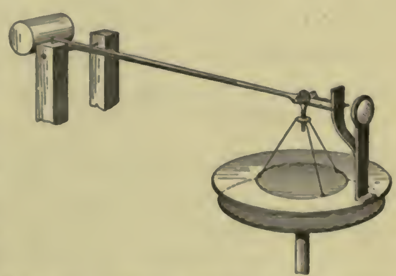


FIG. 356. — Thomson's original form of absolute "trap-door" electrometer.

stant, K, the force of attraction on the disc is given by the formula

$$F = \frac{(V_2 - V_1)^2 AK}{8\pi d^2}.$$

If the plates are, as usual, in air, $K = 1$, and $(V_2 - V_1)^2 = \frac{8\pi d^2 F}{A}$, F, d, and A can all be measured; and so $V_2 - V_1$ is known.

Measurement of Electric Quantities. — The four electrical quantities that have to be measured are quantity, potential, capacity, and dielectric constant. We have just shown how differences in potential may be measured; and, if the potential of a conductor is to be measured, it may be joined to one plate of an electrometer, while the other plate is connected with the earth.

The capacity of a sphere or of a simple form of condenser may be calculated from a knowledge of its dimensions, as explained on page 652. But there is a simple method, due to Cavendish, for determining when the capacity of two condensers is the same; and so, if the capacity of any condenser is desired, it may be compared by this method with a condenser whose capacity may be varied at will but is known, *e.g.* a parallel plate condenser the distance apart of whose plates can be varied.

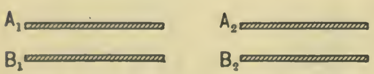


FIG. 857. — Cavendish's method of comparing the capacities of two condensers.

The method is as follows: let A_1, B_1 and A_2, B_2 be the two condensers; charge them by joining A_1 and A_2 to some electrical machine, while B_1 and B_2 are joined to the earth; then disconnect B_2 from the earth, and A_1 and A_2 from the machine and from each other; join B_1 to A_2 by a conductor, and A_1 to the earth (or to B_2). If the capacities are equal, an electroscope in contact with the wire joining B_1 to A_2 will show no effect when A_1 is earthed; for let C_1 and C_2 be the two capacities, and let V be the potential given A_1 and A_2 by the machine; the charge on A_1 is then $+C_1V$, on B_1 is $-C_1V$, on A_2 is $+C_2V$, and on B is

$-C_2V$. When B_1 and A_2 are joined, the two charges, $-C_1V$ and $+C_2V$, do not combine until A_1 is joined to the earth. Then they do, and the final charge, which is distributed over B_1, A_2, and the wire joining them, is $V(C_2-C_1)$; and this will affect an electroscope unless $C_1=C_2$.

This method also permits one to measure K for any dielectric, and was so used by Cavendish. The capacity of the second condenser may be measured when air is the dielectric and again when glass, or sulphur, etc., is substituted. The ratio of the latter capacity to the former is the value of K.

In order to measure a charge, the accepted method is to place it on a condenser whose capacity is known and to measure the resulting potential. Then, since $e = C(V_2 - V_1)$, the value of e is known. There may be a difficulty in making the charge pass to the condenser, but the method described on page 639 may always be used. This is to put the charge inside a conducting vessel which is nearly closed; an equal charge will appear on the outside and this may be measured.

An ingenious method was devised by Lord Kelvin for the measurement of the potential at any point in the atmosphere. Let A be the point of a conductor which is joined to an electrometer, and let some means be adopted to have a continuous current of small conducting particles leave it. Let B be such a particle. Then if the potential of A is higher than that of points in the air near it, a plus charge will be induced on B and a minus one on A; B will carry this charge off as it moves away; and the process is repeated as the stream of particles is maintained. Finally, the potential of A will be lowered by this accumulation of negative charges until it is the same as that of the surrounding air. Similarly, if the potential of A is lower than that of the air near by, it will be raised until it is the same. Therefore, when the potential of A ceases to change, it gives the potential of the air at that point, and may be measured by the electrometer. One means of causing a pointed conductor to give off particles is to use a small flame, because a burning gas is a good conductor. Another method is to have as the conductor A a vessel of water ending below in a small funnel, so that drops of water are continually forming and breaking away. In this manner many interesting facts in regard to atmospheric electricity have been learned; one at least should be noted ; the potential of the lower layers of air is as a rule always higher than that of the earth, and its value is continually changing.

Strains Due to Electrification. — The fact that the main phenomenon of electrification consists in a strain of the dielectric is shown, as has been said before, by the formation of sparks, and in many other ways also. One of the most direct proofs is furnished by what is known as the "residual charges" of a glass condenser. If one is charged to a high potential and then discharged, a second discharge may be obtained after the lapse of a short time; then a third may be obtained, etc., each one being feebler than the preceding one. These are said to be due to residual charges. They depend upon the fact that glass is non-homogeneous; for they cannot be obtained with a homogeneous dielectric. Their explanation is as follows: When the condenser is first charged, the glass is mechanically strained, and when it is discharged, certain parts of the glass lose their strain and, owing to inertia, are strained again in the opposite manner, while other portions of the glass do not relax completely; these two portions, however, balance each other for the moment, and there is no resultant strain; as time goes on, however, these strains, not being maintained by any force, gradually relax, but not to the same degree, so there is again a resultant strain; this causes the second discharge when the armatures are joined, etc.

If it is remembered that there is no field of force inside a conductor, so that such a body cannot maintain a strain, all the phenomena of induction, etc., may be at once explained.

ELECTRODYNAMICS

CHAPTER XLIV

PRODUCTION OF ELECTRIC CURRENTS

Definition of Terms. — The simplest case of an electric current is furnished by the steady discharge of a condenser. (See page 647.) In this, two plates having a difference of potential are joined by a conducting wire; and, as a result of the change, the charges of the two plates disappear. It is noticed further that the temperature of the wire is raised, and certain magnetic effects are produced in the region around the wire. All these phenomena constitute the electric current. We speak of the *current* as being *in the wire;* but this is only a mode of speaking.

As the discharge begins, the plus charge on the plate of higher potential decreases, and so does the minus charge on the plate of lower potential; if by some means these charges may be maintained constant by adding continually the necessary quantities of plus and minus charges, the potentials of the plates will remain unchanged; and the current is said to be "steady." The phenomenon in the conducting wire, which constitutes the current, consists, as will be shown in the next chapter, of a motion of a stream of positively charged particles in the direction from high to low potential in the wire, and of a stream of negatively charged particles in the opposite direction. By *definition* the former direction is called that of the current. If i_1 is the quantity of plus electricity that passes through the cross section of the wire at any point in a unit of time, and i_2 is the quantity of minus electricity that passes at the same time in the opposite direction, the quantity $i_1 + i_2$ is called the "strength of the current" or, more often, "the current." If the current is steady, the quantities i_1t and i_2t pass in an interval of time t; and $(i_1 + i_2)t$ is called the "quantity of the current." If the current is not steady, and if in any interval of time the quantities of plus and minus charges that pass are e_1 and e_2, the quantity of current is $(e_1 + e_2)$. (Thus, in the discharge of the condenser whose plates are charged with $+e$ and $-e$, the quantity of the current is e, because the plates will be discharged if $e_1 + e_2 = e$. If $+e$ passes from one plate to the other; or if $-e$ passes in the reverse way; or if $+\frac{1}{3}e$ and $-\frac{2}{3}e$ pass in opposite ways; etc., the plates are discharged.)

In order, then, to produce a current in a conducting wire it is necessary to have a difference in potential between any two of its points. This difference is called the "electromotive force" (E.M.F.) between the two points.

Work done by a Current. — The passage of a current evidently involves the idea of work. If $V_2 - V_1$ is the difference of potential between two points in a wire, and if $(i_1 + i_2)$ is the current strength, the quantity of positive electricity i_1 moves from a point of high potential, V_2, to one of low, V_1, and therefore the *electric forces do the amount of work* $i_1(V_2 - V_1)$; and similarly, owing to the motion of a quantity of negative electricity in the opposite direction, the same forces do an amount of work $i_2(V_2 - V_1)$. So the total amount of work done in a unit of time by the electric forces is $(i_1 + i_2)(V_2 - V_1)$; or, calling the current strength i and the difference in potential, E, it is iE; and the work done in an interval of time, t, if the current is steady, is iEt. Or, in general, if e is the quantity of current, the work is eE. Ordinarily this work is spent in raising the temperature of the conductor which carries the current; and the necessary amount of energy is furnished by whatever produces the current. (The heat produced in the conductor may be measured if it is in the form of a wire by coiling it in a calorimeter of water. See chapter XII. If the C.G.S. system of units is used in defining the unit quantity of electricity, the product *iEt* is a certain number of *ergs;* and so the heat produced must be expressed in ergs.)

Heating Effect of a Current. — This heating effect of a current is, of course, greatest where there is the greatest amount of work done; that is, where the electromotive force, or drop in potential, is the greatest. This is illustrated in various forms of electric lights, in the electric furnace, in electric heaters, etc.

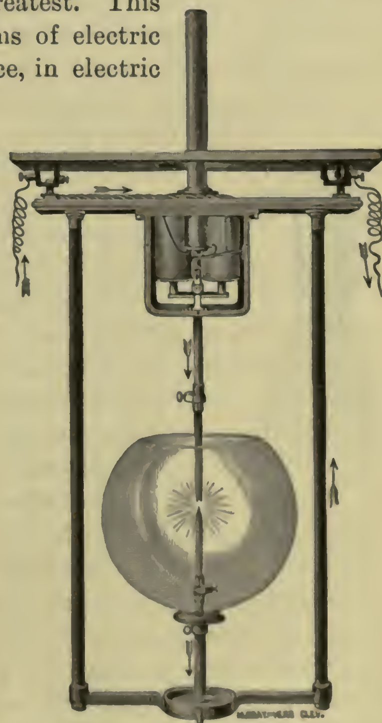


FIG. 858. — The electric arc between two carbon poles.

The arc light, as used for illuminating purposes, consists of two carbon rods which are connected to some source of a current, and which are so controlled by automatic mechanism that when no current is flowing they are in contact, and then as soon as the current begins they are slightly separated. The two rods when loosely in contact offer great opposition to the current, so the temperature rises at the points of contact; this makes the surrounding gas a conductor, and now the rods are drawn apart. The current passes off one rod to the gas, and from this to the other rod. There is great resistance to the current passing off or on a solid; and the temperature of the tips of the rod is raised to a "white heat," if the current is sufficient. This produces the light. Experiments show that more heat is produced at the end of the rod from which the current proceeds to the gas than at the other; this is called the "positive pole."

FIG. 359.—Incandescent lamp.

In the ordinary "incandescent light" there is a glass bulb into which enter two platinum wires connected inside by a fine filament of carbonized wood fibre, and from which the gas has been exhausted as completely as possible. A current is made to flow through the filament, and its temperature is raised to white heat. It does not burn up, because there is no oxygen left inside the bulb.

In the Nernst lamp there is a small filament whose constitution is a commercial secret, which ends in two metal wires; this filament is not a conductor unless its temperature is high, and even then under the action of a current in one direction it decomposes and breaks down. Therefore the process of using the lamp is first to raise the temperature of the filament until it becomes conducting, and then to have it traversed by a current whose direction is reversed at short intervals. If this is done, the filament gives out a brilliant light; and, as it does not oxidize, it may be used in the open air.

In an electric furnace use is made of the high temperature of the arc; and the carbon rods are inclosed in a space whose walls are non-conductors for heat, and in which the pressure of the gas may be increased.

Direction of Current. — In order to determine by experiment the direction of a current it is necessary to ascertain which of the two conductors between which the current flows has the higher potential. The simplest mode of doing this is one invented by Volta. The two conductors whose potentials are V_2 and V_1 are joined by wires to the two plates of a condenser, A_2 and A_1. If $V_2 > V_1$, the plate A_2 becomes charged positively, and A_1, negatively; because lines of force pass across from A_2 to A_1. These charges are on the two faces nearest each other; but if the wire leading to A_1 is broken, and the plate A_2 is then removed, the negative charge on A_1 will spread over the whole plate and may be detected and studied. In Volta's arrangement the plate A_1 was the top plate of a gold-leaf electroscope, and A_2 was a similar plate coated with a thin layer of shellac and carried by a glass handle. Therefore in this apparatus, after the wire leading to A_1 is broken and A_2 is then removed, the negative charge will spread over the plate and the gold leaves, which will then diverge. If now a glass rod which has been rubbed with silk is brought near the electroscope, it will induce a positive charge on the leaves, which will in part neutralize their negative charges, and so they will collapse.

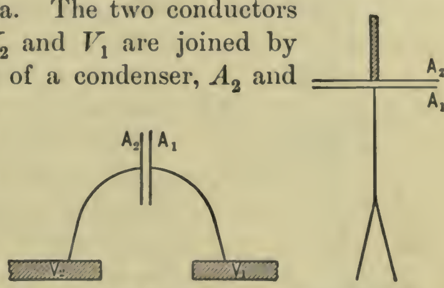


FIG. 360. — Method of determining direction of an electric current. Volta's condensing electroscope.

If, on the other hand, $V_2 < V_1$, the gold leaves will become charged positively, and a charged glass rod will cause them to diverge still farther. In this manner, then, it may be determined whether $V_2 > V_1$, or $V_2 < V_1$; if the former is the case, and if a wire is made to join the two conductors, the direction of the current is from the one at potential V_2 to the one at potential V_1; in the contrary case, the direction of the current is opposite to this.

Detection of a Current. — When an electric current is flowing in a conductor, its temperature rises, as explained above, owing to the work done by the electrical forces against the molecular forces of the conductor. But this fact does not lead to a simple direct means of observing a current, because with a feeble current the change in temperature is small. The magnetic action of a current offers, however, an extremely simple and direct method of detecting and even measuring a current. It was discovered by Oersted, a Danish physicist, in 1819–1820, that a wire carrying a current had a magnetic field around it. We shall take up this question more in detail in a later chapter; but one or two facts may be stated here.

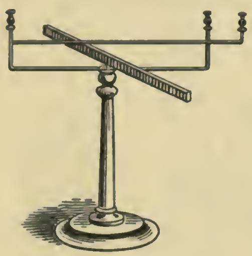


Fig. 361. — Apparatus of Oersted for studying the action of an electric current upon a magnet.

If a magnetic needle is pivoted so as to be free to turn about a vertical axis, it will assume a north-and-south position, and now if a conductor carrying a current is placed parallel to it, but above it, the needle is deflected; if the current is reversed, so is the deflection. Similarly, if the current is parallel to the needle, but below it, it is deflected; but the direction of the deflection is opposite

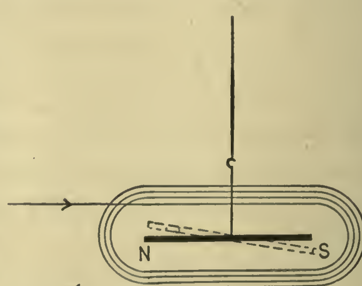


Fig. 362. — Section of a simple galvanoscope.

to what it would be if the current were above the needle. Hence, it follows that if the conductor carrying the current is made in a loop lying in the magnetic meridian and inclosing the magnet, the deflection will be increased; and if many loops are used, forming a flat coil, the deflection will be still greater. This constitutes a "galvanoscope."

FIG. 863. — An astatic combination of magnets.

Another mode of increasing the deflection still more, and at the same time of avoiding, to a large extent, any disturbances of the magnet due to other actions than those of the current in the coils, is to attach rigidly to it another magnet of equal magnetic moment, but turned so that its axis is in an opposite direction. Thus, a north pole of one comes opposite to the south pole of the other. If, now, one of these magnets is inclosed in the coil and the other is either above or below it, the deflective force of the current is in the same direction on both magnets; but the action of any other magnetic field is almost entirely prevented. Such a combination of magnets as this is called an "astatic needle," because if their magnetic moments were exactly equal, and if their magnetic axes were exactly parallel, the system would not be under a directive force due to the earth, and would remain stationary in any position. Actually these conditions are not satisfied; and the earth has an action, but it is extremely small. So, by bringing another magnet near the astatic needle, it may be made to take any position that

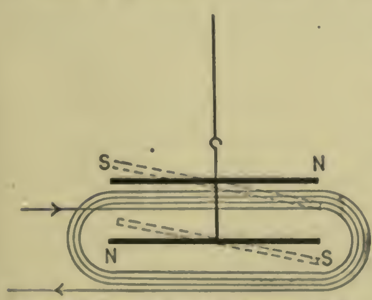


FIG. 864. — Section of a galvanoscope with astatic needle.

is desired, and the action of the earth may be neutralized as completely as is desired. By thus using a "control magnet," then, the coils to carry the current may be kept in any position which is convenient, and the astatic needle may be made to lie in their plane, while the field of force due to the earth and the control magnet may be very small. This last is shown by the period of the magnet becoming very long when it is set in vibration, for $T = 2\pi\sqrt{\frac{I}{MR}}$ (see page 611); and so, if R is small, T is large.

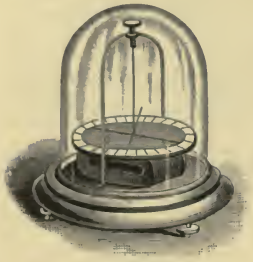


FIG. 865. — Galvanoscope.

The field of magnetic force near any current may be studied by the use of iron filings or of a small magnetic needle, as was described on page 603. It is found that the lines of magnetic force form closed curves around the current; the directions of the current and the lines of force being connected by the right-handed-screw law. Thus, if

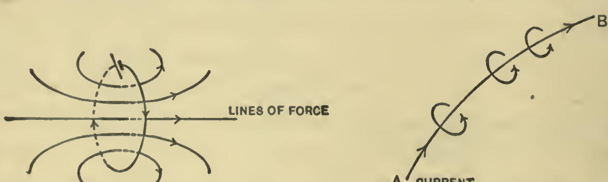


FIG. 866. — Diagram illustrating connection between the direction of a current and that of the lines of magnetic force.

AB is any portion of a conductor carrying a current from A to B, the lines of magnetic force near it are as shown; or, if the total electric circuit is considered, the lines of force pass through it from one side, and return outside. Thus, a current and any one of its lines of magnetic force form two closed links threading each other, like two links of a chain. If a current, then, is passed through the coil of a galvanoscope, the magnetic needle will be deflected; and if the current is reversed in direction, so is the deflection of the magnet. If the current in the coil is in the direction of the motion of the hands of a watch as one looks at the coil from one end, the magnetic force is directed away from the observer, so that a north pole is forced away from, and a south pole is forced toward the observer. By means of such an instrument one can determine, then, the direction of a current, and can roughly estimate its strength.

Tangent Galvanometer. — If the coil of the galvanoscope is a circular one, that is, if the cylinder on which the wire is wound has a circular cross section, the intensity of the magnetic field at the centre of the coil may be proved (see page 711) to vary directly as the current strength and inversely as the radius of the cylinder referred to. Thus if i is the current strength and a is the radius, the intensity of the magnetic force is proportional to $\frac{i}{a}$; it also varies directly with the number of turns of wire in the coil; if this is n and if the turns of wire are so close together as practically to coincide, the intensity may be written $f = c\frac{ni}{a}$, where c is a factor of proportionality. The numerical value of c depends, of course, on the units chosen for the magnetic and electric charges.

If the coil is placed in the magnetic meridian and a current is passed around it, the magnet (*not an astatic one*) suspended at its centre is under the action of two opposing couples, one due to the magnetic field of the earth, and the other to that of the current in the coil; and it comes to rest when these balance each other. If M is the magnetic moment of the magnet, H the hori-

FIG. 867. — Section of tangent galvanometer.

zontal component of the intensity of the earth's field, and N the angle that the magnet makes with the magnetic meridian when it comes to rest, the moment due to the earth's force is $HM \sin N$; and that due to the electric current is $fM \cos N$. Since these must balance each other,

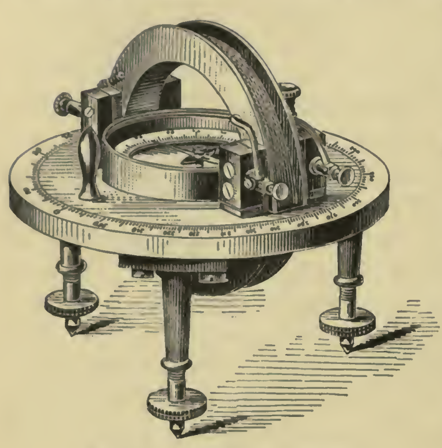


Fig. 368. — Tangent galvanometer.

$$HM \sin N = fM \cos N,$$

or $$\tan N = \frac{f}{H} = \frac{cni}{Ha}.$$

Therefore the strength of the current is measured by the tangent of the angle of deflection; that is, the strengths of two currents vary directly as the tangents of the angles of deflection. Such an instrument as this is called a "tangent galvanometer."

Electro-magnetic Unit Current. — The current strength is defined in terms of the quantities of charge which pass a cross section; but actually these quantities cannot easily be measured directly. So, it is more convenient to define a "unit current" in terms of its magnetic properties, and then from this to deduce the value of a new unit quantity. Thus, a "unit current," or one of unit strength, is defined to be such a current that if flowing in a galvanometer coil of one turn whose radius is 1 cm., the intensity of the magnetic field at its centre equals 2π dynes. (The reasons for this choice of unit current will appear later.) Thus, using the same symbols as in the above formula, $f = c\frac{ni}{a}$, this may be expressed by saying that, when $i = 1$, $n = 1$, and $a = 1$,

$f = 2\pi$; so this definition of a unit current is equivalent to putting $c = 2\pi$ in the formula. This unit current is called the "C. G. S. electro-magnetic" unit, because its definition depends upon the intensity of a magnetic field; that is, upon the force acting upon a unit magnetic charge. Then in a tangent galvanometer the formula becomes

$$\tan N = \frac{2\pi n i}{Ha}; \text{ or, } i = H\frac{a}{2\pi n}\tan N.$$

The quantity $\frac{2\pi n}{a}$ is called the "galvanometer constant"; writing it G, $i = \frac{H}{G}\tan N$. H and G may be measured, and N observed; so the strength of a current may be *measured*.

The C. G. S. electro-magnetic unit *quantity* of electricity is, then, the quantity carried past any cross section of the conductor in one second by a unit electro-magnetic current. There must, of course, be a constant relation between this quantity and the C. G. S. electrostatic unit quantity as defined on page 640; and experiments prove that one C. G. S. electro-magnetic unit charge equals 3×10^{10} C. G. S. electrostatic units of charge. (It should be noted that this ratio of the units equals the velocity of light.)

The work required to carry a unit electro-magnetic quantity of charge from one point to another is the difference of potential between them expressed on the C. G. S. electro-magnetic system. Experiments show that the numbers which would be required to express ordinary potential differences on this system are so large that they are inconvenient. Consequently other units are used in practice. Thus a potential difference on the C. G. S. electro-magnetic system of 100,000,000, or 10^8, is called one "volt."

Measurement of Quantity of Current. — It is often necessary to measure the total quantity of current which flows in a short time, *e.g.* when a condenser is discharged. In this case the current is not constant, and further the time of flow is so short that it is not possible to secure a permanent deflection of the galvanometer needle. If such a current is passed through a galvanometer, the needle will be acted on by an *impulse* and will have a certain "fling"; that is, it will be deflected from its north and south position and will proceed to make a number of oscillations before coming to rest again in its former position. If the period of vibration of the needle is so long and the time of passage of the current so short that we may consider the current as over before the needle is deflected an appreciable amount, the maximum angle of deflection measures the quantity of current. The exact formula may be deduced without difficulty. If e is the quantity of the current on the C. G. S. electro-magnetic system; N, the angle of fling from a north and south direction; G, the galvanometer constant; H, the horizontal component of the earth's magnetic force; T, the period of vibration of the magnet,

$$e = \frac{HT}{G\pi} \sin \frac{N}{2}.$$

An instrument specially designed to measure quantities of current, as distinguished from current strengths, is called a "ballistic" galvanometer.

Measurement of Electro-motive Force. — Since an electromotive force is a difference of potential, it may be measured by any electrometer. (See page 658.) But, in general, other methods are adopted. One is to join the two points which are at different potentials to a condenser of known capacity, and then to discharge it through a ballistic galvanometer. If E is the difference of potential and C the capacity of the condenser, the quantity of current measured will be CE. (The value of C on the electro-magnetic system must be used, if E is to be *measured;* but, if two electromotive forces are to be compared, it is not necessary to know the value of C.)

Another method of comparing differences of potentials depends upon the fact, which will be discussed more fully later, that in the case of a steady current its strength is directly proportional to the E. M. F. producing it; but, if the E. M. F. is applied at the ends of a long, fine wire, the current is small, while if the wire is short or thick, the current is large. In the former case there is said to be a great "resistance"; in the latter, a small one. Thus, if a steady current is flowing through the conductor *PQ*, and the value of the difference of potential between two points *A* and *B* is desired, these points may be connected by wires to a galvanometer, *G*, through coils of wire, *R*, which are so long and so fine that they offer such an opposition to the passage of a current that practically none flows from *A* around through *G* to *B*, and so no difference is made in the conditions at *A* and *B*. If, however, the galvanometer is sufficiently sensitive, it will measure this minute current; and its strength is directly proportional to the difference of potential between *A* and *B*. Other methods may be found described in laboratory manuals.

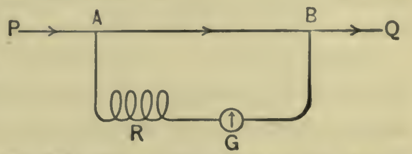


FIG. 369. — Diagram illustrating a method of measuring difference of potential.

Steady Current. — If by any means, mechanical, chemical, thermal, etc., it is possible to maintain a *constant difference* of potential between two conductors, a steady current may be produced by connecting these two conductors by a wire or other conductor. There are at least four methods by which this constant difference in potential may be produced. If an electrical machine such as described on page 635 is turned at a uniform rate, it may be used to furnish a steady current. If two rods of different metals, such as zinc and copper, are partly immersed in some liquid conductor other than a fused metal, such as a solution of sulphuric acid in water, it is found that the rods are at different potentials. If a closed metallic circuit is made by joining several wires of different material in series, and if the junctions of the different wires are at different temperatures, a current is produced in the circuit. Again, if a closed circuit of some wire is moved about in a magnetic field in such a manner that the field of force through the circuit varies, a current arises; and, if this change in the field continues at a uniform rate, the current is steady: this constitutes a "dynamo."

Primary Cells. — Experiments show that, when a solid conductor is immersed, partly or completely, in a liquid conductor other than a fused metal, there is a difference of potential between them, which is characteristic of the two conductors. So, if two solid conductors dip in the same liquid, they will be at different potentials; and, if they are joined outside the liquid by a wire, a current will flow in it. This fact was first observed by Volta (1800), who used zinc, copper, and dilute sulphuric acid in this manner. This is said to be a "Voltaic cell."

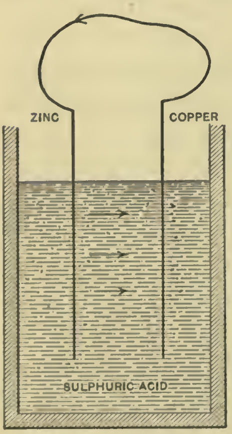


Fig. 370. — Voltaic cell.

It is a question of experiment to determine which of the solid conductors has the higher potential. In the case of the voltaic cell, the copper rod is at a higher potential than the acid, and the acid is higher than the zinc; so the current in the connecting wire outside is from the copper to the zinc. It is observed that, as the current continues to flow, the zinc gradually dissolves away and bubbles of hydrogen gas collect on the copper rod or break loose from it and rise to the surface. It is observed, further, that there is a current also through the dilute acid, and that its direction is from the zinc to the copper. Thus the current flows in a *circuit;*

outside the liquid, from copper to zinc; inside the liquid, from zinc to copper. Since the direction of a current is always from a point of high potential to one of low, it is thus evident that at the boundary separating the zinc and the acid there must be some mechanism which raises the potential; so that the points on the zinc must have the lowest potential in the whole circuit, and contiguous points in the dilute acid must have the highest potential. This phenomenon is evidently connected closely with the dissolving of the zinc in the acid.

If pieces of zinc are placed in dilute sulphuric acid in a tumbler or beaker, it is noted that the zinc dissolves, that hydrogen gas is evolved, and that the temperature of the acid is raised. This proves that, when zinc dissolves, energy is liberated; in the simple chemical experiment this energy is spent in producing heat effects; in the voltaic cell it is spent in raising the potential of points in the acid, and this maintains the current and so heats the conductors, etc.

At the surface of the copper, where the current enters it from the acid, work is required to raise the potential of the plus charges from that of the acid to that of the copper, and to lower that of the negative charges which are going in the opposite direction. This difference of potential at the surface is due to the evolution there of the hydrogen gas.

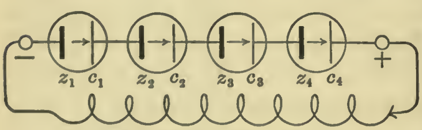
The mechanism of the current through the acid and at the zinc and copper rods will be discussed in the next chapter.

The two solid conductors which dip in the liquid are called "poles"; the one which is at the higher potential is called the positive one, while the other is called the negative one. The latter is always dissolving as the current flows; so if it contains any metallic impurities, *e.g.* if the zinc has particles of iron in it, there will be *local* currents from the zinc to the acid, then to the iron, and thence to the zinc, etc. These currents have no external action; and so should be prevented, if possible, because the zinc consumed in producing them is wasted. This can be done in many cases by rubbing mercury over the zinc rod before it is immersed in the liquid, and thus making a surface of mercury amalgam with the metal, which is practically uniform.

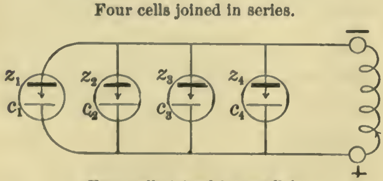
As the current flows in a voltaic cell, hydrogen bubbles collect over the copper pole, and thus hinder the action of the cell. Various devices have been invented in order to prevent this. The most successful is due to Daniell. He made a cell, which bears his name, consisting of a porous cup — such as unglazed porcelain — inside a larger vessel; the cup contains a saturated aqueous solution of copper sulphate, and the outer vessel, dilute sulphuric acid; the zinc rod dips in the latter, the copper rod in the former. When the two rods are joined outside by a wire, the current flows from the copper to the zinc. As it flows, the zinc dissolves as before, but now copper is deposited out of the copper sulphate solution on the copper rod. Consequently there is no change in the nature of the surface of the latter. This cell of Daniell is a typical "two-fluid cell."

Other cells, both one and two fluid, can be made by using other metals than zinc and copper, and other liquids than sulphuric acid. They are called "primary cells" in distinction to "secondary" ones, which will be described presently.

Cells may be joined "in parallel" or "in series." Thus if C_1 and Z_1 are the positive and negative poles of one cell, C_2 and Z_2 those of the second, etc., the cells are said to be in series if C_1 and Z_2, C_2 and Z_3, etc., are connected by wires; while if C_1, C_2, C_3, etc., and Z_1, Z_2, Z_3, etc., are connected, the cells are said



Four cells joined in series.



Four cells joined in parallel.

Fig. 871.

to be in parallel. If the cells are all of one kind, let E be the difference of potential between the two poles of each; then if n cells are joined in series, the difference in potential between C' and Z_n is nE. Whereas, if they are joined in parallel, the only effect is to make what is practically one cell with poles n times as large; this does not affect the difference of potential between the poles.

A mechanical analogy of a simple voltaic cell is furnished by a pump or paddle wheel working in a horizontal tube connecting two tall vertical pipes containing some liquid, such as water. If the pump is open, the liquid will stand at the same level in the two vertical pipes; but, when the pump or wheel is set in action, the liquid will be forced through so as to stand higher on one side. A difference of pressure on the two sides of the pump or wheel is thus produced; and, if sufficient, it will stop the action of the latter. If now a connecting tube between the upper portions of the pipes is opened, the liquid will flow from the one at the foot of which the pressure is the higher over into the other, and a continuous current will be produced. This pipe in which the pressure against the pump or wheel is the greater corresponds to the copper rod in the voltaic cell; the other pipe to the zinc; and the pump or wheel to the energy furnished by the dissolving zinc.

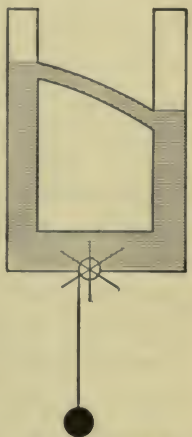


FIG. 872.—Model representing action of voltaic cell.

Thermoelectricity. — If a closed circuit of linear conductors, like wires, includes at least two different substances, there is in general an electric current produced in the circuit if the junctions of these substances are kept at different temperatures. Thus, if two wires I and II make up a circuit having junctions at A and B, there will be, in general, a current if the temperatures of A and B are not the same. The direction and strength of the current depend upon the two substances and upon the difference in temperature. It is found by experiment that, beginning with a condition when A and B are at the same temperature, if that of A is kept unchanged and that of B is continuously *increased*, the current will be in a definite direction and will gradually increase, while if the temperature of B is *decreased*, the current will be in the opposite direction and will gradually increase; as the temperature of B is made to differ more and more from that of A in one direction, — in certain cases when it is higher, in others when it is lower, — there comes a point when the current begins to decrease, and finally one at which the current ceases; while if the difference in temperature is increased still more, a current is produced, but it is in the opposite direction to that which it was before, and as the change in the temperature of B continues, this reverse current increases in strength. If t_A is the temperature of A during the experiment, and if t_I is that of B at which the current ceases, experiments show that their mean $\frac{t_A + t_I}{2}$ is a constant quantity for any two substances: it is known as their "neutral temperature"; and t_I is called the "temperature of inversion," corresponding to t_A.

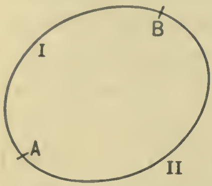


FIG. 373. — A closed circuit made up of two conductors I and II, having junctions at A and B.

In order to explain these thermocurrents, as they are called, it is necessary to assume that at any cross section in the conductors where two different substances come in contact there is a difference of potential. If P and Q are two different substances meeting over a surface, the fundamental experiments of electrification show that, when they are separated, one is charged with plus, the other with minus electricity. This proves that when they are in contact there is some electric force — due to the difference in the electric properties of the molecules — acting at the surface of con-

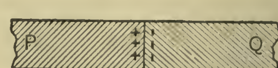


FIG. 374. — Junction of two conductors P and Q.

tact, and resulting in a separation of the plus from the minus charges. Let us suppose that P is the substance which is charged positively; then the direction of the force producing this charge must be across the surface of contact *from Q to P*. As a result of the plus charge on P and the minus charge on Q, the potential of P is higher than that of Q; so that if P and Q are conductors, and if they are joined by some wire, a current would tend to flow, owing to this fact, from P through the wire to Q. This difference of potential at the surface of contact would be maintained by the molecular forces. Calling this difference of potential E, we may say that there is a "contact electro-motive force" E at the boundary. The proof of the existence of this E. M. F. across the surface of contact is afforded if P and Q are conductors, and if an electric current is forced by some source, such as a voltaic cell, across this surface, first in the direction from P to Q, then in the opposite direction. It is found that in the former case the temperature of the junction rises; in the latter it falls. If the current i flows for an interval of time t from P to Q, the electricity is passing from high to low potential, and so *the external electric forces do* the work itE at the junction; and this energy appears in the form of heat effects. If, however, the current is in the opposite direction, the electricity is having its potential raised at the junction, and so the work itE must *be done on the electric forces* at the expense of the energy of the molecules at the junction; and therefore its temperature falls. (Or, we may say that in the former case work is done *against* the molecular forces which produce the electrical separation; while in the latter, these forces *do work* themselves in helping on the current.) These forces at the surface of contact of two substances are called "Peltier electro-motive forces," having been first discovered by him. They can be measured by putting a junction in a calorimeter of water, and measuring the heat produced, the current, and the time. Direct experiments prove that they vary in amount with the temperature. Thus, in the thermo-couple described above there are two such forces, at A and B; and, if the temperatures at these points are different, these forces are unequal.

But there are other similar forces in each conductor between A and B, if the temperatures of these points are different. For, consider either of these conductors, the two ends of which are at different temperatures; if a section is taken across the wire at any intermediate point, the temperatures on its two sides differ *slightly*, and so the condition of the molecules which are in contact across this section is different on the two sides. Therefore, we might expect an electro-motive force at each point in the conductor. This was proved by Lord Kelvin — then Sir William Thomson — by the following experiment: let an electric current be forced through a wire of some definite material whose ends are kept at a higher temperature than its middle point, and let the temperatures be noted at two intermediate points, one in each half, which are such that their temperatures are the same when no current flows; it is observed that, when the current is flowing, the temperature at one of these points rises, while that at the other falls.

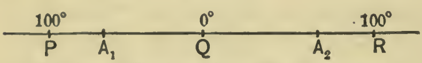


Fig. 875. — Diagram representing Thomson's experiment.

Thus if the wire is $\overline{PQR}$, let the current be from P to R, and the temperature of P and R be higher than that of Q; and let A_1 and A_2 be the two points whose temperatures are the same before the current begins. The molecular forces at A_1 producing the E. M. F. due to the temperature effect just described are either in the direction from Q toward P, *i.e.* from a cold point to a hot one, or from P toward Q. If the former is true, as the current flows from P to Q, the temperature at A_1 rises; while if the latter is true, the temperature at A_1 falls. Similarly, the temperature at A_2 either falls or rises; but, if the temperature at A_1

rises, *i.e.* if the molecular forces are in the direction from a cold point to a hot one, the current at A_1 is in a direction opposite to that of these forces, while at A_2 the forces and the current are in the same direction, and so the temperature at A_2 falls. These forces in a wire which is homogeneous except for differences in temperature are called "Thomson electro-motive forces." In a simple circuit made up of two wires there are then these forces at each point of both.

The electric current produced in a circuit made up of different substances whose junctions are at different temperatures is due to the Thomson and Peltier electro-motive forces. These currents were discovered by Seebeck in 1821, but their explanation was not known for many years. It is evident that, if a sensitive method is known for the detection of an electric current, a means is offered for detecting differences in temperature between two points; for the junctions of a thermocouple may be placed at them. The sensitiveness of the instrument may be increased by joining in series several pairs of the two conductors, as shown in the cut. If the alternate junctions are kept at one temperature, and the other junctions are kept at a different one, the current will be increased; and so a less difference in temperature may be detected. Such an instrument is called a "thermopile." A cut of an actual instrument is shown.

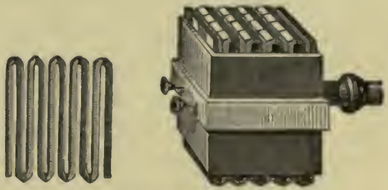


FIG. 876. — A thermopile.

CHAPTER XLV

MECHANISM OF THE CURRENT

Electrolysis. — It is found by experiment that many liquids are conductors, while others are not. A metal in a liquid condition is a conductor, and its properties are exactly like those of the solid conductor. There are, however, certain liquid conductors such that, when a current is made to traverse them, there is an evolution of matter at the points where the current enters and leaves. The liquid must be held in some vessel and two metal rods or wires connected with some source of electric current — such as a series of cells — must dip into it. The conductor, or "electrode," at which the current enters the liquid is called the "anode"; that at which it leaves, the "cathode." Thus the potential of the former is higher than that of the latter; and the direction of the current in the liquid is from the anode to the cathode.

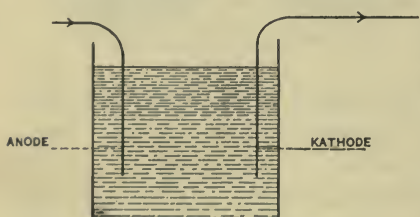


FIG. 377. — An electrolytic cell.

The matter that is liberated at the anode and cathode may bubble off in the form of a gas, it may combine chemically with the metal rods themselves, or it may simply form a solid deposit on them. Liquid conductors which have this property are called "electrolytes"; and the process of conduction in them is called "electrolysis." A careful study of the nature of electrolytes has shown that in every case they are *solutions*, *e.g.* common salt or sulphuric acid in water; and that the solutions are of the kind which exhibit an abnormal osmotic pressure, an abnormal depression of the freezing point, and an abnormal elevation of the boiling point. (See pages 261 and 276.)

Faraday's Laws. — A careful study has been made of the character of the substances which are evolved at the anode and the cathode in different electrolytes. It is found that hydrogen and all metals are liberated at the cathode, while oxygen, chlorine, iodine, etc., are liberated at the anode. Further, the amounts of the substances evolved under different conditions were systematically studied by Faraday. As a result of his investigation, he was able to describe all of his observations in two simple laws which bear his name. Before stating these, however, it is necessary to define a chemical term which was used much more commonly in former days than now, and yet which is convenient. Experiments have established the fact that a molecule of any chemical compound consists of a certain number of smaller parts, called atoms, the atoms of any element being alike in all respects. Thus, a molecule of steam consists of two atoms of hydrogen and one of oxygen, so its symbol is H_2O; a molecule of sulphuric acid may be expressed by the symbol H_2SO_4; one of copper sulphate by $CuSO_4$; one of hydrochloric acid by HCl; etc. A molecule of hydrogen gas has the symbol H_2; one of oxygen gas, O_2; etc. The "molecular weight" of any definite compound has been already defined to be a number which is proportional to the weight of one of its molecules; and a method has been described for the determination of this quantity in certain cases. Other methods are also known. Similarly, the "atomic weight" of any element is a number proportional to the weight of one of its atoms. Thus, the molecular weight of oxygen is 32; so its atomic weight is 16; etc. It is seen from the above illustrations of the composition of molecules that in some cases one atom, in others two, of hydrogen are contained in a molecule. Thus, in hydrogen gas and hydrochloric acid, one atom of hydrogen combines with an atom of hydrogen or one of chlorine respectively; in steam and in sulphuric acid, two atoms of hydrogen combine with one of oxygen or with the "radical" (SO_4); etc. The number of hydrogen atoms which is required to form a stable molecule with the atom of a substance or with a certain "radical" (or group of atoms), is called the "valence" of that substance or of that radical. Thus, the valence of hydrogen and of chlorine is one; that of oxygen and of SO_4 is two; etc. Experiments show that if any molecule is regarded as made up of two parts, the valences of the two parts are the same. Thus, since a molecule of copper sulphate is $CuSO_4$, the valence of copper is two, as is shown also by the fact that the saturated oxide of copper is CuO. (An atom may have a different valence in different compounds; but only one of these is in general a stable molecule.) The ratio of the atomic weight of an element, or of the sum of the atomic weights of the atoms in a radical, to its valence is called its "chemical equivalent."

We can now state Faraday's two laws:

1. The quantity or mass of a substance liberated from any electrolyte at either the anode or the cathode is directly proportional to the *quantity* of the current that passes.

2. The masses of different substances liberated at the anode and cathode in any electrolyte by the passage of the same quantity of current are directly proportional to their chemical equivalents.

If the current flows through several electrolytes arranged in series, let m_1 and m_1' be the masses of the substances liberated at the anode and the cathode in the first electrolyte, and c_1 and c_1' their chemical equivalents; m_2, m_2', c_2, and c_2' be similar quantities for the second electrolyte, etc. Faraday's first law states that any m varies directly as the quantity of current; so, if the current is steady, and if i is its strength, m is directly proportional to the product of i by t, the interval of time taken to liberate the mass m. The second law states that $m_1 : m_1' : m_2 : m_2' :$ etc. $= c_1 : c_1' : c_2 : c_2' :$ etc.

Voltameters. — The first law offers a convenient method for the comparison of the strength of two different currents. An electrolyte is placed in series with the two currents in turn, and the quantities of matter liberated in definite intervals of time at either anode or cathode are measured. If the strength of one current is called i_1, and if the mass it liberates in an interval of time t_1 is m_1, m_1 is proportional to $t_1 i_1$. So, if the strength of the other current is called i_2, and if it liberates a mass m_2 of the same substance in time t_2, m_2 is proportional to $t_2 i_2$;

or $$m_1 : m_2 = t_1 i_1 : t_2 i_2.$$

Hence $$i_1 : i_2 = \frac{m_1}{t_1} : \frac{m_2}{t_2}.$$

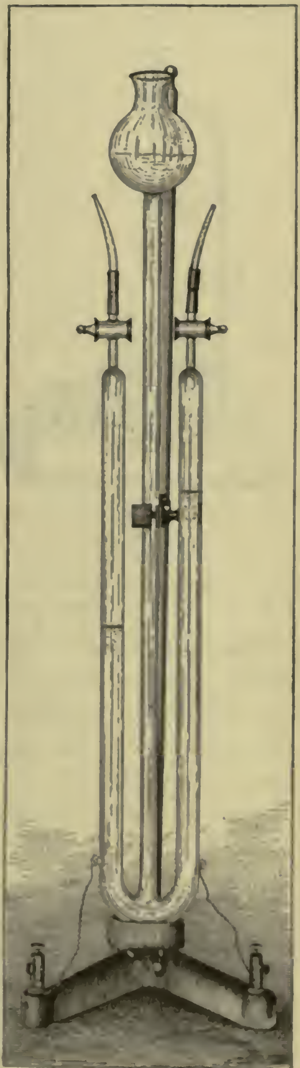


Fig. 878. — A water voltameter.

Special instruments have been designed for this purpose of comparing current strengths; they are called "voltameters." One of the commonest forms is shown in the cut. It is called a "water voltameter"; and in it the electrolyte is dilute sulphuric acid, and the anode and cathode are sheets of platinum inserted at the lower ends of the two arms of a long U-tube. The substances liberated are oxygen gas at the anode, and hydrogen gas at the cathode; these collect in the upper portions of the closed arms of the instrument; and their volumes may be measured. (The middle tube, as shown in the cut, is open to the air at the top, and the quantity of liquid in it may be varied so as to make the pressure on either gas equal to that of the atmosphere, when its volume is measured. From a knowledge of its pressure, volume, and temperature, and of its density as given in tables its mass may be calculated.)

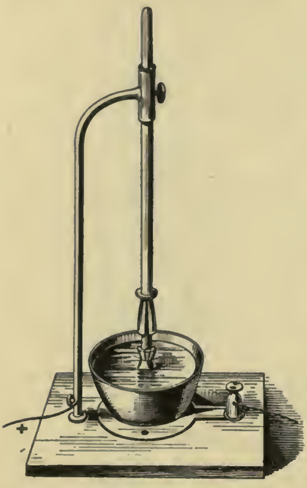


FIG. 379. — A simple form of silver voltameter.

In another form of instrument the electrolyte is a solution of copper sulphate, and the anode and cathode are both sheets of copper; the quantity of copper deposited on the cathode is determined by weighing it before and after the current passes. This is called a "copper voltameter." In the standard instrument the electrolyte is a solution of silver nitrate in water, of a definite concentration; the anode is a plate of silver and the cathode which receives a deposit of silver is the platinum bowl which holds the silver nitrate; this bowl is weighed before and after the current passes, and suitable precautions are taken to prevent any mechanical currents in the liquid, for it is found that they, for purely chemical reasons, affect *slightly* the quantity deposited.

Electro-chemical Equivalent. — By means of Faraday's second law we can calculate the relative amounts of different substances which are liberated by the same current in the same time. The quantity, *i.e.* the mass, of any substance which is set free at either anode or cathode as a unit C. G. S. electro-magnetic quantity of electricity passes is called the "electro-chemical equivalent" of that substance. It is a matter of experiment to determine its value for any one substance; but, being known for this, its value for any other is given at once by the second law. Careful experiments show that the electro-chemical equivalent of any substance apparently varies with the kind of voltameter used; but the variations are undoubtedly due to secondary reactions or causes. If the voltameter is so constructed as to avoid these, the electro-chemical equivalent of silver is found to be 0.011175 g. The chemical equivalent of silver is 107.93; so calling the electro-chemical equivalent of any other substance *m* and its chemical equivalent *c*, we have the relation

$$0.011175 : m = 107.93 : c,$$

or

$$m = \frac{0.011175 \times c}{107.93} = \frac{c}{9658}.$$

Thus, for hydrogen, $m = 0.00010354$; for copper, $m = 0.0032929$; for zinc, $m = 0.0033857$; etc.

Since *m* is the quantity of any substance liberated by a unit electro-magnetic quantity of electricity, the quantity liberated by 9658 such units is a number of grams equal to *c*, the chemical equivalent.

Ions. — The explanation of electrolysis which was advanced by Faraday was that in any electrolyte there are present certain charged particles, some with +, others with − charges, and that these particles are driven by the electric force in one direction or the other. Positively charged particles will move in the direction of the force, that is toward the cathode; while those negatively charged will move in the opposite direction, toward the anode. These charged particles Faraday called "ions"; and those which move toward the cathode he called "cations," while those which move toward the anode he called "anions." Thus all cations are positively charged; all anions, negatively. The electric current in the electrolyte consists, then, from this point of view, of the passage in opposite directions of these two sets of charged bodies. When they reach the electrodes, they give up their charges and in some manner cause the liberation of *uncharged molecules* or ordinary matter. Hydrogen and all metals are liberated at the cathode; therefore we must consider a hydrogen ion or any metallic ion as being positively charged. Similarly, we must consider an oxygen or a chlorine ion as being negatively charged.

Since the electrolyte itself is not electrically charged, any volume of it must contain as much positive electricity on its cations as negative on its anions. The current consists of the passage across any cross section of the electrolyte of these ions; if the current strength is i, and if i_1 is the positive charge carried on the cations, and i_2 is the negative charge carried on the anions, $i = i_1 + i_2$. In the main body of the electrolyte the positively and negatively charged ions balance each other; but in the immediate neighborhood of the cathode, positively charged ions carrying a charge i_1 come up in a unit of time, and negatively charged ions carrying a charge i_2 leave in the same time; so this space gains in this time cations carrying charges $(i_1 + i_2)$ which are not balanced by anions; and these give up their charges to the cathode. Similarly, at the anode, anions carrying a charge i_2 come up, and cations carrying a charge i_1 leave, thus causing a concentration, as it were, of anions carrying a charge $(i_1 + i_2)$ which are not balanced by cations. In other words, the current of strength i enters the electrolyte owing to the fact that anions carrying a charge i are liberated at the anode; and it leaves the electrolyte owing to the fact that cations carrying a charge i are liberated at the cathode. This fact may also be expressed by saying that the *same quantity of electricity* — not regarding its sign — *is carried on that number of ions of any substance, whose mass equals its electro-chemical equivalent.* This may be described differently:

When a *unit* quantity of electricity passes, $i_1 + i_2 = 1$; if m_1 and m_2 are the electro-chemical equivalents of the two sets of ions, m_1 and m_2 grams are liberated at the two electrodes in this interval of time. The liberation of the m_1 grams at the cathode is due, as has been shown, in part to the bringing up of the cations carrying the charge i_1 and in part to the withdrawal in the opposite direction of the anions carrying the charge i_2; so the effect is the same as if there were no anions, but only cations, carrying a unit charge. In other words, a mass of cations equal to m_1 carries a unit electro-magnetic charge; and similarly a mass of anions equal to m_2 carries a unit charge.

Thus, in the case of hydrogen, the electro-chemical equivalent equals $\frac{1}{9658}$; and therefore this number of grams of hydrogen ions carries a unit charge, or one gram of hydrogen ions carries a charge equal to 9658. Similarly, a number of grams of ions of any substance equal to its *chemical equivalent* carries a charge equal to 9658. (The ratio, then, of the charge carried on a hydrogen ion to its mass equals $\frac{1}{9658}$, or, approximately, 1×10^{-4}.)

Faraday showed that both of his experimental laws could be explained if it were assumed that the ions of any one substance were all alike, and that the charge on any ion was proportional to the valence of the substance. For, if these assumptions are true, it is evident that the quantity of electricity carried through the liquid must vary directly as the mass liberated at either electrode; this is the first law. Further, if the same quantity is being carried by two different sets of ions, the masses of these substances liberated, if the charges carried by all the ions are equal, are proportional to their atomic weights; whereas, if each ion of one set carries twice the charge carried by each ion of the other set, — that is, if the valence of the former set is twice that of the latter, — only one half as many of the ions of the former are involved in the current as of the latter, and so the ratio of the mass of the former to that of the latter is equal to that of their chemical equivalents; this is the second law. On this assumption, the charge carried on an ion whose valence is one is the smallest charge involved in electrolysis. It is called an "atom of electricity."

Nature of Ions. — The question as to the nature of the ions is a most important one. As was said above, all electrolytes are solutions which show abnormal osmotic pressures, depressions of the freezing point, etc.; and it is shown in treatises on Physical Chemistry that these various abnormal phenomena can all be explained if it is assumed that in these solutions a certain proportion of the dissolved molecules are dissociated into simpler parts. It is natural to expect that, if a molecule is broken up into parts, they should be electrically charged, so that equal amounts of positive and negative electricity are produced. If this is the case, it is seen at once that the charge on any atom or radical is proportional to its valence. Thus, if a molecule of hydrochloric acid, HCl, breaks up into two parts, H and Cl, and if one is charged positively, the other will have an equal amount of negative electricity; and the valences of hydrogen and chlorine are the same. Similarly, if a molecule of sulphuric acid, H_2SO_4, dissociates into three parts, H, H, and (SO_4), the two hydrogen atoms will be charged alike, and therefore the radical (SO_4) will have an opposite charge equal numerically to twice that on a hydrogen atom; and the valence of (SO_4) is two.

We assume, then, that when an electrolytic solution is made, a certain proportion of the dissolved molecules are dissociated into simpler parts, and that these parts are electrically charged, some positively, some negatively, so that the total charge is zero. (This condition of dissociation is not to be thought of as a static one, but as dynamic; molecules are constantly dissociating, and others are being formed by combinations of the parts, which are moving about in the solution; but at any temperature and concentration a certain definite proportion of the molecules are in a state of dissociation.) These charged fragments of molecules form the ions when two electrodes at different potentials are lowered into the solution; the positively charged ions move toward the cathode during their intervals of existence, before they combine with other ions and form electrically neutral molecules; the negatively charged ones move toward the anode. The fact should be emphasized that the ions are produced in the act of solution, not by any action of the electric current:

the current merely liberates the matter at the electrodes. It should also be emphasized that an ion is an electrically charged atom or radical, and is not a molecule; and that the properties of matter as we observe them, *e.g.* gases, liquids, etc., are the properties of molecules or groups of molecules. Thus, there is no connection between the general properties of a hydrogen ion and those of a hydrogen molecule. Again, as an ion moves through a solution, it is extremely probable that it carries with it a certain number of *molecules*, and that the number associated with a negative ion is not the same as that associated with a positive ion. If this is true, the effective mass of an ion is much greater than its actual mass, considered merely as a fragment of a molecule.

The student should consult Jones, *The Modern Theory of Solution*, New York, 1899, for the original memoirs of Van't Hoff, Arrhenius, and others.

The question as to whether an ion is positive or negative is settled by observing whether it is a cation or an anion. Thus an ion of hydrogen or of any metal is positive; while one of oxygen, or chlorine, etc., is negative. Again, since the same masses of any substance are liberated by the same quantity of electricity, regardless of the nature of the electrolyte, *e.g.* when hydrogen is liberated from dilute sulphuric acid, or nitric acid, or hydrochloric acid, etc., it is proved that an ion of any one substance always has the same charge in an electrolyte, no matter to what molecule it owes its origin. Thus a hydrogen ion in a liquid always has a definite plus charge; etc.

We shall now consider in detail one or two cases of electrolysis. If sulphuric acid, H_2SO_4, is dissolved in water, let us consider the ions as being H, H, (SO_4); where the first two are charged with equal amounts of positive electricity, which we may call $+e$, and the last has a charge $-2e$. Under the action of the electrical force the hydrogen ions move in the direction of the cathode, they combine with SO_4

ions, other molecules dissociate, etc.; but, as the current flows, hydrogen ions continuously come up to the cathode, give up their charges, combine with other hydrogen atoms to form molecules of hydrogen gas which is liberated. The (SO_4) ions in a similar manner migrate toward the anode; but, since a molecule of (SO_4) radicals cannot exist under present conditions of temperature, pressure, etc., when these ions reach the anode and give up their charges, there is a reaction with the molecules of the water near the anode, which takes place according to the following formula:

$$SO_4 + H_2O = H_2SO_4 + O.$$

Consequently for each (SO_4) radical an oxygen atom appears, and these oxygen atoms form molecules of oxygen gas which may bubble off at the surface of the electrolyte or may act chemically upon the anode and oxidize it.

Again, let the electrolyte be a solution of copper sulphate, $CuSO_4$, in water; and let both the electrodes be copper plates. A molecule of copper sulphate dissociates into two ions, Cu and (SO_4); the former is charged positively, the latter negatively. When the copper ions reach the cathode, they form molecules and are deposited on it. When the (SO_4) ions reach the anode, they react upon the copper molecules of the plate in such a manner that the copper goes into solution. Since metal ions are charged positively, the copper dissolves in the form of positive ions; so the current is carried off the anode by them. These + ions serve, then, to balance the $-(SO_4)$ ions which are being continually brought up to the anode.

This obviously offers a method for "copper plating" an object. Its surface must be so prepared that it is a conductor and that copper will adhere to it; and it then must be used as a *cathode* in an electrolytic bath of copper sulphate, the anode being a plate of copper.

Similarly, in an aqueous solution of silver nitrate, $AgNO_3$, the ions are Ag and (NO_3); the former is the cation, the latter the anion. If a plate of silver is the anode, it dissolves, and silver is deposited on the cathode. This offers a method of silver plating.

In the last two illustrations it is seen that the mechanism by which the current enters the electrolyte from the anode consists in the copper or the silver plates dissolving; this is done by the *positively charged ions* of copper or of silver leaving the plates and entering the liquid. In the water voltameter, where the anode and cathode are platinum plates, the case is not quite so simple. The negative ions of oxygen are formed at the anode by the reaction of SO_4 upon the water molecules; and in some manner positive charges pass from the anode to certain of these ions, first neutralizing their negative charges and then giving them positive ones; and, after this takes place, a positive oxygen ion combines with a negative one and forms a molecule of oxygen gas, which bubbles off at the surface. At the cathode, the mechanism is similar. Thus, in the case of copper sulphate, the positive copper ions reach the cathode; under the electric force *negative* charges pass from this upon certain of the copper ions, making them negative; then a negative copper ion combines with a positive one to form a copper *molecule* which is deposited on the cathode.

Polarization. — If two platinum electrodes dip in a solution of sulphuric acid, and if a very small electro-motive force is applied to these electrodes, a current will flow, but will soon cease. This is owing to the fact that under the action of the electrical force the positively charged hydrogen ions collect at the cathode, and the negatively charged anions at the anode; so that, if there is not sufficient force to make the necessary charges pass from the electrodes to the ions and then to form the molecules, these charged particles will lower the potential at the anode and raise it at the cathode until there is no longer any electrical force in the electrolyte, and the current stops. As the applied electro-motive force is gradually increased, a value is reached which will cause the evolution of the gases at the electrodes; and the current will now continue to flow. The same description applies in general to any case of electrolysis; a definite E. M. F. must be applied before electrolysis begins; but its value is different for different electrolytes. This may be calculated, however, from a knowledge of the heats of combination of the compounds which are separated by the electrolysis and of their electro-chemical equivalents.

Thus, experiments prove that, when 18 g. of water are formed by the combination of 2 g. of hydrogen and 16 g. of oxygen, 68,800 calories of heat energy are evolved; and, therefore, when 18 g. of water are broken up into 2 g. of hydrogen and 16 g. of oxygen, an amount of work equal to the mechanical equivalent of 68,800 calories must be done, *i.e.* $68,800 \times 4.2 \times 10^7$ ergs. As a quantity of currents equal to e is passed through the electrolyte consisting of H_2SO_4, the quantity of hydrogen evolved is me, where m is the electro-chemical equivalent of hydrogen; and an "equivalent" amount of oxygen is liberated at the anode. Since it requires $68,800 \times 4.2 \times 10^7$ ergs to liberate 2 g. of hydrogen, that required to liberate me g. is $\frac{me\, 68,800 \times 4.2 \times 10^7}{2}$. If E is the E. M. F. applied to the electrodes, which just causes the electrolysis, the work required to pass a quantity of electricity e between the electrodes is Ee. Therefore, since this work is spent in liberating the hydrogen and oxygen, — neglecting the work done in heating the electrolyte, and assuming that there is no other source of energy, — we have the equation

$$Ee = \frac{me\, 68,800 \times 4.2 \times 10^7}{2},$$

or

$$E = m\, 68,800 \times 2.1 \times 10^7.$$

For hydrogen, $m = 1.035 \times 10^{-4}$, and therefore $E = 688 \times 1.036 \times 2.1 \times 10^5 = 1.5 \times 10^8 = 1.5$ volts. The E. M. F. of a Daniell cell is approximately 1.1 volts; so at least two Daniell cells are required to decompose water.

Calculation of the E. M. F. of a Primary Cell.—We may consider the mechanism of a voltaic or of a Daniell cell from this standpoint of ions. In the former, the cause of the current is the solution of the zinc in the acid, that is, the passing off of positive zinc ions into the liquid. The chemical action is the solution of the zinc in the acid and the evolution of hydrogen at the copper pole; and experiments have proved that when 65.4 g. of zinc are dissolved in dilute sulphuric acid, 38,066 calories are evolved. The electro-chemical equivalent of zinc is 0.00338. So when a quantity of current e passes off the zinc rod into the acid, the mass of zinc dissolved is 0.00338 e; and the energy liberated is

$$\frac{38{,}066 \times 4.2 \times 10^7}{65.4} \times 0.00338\, e.$$

If E is the difference of potential between the copper and the zinc electrodes, it follows that

$$Ee = \frac{38{,}066 \times 4.2 \times 10^7 \times 0.00338}{65.4} e,$$

or $$E = \frac{38{,}066 \times 4.2 \times 10^7 \times 0.00338}{65.4} = 0.83 \times 10^8 = 0.83 \text{ volts.}$$

In this calculation we neglect the loss of energy of heating the liquid, and we assume that the only source of energy is that furnished by the solution of the zinc.

In a similar manner, when a current is produced by a Daniell cell, zinc dissolves at the zinc plate and copper is deposited at the copper electrode. When 63.6 g. of copper dissolve in sulphuric acid, 12,500 calories are evolved; and the electro-chemical equivalent of copper is 0.00329. So the E. M. F. of the cell is found by calculation to be 1.1 volts, making the same assumptions as before.

Storage Cell. — In certain cases of electrolysis the anode and cathode are so modified by the passage of the current, and the consequent liberation of matter at them, that they may be used afterward to form a cell for the production of a current. Therefore, if the battery of cells or the dynamo which was producing the current through the electrolyte is removed, and if the electrodes are joined by a wire, a current will flow through it. This action does not continue indefinitely, for in producing a current, those modifications which were the result of the electrolysis are reversed, and the electrodes return to an inactive condition. If the battery of cells or the dynamo is again used to send a current through the electrolyte, the process may be repeated. Such a combination of electrodes and electrolyte is called a "storage cell" or a "secondary cell"; and when it is in a condition to produce a current itself, it is said to be "charged" — but it is evident that this expression has no connection with what has been called a "charge" in previous chapters.

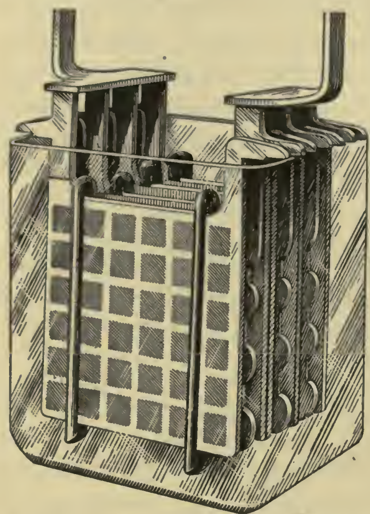


Fig. 880. — The ordinary form of storage cell.

The commonest form of storage cell has dilute sulphuric acid as the electrolyte, and as the anode and cathode two lead grids whose interstices are filled with a paste of lead sulphate. When such a cell is charged, it has an E. M. F. at first of about 2.1 volts; but this falls to about 1.8 volts as the current flows.

Conduction of Electricity through Gases. — In speaking in a previous chapter of the passage of sparks through air or any gas, occasion was taken to state that the discharge consists in the disruption of molecules into simpler parts, and that the path of a spark is an excellent electric conductor. We are thus led to believe that a gas becomes a conductor owing to the presence in it of small parts of molecules, which we may again call ions, although there is no reason for believing that these ions are the same as those in evidence in the electrolysis of a liquid. The charged particles may be the same, but the masses associated with them are different (see page 693). All the known facts in regard to conduction through a gas are in support of this idea of the ionic nature of the process.

The method of determining whether a gas is a conductor or not is, of course, to immerse in it two electrodes, at a small distance apart, and to observe whether a current flows when a difference of potential is produced between these plates by some cell or combination of cells. It is found that a pure dry gas is an extremely poor conductor; but it may be made conducting by many means. A few will be mentioned. By passing a spark through any portion of the gas, all other neighboring portions have their conductivity increased. In many cases of complex gases, a sufficient rise in temperature makes them conducting. If the ultra-violet waves from a source of light pass through a gas, it becomes a conductor. There are many solids, *e.g.* alkaline earths and metals, which when illuminated with ultra-violet light make the gas near them conducting. This is called "photo-electric" action. Further, if metals (or carbon) are made very hot, they make the surrounding gas conducting. If the cathode rays or the X-rays (see below) traverse a gas, it is made conducting. Certain bodies, known as "radio-active" substances, have the power of emitting radiations which make the gas through which they pass a conductor; these will be discussed more fully below. In all these cases the gas is said to be "ionized."

Spark and Arc Discharge. — If two electrodes immersed in a gas at not too great a distance apart are raised to a sufficient difference of potential, a spark will pass between them, as already stated. The character of the discharge and the potential difference required depend upon the nature and the condition of the gas: its pressure, its temperature, its purity, etc. If the two electrodes are brought sufficiently close together and the gas is a sufficiently good conductor, so that a large current passes, the discharge is called an "arc," as is illustrated by the ordinary arc lights in the streets. In both the ordinary spark and the arc the electrodes are gradually vaporized, and their vapors, as well as the gas itself, are rendered luminous.

The phenomena of the discharge are so varied that some special treatise on the subject should be consulted. The best is J. J. Thomson, *Conduction of Electricity through Gases*, London, 1903. We shall describe here only one or two cases of special interest.

Vacuum Tube Discharge. — If the gas is inclosed in a glass bulb into which two metallic electrodes enter, and if the pressure is gradually lowered by means of an air pump, the

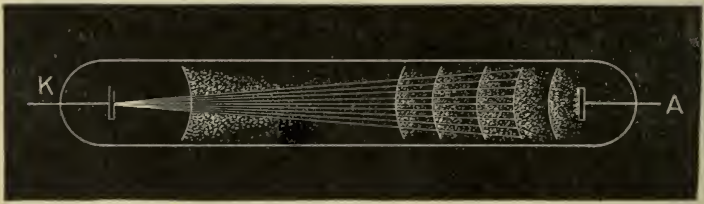


Fig. 381. — Vacuum discharge tube.

character of the discharge changes in a most marked manner. When the pressure is about that of 1 mm. of mercury, it is observed that at the cathode there is a velvety light covering it entirely or in part, which is separated by a dark space from a luminous region, and that this is separated by a second dark space from a luminous striated column extending to the anode, the end of which — if it is a wire — has a bright spot of light. The first dark space near the cathode is called the "Crookes dark space"; the second one, the "Faraday dark space"; the region separating them, the "negative glow"; and the striated portion near the anode, the "positive column." As the bulb is more and more exhausted, the Faraday dark space extends farther down toward the anode, and another phenomenon becomes most prominent. There is a radiation of something from the cathode, proceeding in straight lines *perpendicular to its surface*, quite independent of the position of the anode. This radiation produces a faint luminescence of the traces of gas left in the bulb; and so the path of the rays through the bulb may be seen. They are called the "cathode rays." Where they strike the walls of the bulb, it is made luminous and its temperature rises; the color of the light which is thus produced depends upon the material of the bulb, but ordinary glass emits a greenish yellow light. (If certain other solids are introduced in the bulb in the path of the rays, *e.g.* coral, they emit characteristic colors.) The fact that the rays proceed in straight lines is proved by introducing in the tube solid bodies which are opaque to the rays; for it is observed that they cast sharp shadows on the walls of the tube. The radiation passes directly through metallic films, if sufficiently thin. The path of the rays in the tube is deflected by a strong electric field, provided the vacuum is perfect enough; and the deflection is in such a direction as to lead one to believe that the radiation consists of negatively charged particles. In fact, all experiments lead to this con-

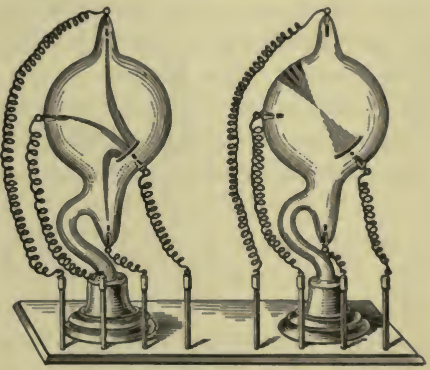


FIG. 882.—Two forms of vacuum tube discharge: *a*, moderate vacuum; *b*, high vacuum.


FIG. 883.—Shadow of maltese cross produced by cathode rays.

clusion. If the rays are made to enter a hollow cylinder which is connected to an electrometer, it is seen that the cylinder is receiving a negative charge; and the fact that, when the rays strike a solid its temperature is raised, is explained by assuming that the rays consist of material particles. Again, as will be shown in a few pages, a charged particle in rapid motion has the same action on a magnet as does an electric current; and, since a wire carrying a current may be made to move under the action of a magnet (which is simply the reverse of the fact that a current can move a magnet), so a particle in rapid motion may have its direction altered if it is charged; and experiments show that the cathode rays may be deflected by a magnet in exactly the manner which one would expect if they were negatively charged particles in rapid motion. The velocity of these rays may be measured in many ways; and it is found to depend upon several conditions: difference of potential between the electrodes, pressure of the gas, etc.; its value is not far from one tenth of the velocity of light. The masses of these particles and their electrical charges may also be measured with a fair degree of accuracy; and it is believed that the charge of any particle is the same as that on a hydrogen ion in ordinary electrolysis, while its mass is approximately one thousandth of that of a hydrogen atom. So far as experiments can prove, these particles which constitute the cathode rays are the same no matter what gas is put in the tube.

If the cathode is in the form of a metal plate with many small openings in it, it is observed that in addition to the cathode rays which are emitted from one side there are rays proceeding in the opposite direction apparently through the holes in the cathode. These are called "canal rays," and have been proved to consist of positively charged particles, moving much more slowly than the cathode rays. Their charges are probably the same as those of the latter rays; but their masses are comparable with those of an atom.

The gas which is traversed by either the cathode or the canal rays is ionized; that is, becomes a conductor. This was proved for the former rays by Lenard, who constructed a glass bulb in such a manner that a portion of the wall which was struck by the rays had an opening in it which was covered by a thin layer of aluminium. Surrounding this bulb was another which could be exhausted if desired. So, as the cathode rays passed through the aluminium window and entered the outer bulb, their action could be studied.

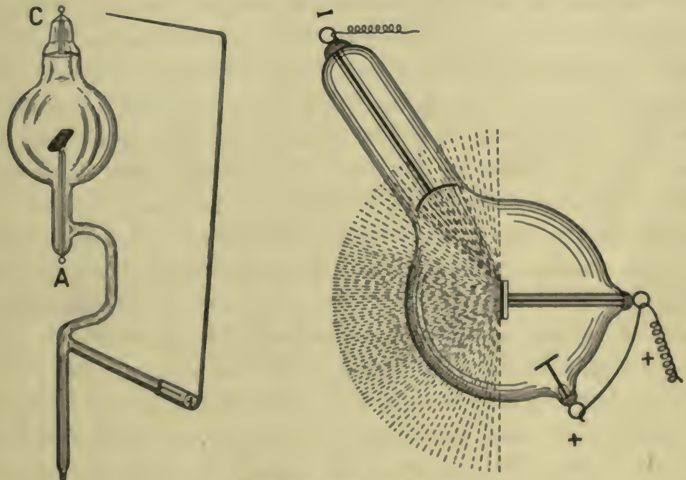


FIG. 384. — An X-ray tube.

The aluminium itself emits cathode rays, as a result of being struck by the interior cathode rays, in addition to transmitting some; and the total radiation from the window is called "the Lenard rays." This includes also a type of radiation entirely distinct from cathode rays, as was first shown by Röntgen. He proved that whenever cathode rays strike a solid surface, a radiation is produced which has most interesting properties, and to which he gave the name X-rays. These rays proceed in straight lines from their source; they are not reflected, refracted, or diffracted; they are not affected by an electric or magnetic field; they cannot be polarized; they affect a photographic plate; they cause many bodies to become luminous; they pass with ease through many bodies opaque to light, *e.g.* wood, aluminium, the human flesh, etc.; they are absorbed by other bodies much more than is light, *e.g.* glass; they ionize a gas. It is believed that they are transverse *pulses* in the ether, in distinction to being trains of waves; and all their properties may be explained on this assumption. If we consider what we would expect to happen when a moving charge is suddenly brought to rest, as the cathode rays are by the solid which they strike, we can understand how these pulses are produced, and how irregular they are. Röntgen's original papers and Stokes' memoirs giving an explanation of the X-rays are reprinted in Barker, *Röntgen Rays*, Scientific Memoir Series, New York, 1899.

Radioactive Bodies. — Certain substances of high atomic weight, namely uranium, thorium, radium, and one or two others, possess, in common with their compounds, such as thorium nitrate, etc., the most remarkable property of emitting spontaneously a radiation which can ionize a gas; they are said to be "radioactive." This radiation is complex, consisting of what are called "α rays," which are positively charged particles analogous to canal rays, and "β rays," which are negatively charged particles analogous to cathode rays. Radium also emits rays which are like X-rays in many respects; they are called "γ rays."

This radiation is accompanied by changes inside the molecules of the substances which emit it; and in some cases the products of the molecular changes are gaseous. Thus, if thorium nitrate is dissolved in water and ammonia is added, a precipitate is formed, which may be separated by filtration. The precipitate is not radioactive at first but gradually becomes so, in fact returning to the same condition as the original thorium nitrate. The filtrate, on the other hand, is most radioactive, but loses this property in time; as it does this, it gives off a gaseous emanation, which is radioactive. This emanation is at first uncharged, but by losing its negative charges it becomes positively charged and may be attracted to any negatively charged body. It undergoes further changes; and during each, radiations are emitted.

The emanation of radium finally decomposes into helium gas. During these processes, further, heat energy is evolved, and the temperature is raised several degrees Centigrade.

Electrons. — When a gas is ionized, either by the action of X-rays or by the rapid motion of minute positively or negatively charged particles through it, all the negative ions are found to be alike in all respects and to be the same as the cathode rays. It is therefore believed that, as such charged particles or as pulses pass through a gas, they break off from its atoms these negative ions. A theory of the constitution of an atom has been based on this idea. An atom is thought to contain within it a great many minute negatively charged particles which are making rapid revolutions — not unlike the constitution of the solar system, the other portions of the atom making up the positive charge. Then ionization would consist in causing one or more of these negative particles to leave the atom. These particles while inside the atom — and when outside also, provided they have no other material particles clinging to them — are called "electrons." Their vibrations inside the atom give rise to waves in the ether. So far as is known a moving piece of uncharged matter does not affect the ether; but, if charged, it does; and, if its motion has *acceleration*, waves are produced.

It is known from theoretical considerations that when a charged particle is in motion, its kinetic energy is greater than it would be if it were not charged; and, therefore, an electric charge in motion by itself — quite apart from matter — would have kinetic energy; that is, would have mass. The question then arises, Is not the inertia of matter really due to the motion of electric charges in its minute parts? In other words, is not a moving charge the fundamental fact in nature? This question is fully discussed, and a most interesting description of the general properties of electrons is given, in a series of papers in the *London Electrician* during 1902–1903, by Sir Oliver Lodge.

Conduction in a Solid. — It has been proved in what has gone before that conduction in an electrolyte and in a gas consists in the actual motion of charged particles, called ions.

In the case of a solid conductor the question as to the mechanism of the conduction of a current is not so simple, owing primarily to the fact that the particles of a solid have so little freedom of motion and can only vibrate. But there is every reason for believing that in a solid also the process of conduction is by means of ions. The existence of free electrons moving about inside the solid conductor from atom to atom may be proved by many experiments; and the evidence in favor of this explanation of conduction is accumulating continually.

Convection Currents. — It was Faraday who first conceived the idea that the essential feature of an electric current was the rapid motion of an electric charge; but the first to prove by direct experiment that such a charge in motion had the same magnetic action as an ordinary current produced by a voltaic cell was the late Professor Rowland. He charged a circular metallic disc and caused it to rotate rapidly on an axle perpendicular to its faces; he observed that when a magnetic needle was brought near the disc, it was deflected exactly as if electric currents were flowing in circles in the disc. He was able to prove that, within the range of velocities used, a charge e moving with a velocity v is equivalent to a current whose strength is ev. It is probable that this is true, even if v is very great, much greater than it is possible to attain by any mechanical means. A current due to a moving charge is called a "convection current."

CHAPTER XLVI

MAGNETIC ACTION OF A CURRENT

General Description. — In a previous chapter a general description of the magnetic field due to an electric current was given; and it was seen that a conductor carrying a current is surrounded by a field of magnetic force, such that the lines of force form closed curves around the current. The relation between the direction of the current and that of the lines of force is given by the right-handed-screw law. The magnetic field, then, due to a circuit carrying a current is the same as if a great many minute magnets, of the same length and strength, are taken and placed side by side so that their north poles are all turned one way and their south poles in the opposite direction, thus forming what is called a "magnetic shell," having the same contour as is made by the conductor carrying the current. For, lines of force would proceed out from the north poles of the shell and all return to the south poles. We can thus speak of the "north face" of a circuit carrying a current and of its "south face."

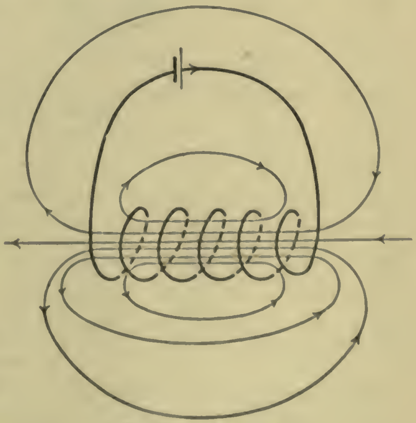


Fig. 385. — A solenoid.

Again, if a wire — or other conductor — is wound in the form of a helix, and if a current is passed through it, thus forming a "solenoid," the magnetic field enters one end and emerges from the other exactly as if it were a bar magnet.

Electro-magnets. — If, then, a bar of iron or of any magnetic substance is inserted in a solenoid, it is magnetized because each little molecular magnet turns and places itself along a line of force. A bar of iron wrapped with a helix of insulated wire is called an "electro-magnet." The bar is

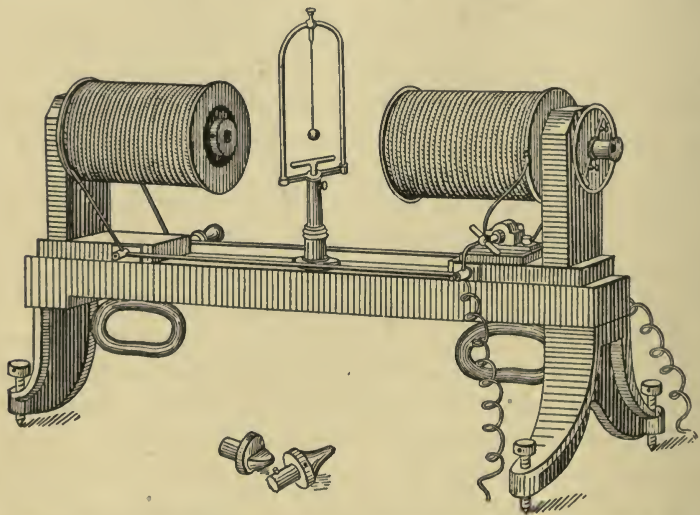


Fig. 386. — A powerful electro-magnet arranged to show magnetic or diamagnetic action on suspended sphere.

usually made in the shape of a horseshoe, or of the general form shown in the cut. These are used in a countless number of instruments, such as call bells, telegraph instruments, etc. The first electro-magnet was made by Sturgeon (1825); but the idea of wrapping several layers of wire, like thread on a spool, is due to Joseph Henry. By this means the intensity of a magnet may be greatly increased.

Electro-magnetic Forces. — Since a solenoid is equivalent to a bar magnet, it will, if suspended free to move, turn and place itself in the magnetic meridian; and, if another solenoid or a bar magnet is brought near it, it will be seen that like poles repel and unlike ones attract. These motions, produced owing to the magnetic field of a current, are said to be due to "electro-magnetic forces"; and we can describe them in another manner which is perhaps more definite. Since the south end of a solenoid is the one which tubes of force enter, and since it is the one which is attracted by the north pole of a magnet or a solenoid from which tubes of force proceed, it is evident that by the act of approaching one another, more tubes of force enter the south end of the solenoid than before. Other cases of attraction and repulsion may be considered in a similar manner; and it is easily seen that they may all be described by saying that motions of conductors carrying currents take place in such a manner that as many tubes of force as possible emerge through their north ends or faces. This is equivalent to saying that the electro-magnetic forces are in such a direction as to produce these motions.

If a loosely wound spiral spring is suspended in a vertical position with its lower end just dipping in a cup of mercury, and an electric current is passed through it, the separate turns of the coil will attract each other, because by coming closer together more tubes of force pass through them, instead of escaping out from the sides; as the spring thus contracts, the electrical connection at the bottom is broken, and so the force of attraction vanishes; the spring then drops, connection is again made, the current flows, etc. The action is increased by inserting an iron rod inside the coil.

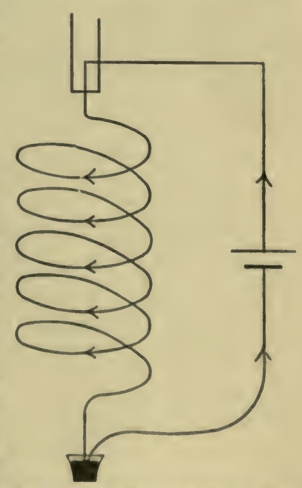


FIG. 387. — Electro-magnetic attraction of parallel currents.

These phenomena of electro-magnetic forces were discovered by Ampère, and he invented many most beautiful experiments to illustrate them. They may be found described in many text-books. He also proposed a theory of magnetism based upon his observations. He advanced the hypothesis that in each molecule of a magnetic substance there is an electric current, flowing in a fixed channel; and therefore if a bar of such a substance is brought into a magnetic field, — due either to a magnet or to a solenoid, — the molecules will all turn so as to include as many tubes of magnetic force as possible. The bar will be saturated when the molecules have so arranged themselves that their currents are parallel to each other and to the ends of the bar; and under these conditions the lines of force due to these currents will emerge at one end and return into the other, exactly like a solenoid.

Fig. 388. — South pole of magnet. Ampère's theory of magnetism.

If two parallel wires or rods are placed in a horizontal plane and are joined by a fixed wire $\overline{BC}$, containing a cell, and by a movable wire $\overline{PP'}$, which can roll on the wires,

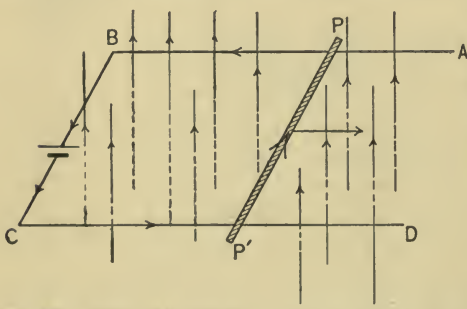


Fig. 389. — Electro-magnetic force: magnetic field is upward through circuit.

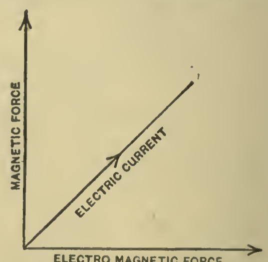


Fig. 390. — Relation between directions of current, magnetic field, and electro-magnetic force. Each is perpendicular to the other two.

this movable wire will be set in motion if there is a magnetic field through the space between the parallel wires, because by so moving a change is made in the number of tubes of force which pass through the circuit ($PP'CB$). If the current is in the direction shown in the cut, the upper face is the north one; and, if the magnetic field of force is in the direction shown, the cross wire $\overline{PP'}$ will move toward the right, so that the circuit incloses more tubes coming out from its north face. This law of force is given by the diagram which describes the connection between the directions of the magnetic force, the electric current, and the electro-magnetic force.

Magnetic Force Due to a Current. — Since the magnetic lines of force due to a current form closed curves around it, it would require a certain amount of work to carry a unit north pole around such a line of force in a direction opposite to that of the line itself; and it is evident that this work must vary directly as the strength of the current. It may be shown by experimental methods that the amount of work required to carry a unit pole around any closed curve encircling the current is the same for all paths. Thus, calling this work W, we may write $W = ci$, where c is a factor of proportionality depending on the system of units adopted for the measurement of the current. The C. G. S. electromagnetic system is based on the definition of such a unit current as will make this factor equal to 4π. (The reason for this choice depends upon the connection between a current and a magnetic shell, and need not be explained here. It may be found in any advanced text-book.) If the unit pole is carried around the circuit m times, the work done is evidently $4\pi mi$. One consequence of this definition of a unit current is that the intensity of the magnetic field at the centre of a circular coil of radius a, of n turns, carrying a current of strength i, is $\frac{2\pi ni}{a}$. (See page 673.)

If the conductor is in the form of a helix, the magnetic force inside it, owing to a current through it, may be deduced at once. The work done in carrying a unit pole around a closed path *abcda*, as shown in the cut, where *ab* is parallel to the axis of the helix and inside it, *bc* and *da* are perpendicular to this axis and *cd* is outside the helix, equals the product of the intensity of the magnetic field inside and the length of the path *ab*, because the lines of force are along the axis, and so no work is required to traverse the portions *da* and *bc*, and the force outside is so small that it may be neglected if the helix is long. So, calling the intensity of the field inside R and the distance ab x, the work is Rx. But, if there are N turns of the helix per unit length, there are xN turns in the length x; and since each of these carries the current i, the work done in threading them by a unit pole is, in accordance with the above definition of a unit current, $4\pi N \times i$. So, $Rx = 4\pi N \times i$, or. $R = 4\pi Ni$; a most important formula.

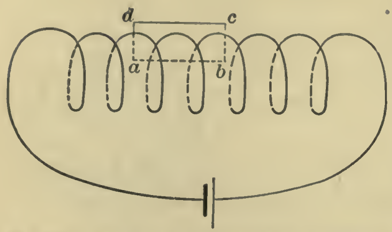


FIG. 891.—Magnetic force inside a long solenoid.

If there is a rod of iron, of permeability μ, filling the solenoid, the number of tubes of induction per unit area passing through the iron is proportional to μR. (See page 615.) Since μ for iron is large, this means that the number of tubes of induction passing through the solenoid is greatly increased by inserting in it a rod of iron. These additional tubes are due to the magnetization of the iron by the current.

Since each of these tubes of induction passes N times through a circuit carrying the current i in a unit distance, it is evident that the magnetic action of the solenoid is proportional to that of a single turn of wire carrying a current equal to N^2i.

Energy of a Current. — The fact that, when a current is flowing in a conductor, forces may be experienced in the surrounding medium, proves that there is a certain amount of energy in this medium due to the current. This energy is not in the form of a strain — no sparks are observed in the medium, etc. So it is natural to think of the energy as being kinetic in its nature; and this idea will become more evident in a later chapter. What will be shown is this: as a current is first started, *e.g.* by joining the two poles of a primary cell, the current does not rise to its full strength instantly, for part of the energy furnished by the source of the current is spent in producing those motions in the surrounding medium which constitute the magnetic field, and it is not until these motions are established that all the energy of the source of the current goes into forcing the current through the conductor, and so heating it. Similarly, if the source of the current is suddenly removed, the current does not instantly cease, because the energy in the medium disappears gradually, being spent in maintaining the current for a short time.

Compare the case of a railway train starting from rest; it does not attain its full speed instantly because the energy furnished by the locomotive is used in producing kinetic energy; but, after the desired speed is reached, the only work done is against the friction of the wheels, the resistance of the air, etc. Similarly, when the train is to be stopped, the power furnished by the locomotive is shut off, but motion continues until the kinetic energy of the train is used up in overcoming friction. The analogy with an electric current is not particularly good, because in the latter case the kinetic energy is not in the conductor, but in the medium, while the work done in producing the heat is spent in the conductor.

From what was said above in regard to solenoids, it is evident that the energy of their currents is proportional to N^2, where N is the number of turns per unit length.

There are many electrical instruments and machines whose actions depend upon electro-magnetic forces; and a few will be described.

The Electric Bell. — This consists of a gong G; an electro-magnet E; a vibrating clapper of spring brass which has a small knob H with which to strike the gong, and a strip of iron facing the poles of the electro-magnet; a stiff piece of brass A, which is attached to the clapper, and presses against a fixed metal stud C. One end of the wire around the electro-magnet is joined to the clapper, while the other

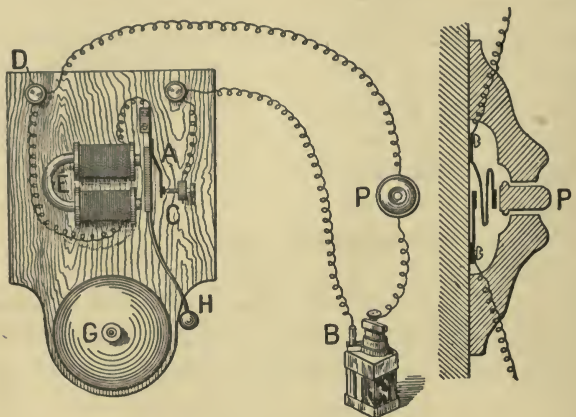


FIG. 392.—Electric bell and push button.

end is connected to a binding post at D. When the bell is in use, D and C are joined to some primary cell B; a contact key P being introduced in the circuit. When the key is pressed so as to make contact, the current flows through the

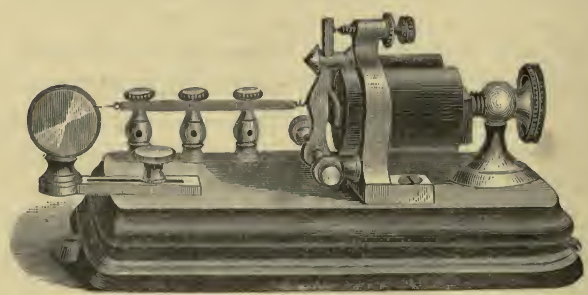


FIG. 393.—The relay.

electro-magnet, attracting the clapper and thus ringing the bell; but as the clapper moves toward the magnet, contact is broken at C, and the current ceases; then, owing to its elasticity, the clapper vibrates back, making contact at *C*, and the current again flows; etc.

Relay. — A relay consists of an electro-magnet in front of which is an iron plate called the armature, carried by a metal rod pivoted at its base. The wire around the electro-magnet may be connected to a primary cell or a battery of cells at a distance, with a key in circuit. So, when the key is pressed, a current will flow around the magnet. Even if the current is extremely feeble, the armature will be attracted; and by means of suitable contact points a second cell may be closed through any circuit; and thus any electro-magnetic effect may be produced, such as ringing a bell, etc. When the key is broken, the current ceases, and the armature is drawn back from the electro-magnet by a spiral spring attached to its rear.

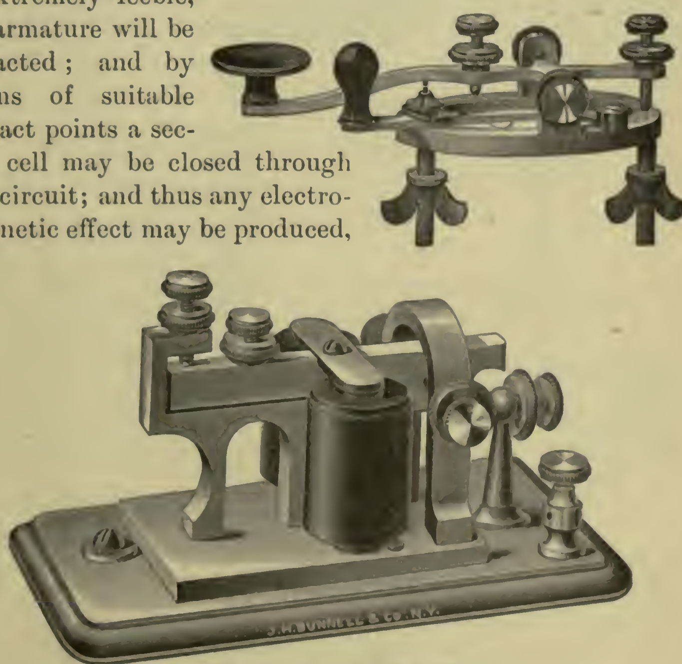


FIG. 894. — Telegraph key and sounder.

If a second electro-magnet with an armature is introduced in the circuit of the second cell, the sound made by the armature as it clicks against the electro-magnet may be clearly heard. This is the principle of the ordinary telegraph system, different letters being distinguished by different combinations of "dots and dashes"; that is, short and long intervals of time between consecutive clicks of the "sounder."

Duplex Telegraphy. — In the "Duplex" system of telegraphy it is possible to receive and send messages from a station at the same time. The instrument consists essentially of a receiving instrument *E*, such as an electro-magnet or a galvanoscope, around which are wound two coils of wire in opposite directions; so, if equal currents are passed through both coils, no effect is produced, while, if a current passes through one coil alone, there is an effect. One of these coils is connected through the "line wire" to the distant station, while the other is joined through several coils of wire of adjustable lengths, *R*, to the earth. The arrangement of the cell and the key *K* is as shown in the cut. The coils *R* are so adjusted that when the key *K* is pressed and makes connection with the cell *B*, equal currents pass around the receiving instrument *E*, and there is no effect; but a current passes over the line wire to the distant station. When a current is received over the line wire, it passes to the earth, entirely regardless of the position of the key *K*, thus affecting the instrument. (If the key *K* is pressed down when a current is being received from the distant station, we may regard the current in the line wire as

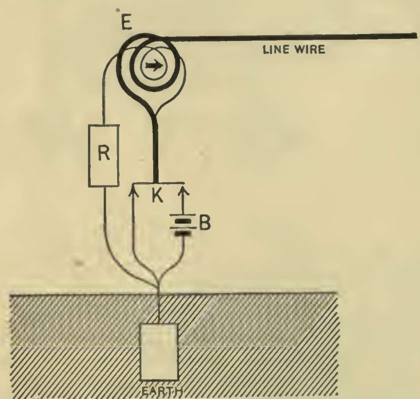


Fig. 395. — Diagram of one form of duplex telegraph instrument.

neutralized, but part of the current from *B* passes around *E* through *R* to the earth.

Telephone. — The ordinary Bell telephone receiver consists of a steel magnet *M*, around one end of which is wound a coil of wire, and in front of which is a thin iron plate called the diaphragm. This is attracted toward the magnet, but is kept from motion as a whole by the frame. If a current is passed through the coil, it will either strengthen or weaken the steel magnet, depending upon its direction; and so the diaphragm is either attracted or repelled. Thus, if the current fluctuates, the diaphragm vibrates.

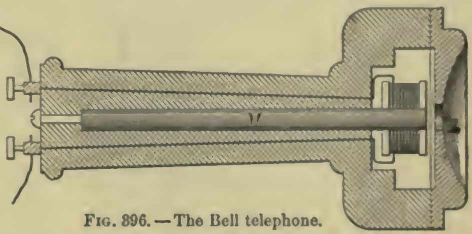


FIG. 896. — The Bell telephone.

Microphone or Transmitter. — This consists of a thin metal diaphragm whose edges are held but which can vibrate like a drumhead; against its centre presses a carbon "button," which is held firmly by suitable metal supports. A cell is connected to the carbon button and to its supports. There is poor electrical contact between these; and it varies in its conducting power as the pressure of the diaphragm against the carbon button varies. If the pressure is increased, more current flows; if it is decreased, less current. So, if the diaphragm vibrates, the current fluctuates in strength.

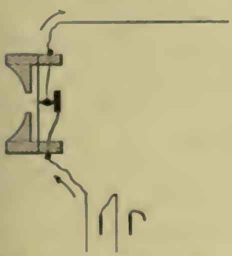


FIG. 897. — Diagram of Blake transmitter.

It is thus evident that, if a telephone is included in the microphone circuit, vibrations of the microphone diaphragm produce corresponding vibrations of the telephone diaphragm. In this manner, vibrations of the former produced by the human voice will cause vibrations of the telephone diaphragm, which will in turn send out sound waves; and thus sounds may be said to be transmitted.

D'Arsonval Galvanometer. — This consists of a permanent steel horseshoe magnet, between whose poles a coil *C* is supported by means of a vertical wire. The wire in this coil is continuous from *A* to *B*, two fixed binding screws. When no current is passing, the coil is held so that its plane is parallel to the line joining the two poles; but if a current is transmitted through the coil by means of *A* and *B*, it will turn so as to include as many of the tubes of force of the magnet as possible. It will be brought to rest by the torsion of the wire; and so its deflections measure the current strength.

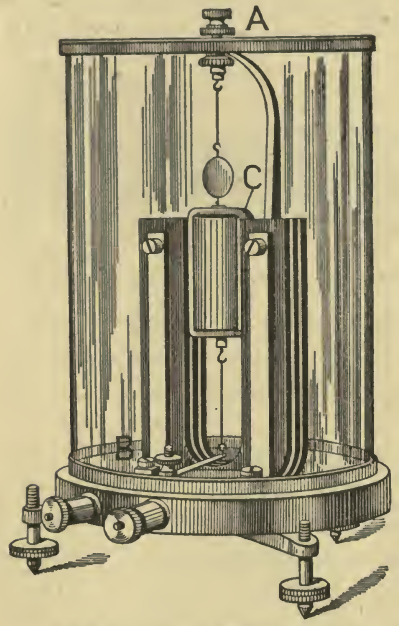


FIG. 398. — D'Arsonval galvanometer.

Practical Instruments. — In most practical work, such as measuring the electric currents of telegraph systems, lighting systems, dynamos, etc., instruments are used which are portable. They are sometimes called "practical instruments." The principle used in them all is to have a permanent steel horseshoe magnet, between whose poles is supported on pivots a coil of wire through which the current to be measured is passed. This coil turns so as to include as many tubes of force as possible; but, as it turns, it winds up a flat coiled spring, and so is finally brought to rest. The angle of deflection is measured by a pointer.

Radio-micrometer. — This is an instrument invented by Professor Boys for the detection and measurement of radiation in the ether. It consists of a thermocouple and a loop of wire, used according to the principle of the coil in the D'Arsonval galvanometer. A loop of copper wire ends in fine strips of bismuth and antimony, *A* and *B*, which are soldered together. This loop is then suspended by a fibre between the poles of a permanent magnet, with its plane parallel to the line joining them. The junction of the two metals is blackened, and is exposed to the radiation; as it absorbs energy, its temperature rises, a current flows in the loop; this is then deflected so as to include as many tubes of force as possible, and it finally comes to rest when this electro-magnetic force is balanced by the torsion of the fibre. This deflection evidently measures the intensity of the radiation absorbed by the blackened junction.

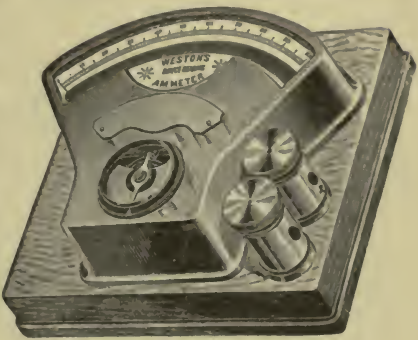


FIG. 399. — Weston's ammeter.

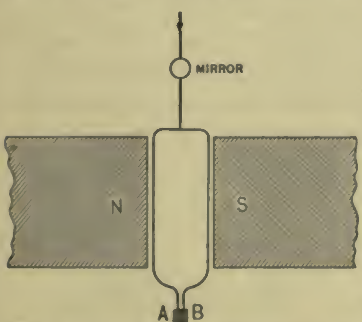


FIG. 400. — Boys' radio-micrometer: *A* and *B* are two different metals forming a thermocouple.

Electric Motor. — This consists of an electro-magnet whose poles are turned to face each other, and of an "armature," which is a shuttle-shaped piece of iron or an iron ring, on which are wound coils of iron. These coils are connected with metal strips, insulated from each other, on the shaft of the armature; and on these "commutator bars," as they are called, rest two metallic rods or "brushes," which are joined to some source of a current such as a dynamo. It is easily seen how it is possible to make the connection of the coils with the bars in such a manner that the current passes through one coil of the armature, which is in such a position that it does not include as many of the tubes of force due to the electro-magnet as it would if the armature turned on its axis; therefore, the armature will turn; and, as it does so, another coil on the armature comes into the position occupied by the previous coil; then the current will pass through this; etc. In this manner continuous rotation of the armature may be secured; and by means of its shaft, work of various kinds may be done; *e.g.* street cars may be moved.

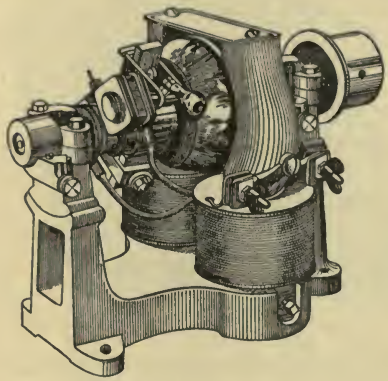


FIG. 401.—An electric motor.

When the motor is doing work, the energy is furnished by the source of the current. If the electro-motive force of this source is E and if the current is i, then the energy it furnishes in a time t is Eit. Part of this goes into heating the conductor, and the rest into the motor. This last may be written eit, where e is called the "back E. M. F." of the motor. So the amount of energy that goes into heat effects is $(E-e)it$.

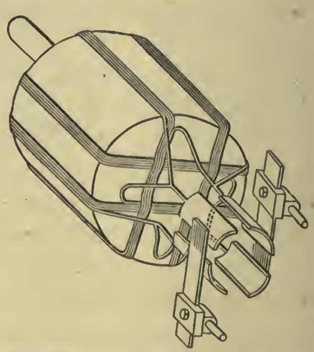


FIG. 401 *a*.—Method of winding coils of wire on a "drum" armature.

CHAPTER XLVII

LAWS OF STEADY CURRENTS

Steady Current. — In the foregoing chapters the various properties of electric currents, viz., heating, magnetic, electrolytic, etc., have been discussed and illustrated; and several methods for the production of currents have been described. A current is called "steady" if these properties remain constant, *e.g.* if a constant deflection of a galvanometer needle is produced, if heat energy is developed at a constant rate, if matter is liberated in an electrolyte at a constant rate, etc.; and experiments prove that, if a current satisfies one of these conditions, it satisfies all. A "variable" current is one that is not steady. In order to produce a steady current one may use a source of constant E. M. F., such as a Daniell's cell, or a thermocouple whose junctions are maintained at constant temperatures, or a dynamo — as will be described in the next chapter.

Uniformity of Current. — One of the most important properties of a steady current is that its strength is uniform throughout the circuit; that is, if the circuit includes conductors of different material, of different sizes, etc., the strength of the current is the same in them all. This may be shown by proving that the magnetic or the heating action of the current is the same for all portions of the circuit. Again, if the current were not the same at all points, there would be accumulations of charges at certain points; and, as these increased, they could be detected; but such is not the case.

Similarly, if at any point of the circuit, it branches so as to form two or more parallel conductors, the strength of the current in the single conductor must equal the sum of the strengths of the currents in the branches. This may be expressed in a formula,

Fig. 402. — A divided circuit.

$$i = i_1 + i_2 + \dots,$$

or,

$$i - i_1 - i_2 - = \dots 0.$$

So, if i_n is the current in any conductor at a branch point, the direction of the current being called positive if it is toward the point, the summation of all the currents at that point is zero, or, in symbols, $\Sigma i_n = 0$.

Ohm's Law. — We have seen that a current flows between two points of a conductor only if there is a difference in potential between them; so that we may in a way regard the E. M. F. as the cause of the current, and it is not unnatural, judging from analogy with the flow of heat in a bar owing to difference in temperature between two points, to advance the hypothesis that the current strength in a conductor varies directly as the E. M. F. between two points, provided the current is steady. (We are considering the case where there is no cell or other source of E. M. F. introduced in the conductor between the two points.) That is, if A and B are any two points in a circuit in which is flowing a steady current of strength i, and if E is the E. M. F. between A and B, the hypothesis is that $E = Ri$, where R is a constant depending upon the nature of the conductor between A and B, but not upon the values of E or i. This hypothesis has been found to be true, so far as experiments can decide; E has been varied by introducing more cells or a dynamo, and the resulting current

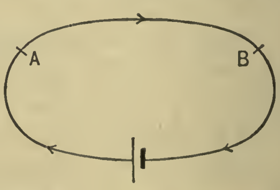


Fig. 403. — Diagram to illustrate Ohm's law.

has been measured. It is called Ohm's law, having been proposed by Georg Ohm in the year 1826.

Resistance. — It is evident from the formula that if R is large, i is small, provided E remains constant; while, if R is small, i is great. For this reason R is called the "resistance" of the conductor between A and B. This law can also be written $i = \frac{E}{R}$; and, if for $\frac{1}{R}$ the symbol C is substituted, the formula becomes $i = CE$. For obvious reasons C is called the "conductance" of the conductor between A and B.

If the conductor between A and B is a uniform wire, it is evident that the E. M. F. between A and B is exactly twice what it is between A and a point halfway to B. Therefore, since the current is uniform, the value of R for the conductor between A and B must be twice that for half the length. So, in general, the value of R for any portion of a uniform conductor of constant cross section varies directly as the length of this portion.

If, while a constant E. M. F. is maintained between the two points A and B, a second conductor is introduced between them, identical with the first one, each will carry a current $i = \frac{E}{R}$, and so the current is doubled or the total resistance is halved. The same would be true if, instead of using two conductors, one of twice the cross section were introduced. So, in general, the resistance of a conductor varies inversely as its cross section.

Direct experiments show that if the same E. M. F. is applied at the ends of conductors of the same length and cross section, but of different materials, the resulting current is different. This and the two previous statements may be expressed in a formula,

$$R = \frac{cl}{a},$$

where R is the resistance of a uniform conductor of length l

and of constant cross section a, and c is a constant for a conductor of any one material, but differs for different ones. This quantity c is called the "specific resistance" of a substance, or its "resistivity." Similarly, the conductance $C = \frac{1}{R} = \frac{1}{c}\frac{a}{l}$, and may be written $C = k\frac{a}{l}$, where $k = \frac{1}{c}$. This constant k is called the "conductivity" of a substance.

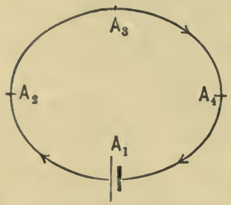


FIG. 404.—Conductors in series.

Illustrations of Ohm's Law.—1. *Conductors in series.* Let the circuit consist of several conductors in series, and let the resistances of the portions $\overline{A_1A_2}$, $\overline{A_2A_3}$, $\overline{A_3A_4}$, be R_1, R_2, R_3; further, let the potentials at the points A_1, A_2, A_3, etc., be V_1, V_2, V_3, etc. Then applying Ohm's law to the separate sections,

$$i = \frac{V_1 - V_2}{R_1} = \frac{V_2 - V_3}{R_2} = \frac{V_3 - V_4}{R_3}.$$

Hence,
$$i = \frac{V_1 - V_4}{R_1 + R_2 + R_3}.$$

The total resistance between A_1 and A_4 is by definition $\frac{V_1 - V_4}{i}$; and it is seen that its value is $R_1 + R_2 + R_3$.

In general, then, the total resistance of a number of conductors in series equals the sum of the resistances of the separate parts. (The fact that R varies as l, the length of a conductor, is a special case of this.)

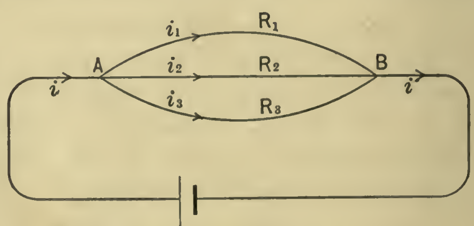


FIG. 405.—Conductors in parallel.

2. *Conductors in parallel.*—Let the circuit branch at any point A into two or more conductors which meet again at B; let the resistance of these branches be R_1, R_2, R_3, etc.; and let the currents

flowing in each be i_1, i_2, i_3, etc. The total current is $i = i_1 + i_2 + i_3 + \cdots$; and the total resistance between A and B is by definition $\frac{V_A - V_B}{i_1 + i_2 + i_3 + \cdots}$. Applying Ohm's law to the various branches, we have

$$i_1 = \frac{V_A - V_B}{R_1};\quad i_2 = \frac{V_A - V_B}{R_2};\quad i_3 = \frac{V_A - V_B}{R_3};\ \text{etc.}$$

Hence $$i_1 + i_2 + i_3 + \cdots = (V_A - V_B)\left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \cdots\right);$$

and therefore, calling the total resistance R,

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \cdots.$$

This may be expressed more simply in terms of conductances, for $C = \frac{1}{R}$; hence,

$$C = C_1 + C_2 + C_3 + \cdots;$$

or, in a branched circuit the total conductance equals the sum of the conductances of the branches. (The fact that R varies inversely as the cross section of a conductor is a special case of this.)

It should be noted, further, that the ratio of the currents in any two branches equals the inverse ratio of the resistances of these branches. Thus,

$$i_1 : i_2 = \frac{1}{R_1} : \frac{1}{R_2}.$$

3. *Wheatstone bridge.* — This is a particular arrangement of six conductors; four form a circuit $ABCD$, and two connect the diagonal points A and C, and B and D. This network of conductors is used in many experimental methods; but only one will be described here. In this a cell is introduced in one of the diagonal branches, say $\overline{AC}$, and

FIG. 406. — Wheatstone's bridge.

a galvanoscope in the other. If the arrangement is such that A is joined to the positive pole of the cell, the potential of A is higher than that of C; and the potentials of B and D are both less than that of A and greater than that of C. So it must be possible to find a point D in the branch $\overline{ADC}$ whose potential equals that of any given point B in the branch $\overline{ABC}$. If this is the case in the actual arrangement, no current will flow across from B to D, and the current flowing from A to B will equal that from B to C; and that flowing from A to D will equal that from D to C. Call the potentials, at A, B, C, and D, V_A, V_B, V_C, and V_D (it should be noted that $V_B = V_D$); the resistances of $\overline{AB}$, $\overline{BC}$, $\overline{AD}$, and $\overline{DC}$, R_1, R_2, R_3, and R_4; and the currents in $\overline{AB}$ and $\overline{BC}$, i_1; and in $\overline{AD}$ and $\overline{DC}$, i_2. Then

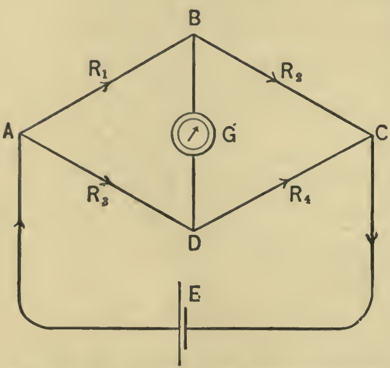


Fig. 407.—Arrangement of Wheatstone's bridge for the comparison of resistances.

$$i_1 = \frac{V_A - V_B}{R_1} = \frac{V_B - V_C}{R_2},$$

$$i_2 = \frac{V_A - V_D}{R_3} = \frac{V_D - V_C}{R_4}.$$

But $V_B = V_D$; hence $\frac{R_1}{R_2} = \frac{R_3}{R_4}$, or $R_1 R_4 = R_2 R_3$.

This formula evidently offers a method for the comparison of resistances. For, suppose $\overline{AB}$ and $\overline{BC}$ are two conductors the ratio of whose resistances is desired. They can be joined in series, and A and C can be joined by a *uniform wire;* then, making the diagonal connections and introducing the galvanoscope and the cell, the end D of the wire connected to the former may be moved along the uniform wire until there is no deflection of the galvanoscope needle. Then $V_B = V_D$, and the bridge is said to be "balanced." As just proved, when

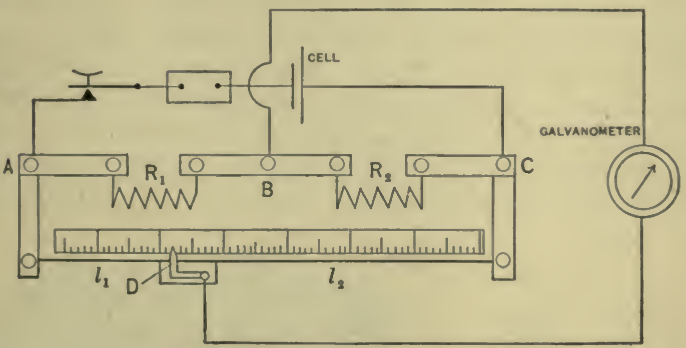
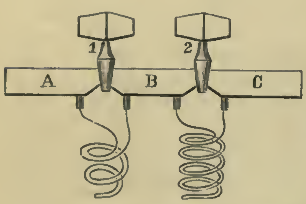


Fig. 408. — A Wheatstone "wire bridge."

this condition is secured, $\frac{R_1}{R_2} = \frac{R_3}{R_4}$; but $\frac{R_3}{R_4}$ equals the ratio of the *lengths* of the portions of the wire $\overline{AD}$ and $\overline{DC}$; and, as these can be measured, the ratio $\frac{R_1}{R_2}$ is known. Similarly, even if the conductor $\overline{ADC}$ is not a uniform wire, but if the ratio $\frac{R_3}{R_4}$ is known, that of R_1 to R_2 may be determined.

As will be explained in the next section, a column of mercury at 0° C., of a uniform cross section of a length 106.3 cm., and having a mass of 14.4521 g., has a resistance which is called an "ohm"; so using a "wire bridge," that is, one in which $\overline{ADC}$ is a uniform wire, and introducing the column of mercury in the "arm" $\overline{AB}$, a wire forming the "arm" $\overline{BC}$ can be made, by altering its length, to have also a resistance of 1 ohm, or of 2 ohms, etc. So, combining these and proceeding in an obvious manner, a series of coils can be made whose resistances are 1, 2, 3, 4, 5, 10, . . . 1000, 5000, etc., ohms. These coils may be so arranged in a convenient box that any one or any combination of them may be used. The ends of each coil are joined to large brass blocks which are insulated otherwise from each other, as shown in the cut; so that there is a continuous circuit from the first block to the last through the coils. These blocks may, however, be directly connected by the insertion of brass "plugs"; and when any one plug is in place, the corresponding resistance coil is "short circuited"; that is, its resistance is so much greater than that of the plug that any current through the box will pass directly through the plug. (It may be assumed for ordinary purposes that the resistance of the blocks and



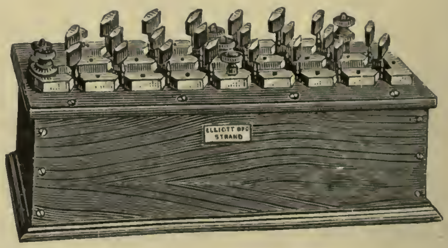


FIG. 409. — Resistance box, showing method of winding coils.

plugs is zero.) These resistance coils are wound, as shown in the cut, in such a manner that the current flowing in one portion is immediately next a current flowing in the opposite direction; so that the coil has no magnetic action. (The magnetic shell to which it is equivalent has an extremely small area. See page 707.)

Definition of the "Ohm," the "Ampere," and the "Volt." — Methods have been described in previous chapters for the measurement of current and E. M. F. on the C. G. S. electromagnetic system, and so the resistance of any conductor may be determined, simply using Ohm's law $E = iR$. (Other methods will be given shortly.) It is found that if the C. G. S. electro-magnetic system of units is used in expressing the values of current and E. M. F. the numbers obtained for the resistances of ordinary conductors of moderate lengths are enormous; and so in ordinary cases a unit of resistance 10^9 times that on the C. G. S. electro-magnetic unit is used. This unit is called an "ohm." As the practical unit of E. M. F. a volt is used (see page 673), which equals 10^8 times that C. G. S. electro-magnetic unit. Therefore, if the value of the E. M. F. applied at the ends of a certain conductor is E_1 volts, $E = E_1 \, 10^8$; and, if the resistance of this conductor is R_1 ohms, $R = R_1 \, 10^9$. So the current strength i is $\frac{E}{R} = \frac{1}{10}\frac{E_1}{R_1}$. Therefore, if the numerical value of the current is to be deduced from Ohm's law, using ohms and volts in which to express resistances and electro-motive forces, a unit current must be defined whose value is one tenth that of a C.G.S. electro-magnetic unit current. Such a current is called an "ampere." For if, in the above experiment, the current strength is i_1 *amperes*, $i = \frac{i_1}{10} = \frac{E_1}{R_1}\frac{1}{10}$; or $i_1 = \frac{E_1}{R_1}$. The quantity of current carried by a current of i_1 ampere flowing for t seconds is called $(i_1 t)$ "coulombs"; and the capacity of a condenser, which when charged with e coulombs is V volts, is said to be $\frac{e}{V}$ "farads." A capacity of one millionth of a farad is called a "micro-farad."

The definitions just given of the ohm and the volt are not, strictly speaking, correct. As may be easily understood, the simplest way of defining a unit of resistance is to select some standard conductor and call its resistance the unit; but, since there are great advantages in using the C. G. S. system, or quantities which may be expressed in terms of it by a certain number of factors 10, the best manner of defining a unit is to select some conductor whose resistance, when determined as accurately as possible in terms of the C. G. S. system, is simply expressed in terms of it, and then to adopt the resistance of this conductor as the unit. Thus, the ohm i defined to be equal to the resistance of a column of mercury at 0° C., of uniform cross section, of length 106.3 cm. and having the mass 14.4521 g. (This column, then, has a cross section of almost exactly 1 sq. mm., accepting the usual value for the density of mercury at 0° C.) The resistance of this column of mercury is equal to 10^9 C. G. S. electro-magnetic units, to within the limits of accuracy of our present experimental methods.

The volt is defined to be the E. M. F. which, steadily applied to a conductor whose resistance is 1 ohm, will produce a current of 1 ampere. It is therefore practically equivalent to 10^8 C. G. S. electro-magnetic units. The E. M. F. of a certain cell, known as the "Clark cell," which can be made in a definite manner, is found by careful experiments to be 1.4322 volts at 15° C. The E. M. F. of the "cadmium cell," which is another standard cell, is found to be 1.0186 at 20° C. (The E. M. F. of the latter cell changes with the temperature much less than that of the former.)

Heating Effect. — It was shown on page 664 that the heat energy developed in a conductor carrying a current of strength i in a period of time t was Eit, where E is the difference of potential at the ends of the conductor considered. As was also noted, the number expressing this quantity of heat is in *ergs*, if the C. G. S. electro-magnetic system is used. If the current is steady, this quantity may be expressed in other ways, for $E = iR$. So, writing $W = Eit$, we have $W = i^2Rt = \frac{E^2t}{R}$. It is seen, then, that the heating effect is independent of the *direction* of the current, because the *square* of the current enters the formula, and it has the same value for either a plus or a minus sign.

If the current is i_1 *amperes* and the resistance is R_1 *ohms*, $i = \frac{i_1}{10}$, and $R = R_1 10^9$; so $W = i_1^2 R_1 t\, 10^7$ ergs. But 10^7 ergs equal 1 joule, and the *power* of a machine which does an amount of work of 1 joule per second is said to be 1 *watt*.

So, if currents are measured in amperes and resistances in ohms, the power of the current is $i_1^2 R_1$ watts.

Measurement of Resistance Absolutely. — By means of a Wheatstone bridge one may determine the *ratio* of the resistances of two conductors, but it does not furnish a method for the measurement of a resistance directly. The formula just deduced for the heating effect of a current does, however, suggest a method. If a coil of wire is immersed in a calorimeter containing water, the heat produced in a given time by passing a current through it may be measured in calories, and since the mechanical equivalent of heat is known, the value in ergs may be deduced. The current strength may be measured by a galvanometer and, since thus W, i, and t are known, the value of R may be determined; for $R = \frac{W}{i^2 t}$. (There are other methods which are more accurate.)

FIG. 410. — Calorimeter for measuring the heating effect of a current.

Temperature Effect. — Experiments show that the resistance of a given conductor varies with its temperature. This is what might be expected, because it was shown that the resistance of any uniform conductor of length l and of cross section a could be expressed $R = c\frac{l}{a}$, where c was a constant, characteristic of the conductor. But, if the temperature of the conductor is altered, the motion and the distance apart of its molecules are affected; and so it is no longer the same substance. It is found that as the temperature is increased the resistance of all solid conductors — with one or two exceptions —

increases, while that of most liquid conductors decreases. (The effect in this last case is complicated by the fact that the extent of the dissociation produced in the act of solution varies with the temperature.) As the temperature of a pure solid is lowered toward absolute zero, its resistance almost vanishes.

This change in resistance of a conductor with change in temperature offers at once a method of making a "resistance thermometer"; and, in fact, a "platinum thermometer" consisting of a coil of platinum wire whose resistance can be measured is at the present time the most satisfactory thermometer in use for accurate work. Similarly, the same phenomenon is made use of in the "bolometer," an instrument for the detection and the measurement of radiation in the form of ether waves. A strip of platinum is covered with lampblack, so that it absorbs as completely as possible all radiation that falls upon it, and is made to form one arm of a Wheatstone bridge. As ether waves are incident upon it, its resistance changes, and the amount of the change measures the intensity of the radiation.

CHAPTER XLVIII

INDUCED CURRENTS

THE discovery by Oersted of the fact that an electric current produced a magnetic field, and the subsequent discovery of methods for making a bar of iron a magnet by means of a current, led many investigators to seek for means of producing an electric current by means of a magnet. The method of doing this was discovered independently by Joseph Henry in America and Michael Faraday in England about 1831.

Experiments of Henry. — Henry's experiments were the earlier. He observed that, if a circuit in which there was a

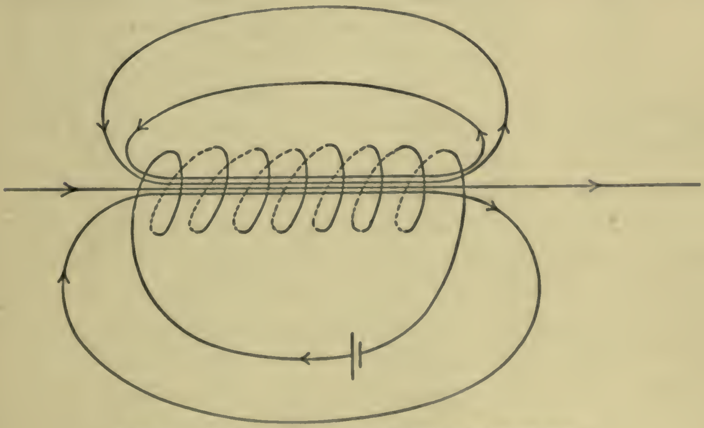


FIG. 411. — A solenoid, illustrating Henry's first experiment.

battery of cells was broken at any point, there was a faint spark; and, further, if the break was made by means of the hands, so that the circuit was completed by the arms and body, a shock was felt. He noticed, too, that both the effects were increased by increasing the length of the conductor and by coiling it up into a helix. There is thus an "extra-current" on breaking a circuit, in addition to the one due to the battery; and Henry's experiments prove that this current varies as the magnetic field of the original current; for, if the conductor forms a helix, the magnetic field is much greater than if the conductor forms simply an approximately circular circuit. A few years later Henry observed that if a wire were wound around the soft iron armature of a horseshoe electro-magnet and if the current were suddenly broken, or if the armature were suddenly removed from the magnet, a shock would be felt, if the two ends of the wire were held in the hands; or, if these ends were joined to a galvanometer, a sudden deflection of the needle would be produced, but the needle would return to its original position. The same effects are produced if the current is again made or if the armature, when separated from the magnet, is brought close to the magnet, but the current in the galvanometer is in the opposite direction. The quantity of the current in the galvanometer, or the shock received by the arm, varies with the number of turns of wire on the armature; and the shock varies with the suddenness of the motion of the armature; the current also varies with the material of the conductor, while the shock does not. It is evident that these "induced" currents, as they are called, are due to the change in the number of tubes of magnetic induction which pass through the coil of wire wound on the armature.

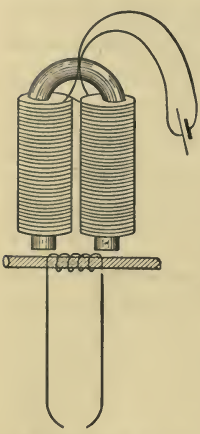


FIG. 412.—Diagram representing Henry's second experiment.

Experiments of Faraday. — Faraday's experiments were somewhat different. He had two separate coils of wire wound on the same iron ring, one coil being joined to a cell, the other to a galvanometer; and he observed that, if he broke the current or made it again, there was a sudden fling of the needle, but that the current was only a transient one. Here, again, the induced current is due evidently to the change in the number of tubes of magnetic induction through the circuit which is joined to the galvanometer. Faraday then showed by a series of most brilliant experiments that if the number of tubes of magnetic induction inclosed by any closed conducting circuit is varied in any manner, *e.g.* by bringing up or removing a magnet or another circuit carrying a current, there is an induced current, whose strength varies directly as the change in the number of tubes of magnetic induction and as the rate of this change, and also depends upon the material of the circuit. If iron is inside the circuit, it is magnetized by the current; and thus the induction is changed. (It was, in fact, owing to this study of induced currents that Faraday was led to his conception of tubes of induction and to the idea of these tubes being continuous through a magnet. See page 615.)

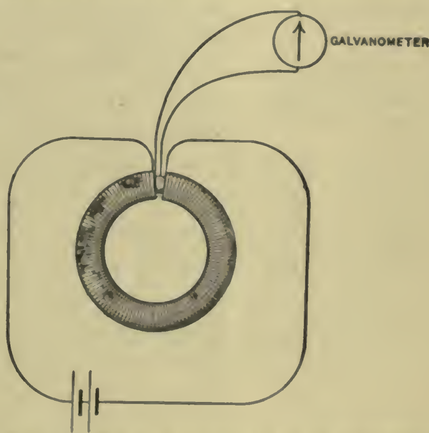


FIG. 413. — Faraday's double coil in his first experiment.

Many years later Faraday rediscovered the phenomena of the extra current on breaking a circuit. He arranged his apparatus as shown in Fig. 414, where E is a cell, C is a helix or electro-magnet, and A and B are the two ends of a broken wire in parallel with the helix. He observed that, if A and B are held in the hands and the electrical current is broken at E, a shock is felt. Similarly, if A and B are joined to a galvanometer, there is a sudden fling of the needle when the circuit is broken. Just before the current is broken, there is a magnetic field through the helix; but, when the circuit is broken at E, there is still a closed circuit around the helix and through $\overline{AB}$; and the magnetic field in this now decreases, since the cell is out of circuit, and so there is no E. M. F. to maintain the current. Owing to the change in the number of tubes of induction in this circuit there is the extra current.

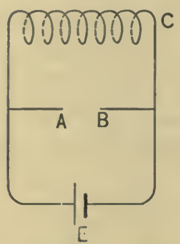


Fig. 414.—Diagram representing Faraday's second experiment.

Law of Induced E. M. F. — All of the facts discovered by Henry and Faraday in regard to the strength of induced currents may be expressed by saying that, when the number of tubes of magnetic induction inclosed by a closed conducting circuit is varied, there is an induced E. M. F. in this circuit whose value is proportional directly to the change in this number, and inversely to the time taken for the change. If there are n turns of the wire, as in a helix or coil, the tubes pass through each, and the induced E. M. F. is n times as great as if there was but one turn. Thus, calling the change in the number of tubes of magnetic induction ΔN, and the time taken for this change Δt, the induced E. M. F. during this change equals $c\frac{\Delta N}{\Delta t}$, where c is a factor of proportionality. It may be proved by methods of the infinitesimal calculus that using the C. G. S. electro-magnetic system to express the E. M. F. and the C. G. S. definition of a unit pole as the source of a unit magnetic tube, this factor c has the numerical value 1. Thus, we may write $E = \frac{\Delta N}{\Delta t}$.

If R is the resistance of the circuit, the strength of the induced current is $\frac{E}{R}$, or $i = \frac{1}{R}\frac{\Delta N}{\Delta t}$. Hence the *quantity* of the current in time Δt, or $i\Delta t$, is $\frac{\Delta N}{R}$. The current strength varies as the change is made, but the total induced quantity, or the summation of $i\Delta t$ during the entire change, equals the summation of $\frac{\Delta N}{R}$; that is, it equals the total change in N divided by R. It must be remembered that if the wire is coiled up, this quantity N varies directly as the number of turns. Thus the induced E. M. F. varies with the *suddenness* of the change, while the induced quantity does not; the latter depends upon the resistance of the circuit, and upon the total change in the number of tubes; *i.e.* upon the field of magnetic force, and upon the area and number of turns of the coils. These facts are shown by Henry's and Faraday's experiments; because the shock received by one's arms is conditioned by the induced E. M. F., while the fling of the galvanometer needle measures the induced *quantity* of the current.

It must be particularly noted that there is an induced current in the circuit only so long as there is a *change* in the number of tubes of magnetic induction through the circuit.

The direction of this current was investigated by both Henry and Faraday. Their conclusions may be expressed by saying that, if the change in the magnetic field through the circuit is an *increase* in the number of tubes of induction, the induced current is in such a direction as by its own magnetic field to *decrease* the number; or, if the change in the field is a *decrease* in the number of tubes, the induced current is in such a direction as to *increase* the number. In general, then, the induced current produced by any change in the magnetic field through it is in such a direction as to tend to neutralize this change. (If this were not true, an increase in the magnetic field would induce a current in such a direction as to increase the field still more; this second increase would produce a second induced current in the same direction, etc.; so conditions would be unstable.) If there is already a current flowing in the circuit, the induced current is superimposed upon it, either increasing or decreasing it.

Special Cases. — A few simple cases will be considered; if a current is flowing in a circuit, and if a bar magnet is made to approach it or to recede from it, the direction of the induced current may be at once predicted. If the north pole of the magnet is nearest the south face of the circuit, some tubes due to the magnet pass out of the north face of the circuit. So, if the magnet is brought nearer the current, more tubes will pass through it, and the induced current will be in such a direction as to oppose this change; *i.e.* it will be in a direction opposite to that of the original current. Thus the current in the circuit is decreased as long as the magnet is approaching. This means, expressed in other language, that work is required to move the magnet, and since this is done by the current, only part of the energy of the cell is available for forcing the current around the circuit. Conversely, if the magnet is withdrawn, the field of force through the circuit is decreased, and the induced current is in the same direction as is the original current. This means that work is being done by whatever agency moves the magnet; and this work appears as an increased current. The case when the bar magnet is turned with its south face toward the south face of the circuit may be treated in a similar manner.

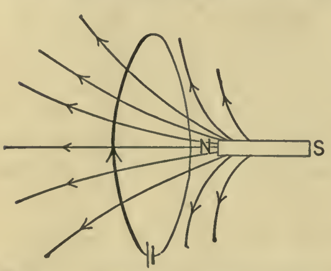


FIG. 415. — Diagram to illustrate induced currents.

Earth Inductor. — If a coil of wire is arranged so as to turn on an axis parallel to its plane faces, it may be so placed that this axis is vertical; and then, if the face of the coil is perpendicular to the magnetic meridian, it will include a field of magnetic force due to the earth. If the coil is turned into the magnetic meridian, there is no field through the coil; and, if it is turned 90° farther, the original field of force will pass through it, but in the opposite direction with reference to the coil. So it is just the same as if the coil had remained stationary and the field of force had changed from N to 0 to $-N$. The total change, then, is $2N$. If A is the area of the face of the coil, if there are n turns of wire in the coil, and if H is the horizontal component of the earth's magnetic field, $N = nAH$. So, if the terminals of this coil are joined to a ballistic galvanometer, the quantity of current measured when the coil just described is suddenly turned through 180°, from a position perpendicular to the magnetic meridian, equals $\frac{2nAH}{R}$, where R is the resistance of the circuit. Similarly, if the coil is so turned that its axis of revolution is in the magnetic meridian and its faces are horizontal, $N = nAV$, where V is the vertical component of the earth's magnetic force; and if the coil is turned on the axis through 180°, the quantity of the induced current is $\frac{2nAV}{R}$. Therefore the ratio of these two quantities equals $\frac{V}{H}$, which is the tangent of the angle of dip. (See page 619.) Such an instrument is called an "earth inductor." It was invented by the great German physicist, Weber.

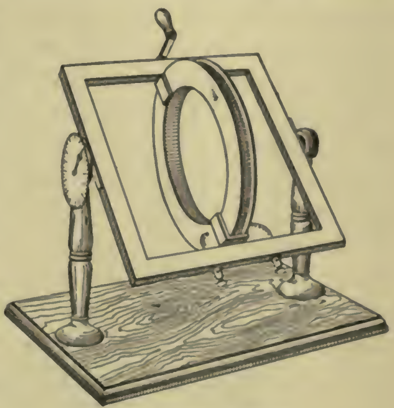


FIG. 416. — Earth inductor.

Induction Coil. — A case of special interest is one studied by both Henry and Faraday: a coil of wire is wound on a spool or cylinder of such a size that it will slip inside another

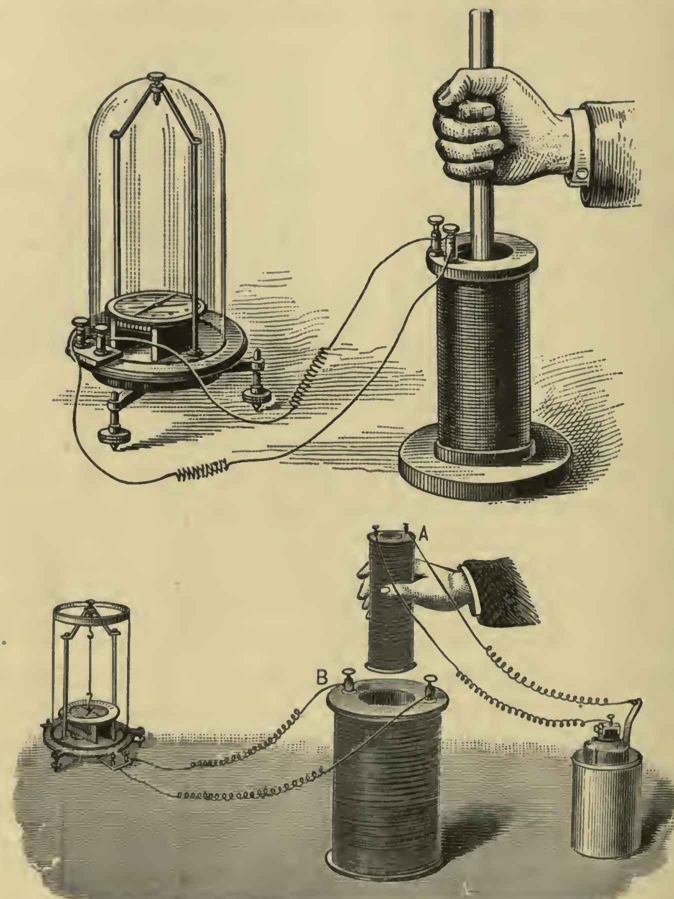


Fig. 417. — Methods for production of induced currents.

cylinder, on which is wound another coil. Call the first coil *A*, the second *B*. If the terminals of *A* are joined to a cell, and those of *B* are connected to a ballistic galvanometer, no current will flow in the latter until the current in A is varied; but, if this is done, a current is induced. If there is a rod of soft iron inside A, the induced currents are greatly increased when the current is made or broken; because when the current is flowing, this rod is magnetized, and so the magnetic field is increased. If a magnet is brought near this iron rod or is taken away, there are also induced currents. Similarly, if the terminals of B are held in the hand, a shock is felt when the current in A is made or broken; and this shock is increased by inserting a piece of iron in A. The shock on breaking the circuit is greater than on making, because the time taken for the change in the magnetic field is less in the former case, and so the induced E. M. F. is greater. The induced *quantity* is the same in both cases. If a thin copper (or conducting) cylindrical tube is interposed between A and B, it prevents almost completely the shock felt at the terminals of B, but does not affect the quantity of current as shown by a ballistic galvanometer. The reason is that, as the current in A is changed, electro-motive forces are induced both in B and in the copper tube; and so, as the current in the *tube* changes, it also induces an E. M. F. in B; these two induced electro-motive forces are in opposite directions and so the resultant effect is small.

If the current in A is large, and there are enough turns of wire on B, an E. M. F. may be induced in B, when the current in A is broken, sufficient to spark across considerable distances in case the terminals of B are separated. This makes the ordinary "induction coil," as shown in Fig. 418. There is a mechanical arrangement for automatically breaking and making the circuit in A, whose principle is evident. The iron core of such a coil is always made of iron wires insulated from each other and is not a solid rod, because as the current in A changes, induced currents would be produced in a solid iron rod in circles around the axis of the rod, and these are prevented by the division of the rod into wires. (These currents produced in a solid core are called "eddy," or Foucault currents.) A condenser is always introduced in the battery circuit in parallel with the "primary" coil *A*. One of its chief functions is to prevent sparking

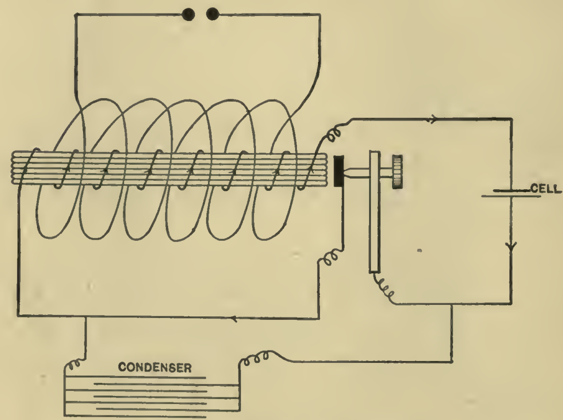


FIG. 418. — Diagram of induction coil.

at the points where the circuit is broken; it does this by diverting the extra current in the primary from the two points where the circuit is broken into the two plates of the condenser. (In other words, to produce a spark, a definite potential difference is required, depending upon the distance; and the difference of potential of the two plates of the condenser does not rise sufficiently high to allow a spark to pass, provided the capacity is great; for $V_1 - V_2 = \frac{e}{C}$. See page 652.) Thus, if the extra current on breaking the primary circuit is prevented, the *change* in the field through the "secondary" circuit *B* is very sudden, and the induced E. M. F. is intense.

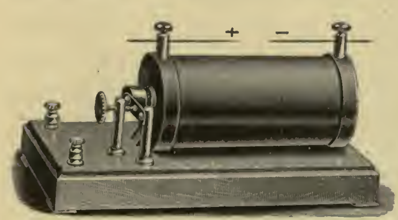


FIG. 418 *a*. — Induction coil.

When the current in the primary is again made, the change in the magnetic field is comparatively slow, and so the induced E. M. F. in the secondary is not great.

Self and Mutual Induction. — 1. *Self-induction.* If a current is flowing in a circuit, it has a field of force of its own: if i is the current strength, the number of tubes of magnetic induction which thread this current is proportional to it and may be written Li, where L is a constant for the given conductor and for a given medium surrounding it. L is called the "coefficient of self-induction" or the "inductance." It is evident that L varies directly as μ, the permeability of the medium; for the magnetic induction equals μ times the magnetic force, and the latter depends simply upon the current and the shape of the conducting circuit. (Thus the effect of introducing an iron rod into the circuit is explained. In the case of iron it must be remembered that μ is not a constant, for it depends upon the intensity of magnetization. So L is not a constant unless the medium is kept the same.)

Further, L must increase as the area of the circuit increases, because the circuit will include more tubes. In the case of a solenoid which has N turns per unit length, the magnetic force inside has been shown to be $4\pi Ni$; therefore, if l is the length of the solenoid, each tube of force passes through the current Nl times; and, if A is the area of the cross section of the solenoid, $L = 4\pi N^2 lA$ if the medium is air, and equals $4\pi\mu N^2 lA$ in general.

Therefore, if the current is varied in any way, *e.g.* by altering the E. M. F. of the cell, there will be an induced E. M. F. whose value equals the rate of change of Li; and the greater L is, so much the greater is the induced E. M. F. The induced quantity of current equals the total change in the number of tubes of induction divided by the resistance of the circuit. If the applied E. M. F. is increased, so as to tend to increase the current, and thus increase the field of force, the induced current must be in the opposite direction; and as a consequence the current does not rise *instantly* to the value corresponding to the applied E. M. F. Similarly, if the applied E. M. F. is decreased, the induced current is in the direction of the original current; and so this does not decrease instantly to its final steady value. These facts may be expressed in a formula by writing the induced E. M. F. $= -\frac{\Delta N}{\Delta t}$, where the minus sign means that if ΔN is positive, the induced E. M. F. is negative, while if ΔN is negative, it is positive.

Particular cases of these changes are when the circuit is suddenly broken and when it is suddenly made, *e.g.* by removing one of the electrodes from the cell and by then plunging it in. In the former case, the current does not instantly fall to zero; there is the extra current, as shown by the spark, etc., as observed by Henry and Faraday. In the latter case, the current does not rise instantly to its fixed value. The time taken for these changes evidently varies directly as L; so that L measures what may be called the "inertia of the current."

When the circuit is broken, the energy of the magnetic field is no longer maintained by the cell, and it returns into the conductor, continuing the current until all the energy is consumed in heating the conductor. Then the current ceases. Similarly, when the circuit is closed, part of the energy furnished by the cell is spent in producing the magnetic field, and only a portion of it is available for producing the current in the conductor. It is not until the magnetic field is established, then, that all the energy supplied by the cell goes into maintaining the current. As a consequence it takes time to produce a steady current. These intervals of time required for a current to come practically to rest when a circuit is broken, or to be produced, are, as a rule, extremely short, a few millionths of a second; but if the circuit has a large value of L, *e.g.* if it is in the form of a solenoid inclosing a rod of iron, the time may be as great as a second. The energy of the current, *i.e.* of the magnetic field due to it, is thus in the surrounding medium.

2. *Mutual induction.* — If a circuit carrying a current is near another circuit also carrying a current, some of the tubes of force due to each current will pass through the other circuit. Thus if the currents in the two circuits are i_1 and i_2, the number of tubes due to the first current which pass through the circuit carrying the second one may be written $M_1 i_2$. (It must be noted that if the second circuit has n turns, the tubes pass through its current n times.) Similarly, the number of tubes due to the second current that pass through the first may be written $M_2 i_2$. It may be proved by the infinitesimal calculus that $M_1 = M_2$, and that this quantity is a constant for the two circuits, depending upon their shape, size, number of turns, and relative positions, and also upon the permeability of the surrounding medium; it is called the coefficient of "mutual induction," or the "mutual inductance." (The unit of induction is called the "Henry." It is the induction which a coil has if it is of such a size and shape that a variation in it of a current at the rate of one ampere per second produces an induced E. M. F. of one volt.)

The value of M may be calculated *provided all the tubes due to one current pass through all the coils of the other helix.* This condition may be secured practically in two ways: (1) by having one helix inclose the other and making them as nearly as possible of the same length and cross section, and then introducing a long rod of iron; (2) by winding the two helices on a ring of iron, as shown in the cut; for iron has such a great permeability compared with air that all the tubes practically stay

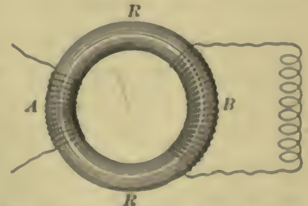


Fig. 419. — Diagram of transformer: two coils wound on an iron ring

in the iron and do not escape into the air. The former piece of apparatus constitutes an "induction coil" and has already been described; the latter is called a "transformer" and is used for many practical commercial purposes.

FIG. 419 *a*.—Section of a commercial transformer.

In the case of an induction coil, the intensity of the magnetic field at a point inside the primary far from the ends owing to a current i_1, is $4\pi N_1 i_1$, where N_1 is the number of turns per unit length. If there are N_2 turns per unit length in the secondary, there are $N_2 l$ turns in a length l; and if A is the area of the cross section of the primary, $4\pi\mu N_1 N_2 l A i_1$ tubes of induction thread these coils; so for a length l of the induction coil — near its middle part — $M = 4\pi\mu N_1 N_2 l A$. It is thus seen how the induced E. M. F. in the secondary is increased by introducing an iron core — so as to have μ large, and also by having N_1 and N_2 large.

Transformer. — In the case of the transformer the case is somewhat different, owing to the fact that it is assumed that none of the tubes escape into the air, and an exact calculation can be made of the effect of all the coils, not simply of those near the middle. If n_1 is the total number of turns in the primary, and n_2 that in the secondary, the coefficient of self-induction of the primary is proportional to n_1^2; and the coefficient of mutual induction of the two coils is proportional to $n_1 n_2$. So, if the current in the primary suddenly ceases, or if it is reversed in direction, the E. M. F. induced in the primary is proportional to n_1^2, and that induced in the secondary is proportional to $n_1 n_2$. As will be shown in a few pages, it is possible to construct a machine that produces an E. M. F. that is rapidly reversed in a continuous manner; this is called an "alternating" E. M. F. A particular case is one that may be written $E = E_1 \cos pt$, where E_1 is a constant. In this case the E. M. F. obeys a "sine curve," rising to a maximum value E_1, decreasing to 0 and then being reversed to $-E_1$, etc. If such an E. M. F. is applied to the primary of a transformer, it will produce a similar induced E. M. F. in the secondary, which may be written $E_2 \cos(pt-N)$. The phenomena are all then periodic, with a period $2\pi/p$, or a frequency $p/2\pi$. From what has been shown above $E_1 : E_2 = n_1^2 : n_1 n_2 = n_1 : n_2$. Thus, if $n_1 = 100\, n_2$, $E_1 = 100\, E_2$. Therefore by means of a transformer an alternating current with a large E. M. F. may be made to produce another alternating current with a small E. M. F. This plan is used in lighting houses with lamps rendered incandescent by means of alternating currents; the street current has a large E. M. F., but by means of a transformer the current produced in the house has a small E. M. F., which is not dangerous to life or property.

Since the energy supplied by the primary current is proportional to $E_1 i_1$, and that spent in maintaining the current in the secondary is proportional to $E_2 i_2$, it follows that, if we neglect any losses, $E_1 i_1 = E_2 i_2$, or the ratio of the current equals the inverse ratio of the electro-motive forces.

Large electro-motive forces are used in the street circuits because if such is the case, the conductors may be made of smaller sized wires than would be possible if the E. M. F. were small. The conductor must be able to furnish a large amount of energy, *i.e.* $E_1 i_1$, must be large; but the energy spent in heat is proportional to $i_1^2 R$, where R is the resistance; and, in order to have this small, i_1 must be as small as possible, and therefore E_1 is the factor of $E_1 i_1$, which is made large.

Dynamos. — The simplest case of a so-called dynamo is that of the "Gramme-ring" type. It consists of two parts, — the magnet and the armature. The magnet is one so made that the north and south poles come opposite each other, as is shown in the cut. In practice it is always an electro-magnet which is magnetized by an electric current. The armature consists of a soft iron ring which is made up of insulated iron wires bent into circles, and around which is wound a continuous copper wire carefully insulated from

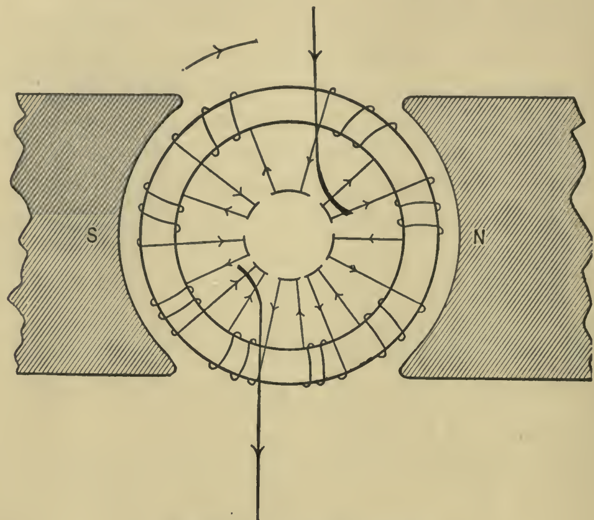


FIG. 420.—Gramme-ring dynamo.

the iron. The armature is rigidly fastened to a shaft perpendicular to its plane, and the shaft is placed perpendicular to the magnetic field of force between the poles of the magnet. If the shaft is revolved, currents will of course be induced in the coils of wire wound around the ring, because the field of magnetic force through them is constantly changing as the ring rotates. On the shaft of the armature is fastened what is called the "commutator," which consists of metal strips or "bars" along the shaft, each insulated from its neighbors, and resting across these bars are two so-called "brushes," which are metal strips, one on one side

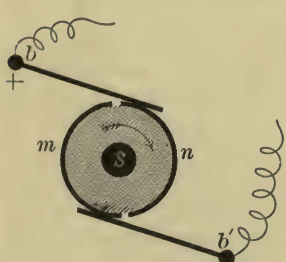


FIG. 420 *a*.—Simplest form of commutator.

of the commutator, the other diametrically opposite, so arranged as to touch opposite bars of the commutator at the same instant. These brushes are held stationary as the commutator revolves, and they are joined by a conductor through which a current is desired. Each of the bars of the commutator is connected by a wire to different points of the wire which is wound around the iron ring. Consequently, as the armature is turned by means of the shaft, the brushes are always joined to those portions of the wire around the iron ring which occupy in turn the same positions in the magnetic field.

The lines of force from the north pole of the magnet pass into the iron ring, and around through the ring to the side opposite the south pole, where they pass out and cross the air gap to the south pole. They do not pass across the ring, but are divided, as it were, into two sections, which crowd through at the top and bottom of the ring. Consequently, as the armature revolves, turns of wire which are horizontal have no field of force through them; but as they reach the top or bottom of their path, and so are placed vertical, there is a strong field through them. Currents are, therefore, induced in them, and their directions are easily deduced. If the magnetic poles are as shown in the cut, and the armature is being turned in the direction indicated, it is seen that the currents in all the coils on the ascending half of the ring are downward, while those in the other half are also downward. So, if the brushes are connected with those two bars of the commutator which are

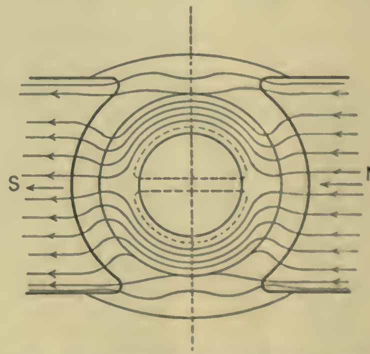


FIG. 421.—Lines of magnetic force between pole pieces of dynamo when there is no current in armature.

joined to the top and bottom coils of the armature, a current will be produced which will flow from one brush around to the other through the external wire, thus reaching the top coil, where it will divide into two branches which meet at the bottom coil; it then flows back to the other brush, etc. This process is perfectly continuous, and a steady current will be produced.

In practice, the brushes are not connected with the coils at top and bottom, but with those a little farther in advance, in the direction of rotation. This is because, as the current flows in the armature coils, it produces a magnetic field which so influences the field due to the magnets that the coil where the current tends to branch is no longer exactly at the top, but is a little in advance.

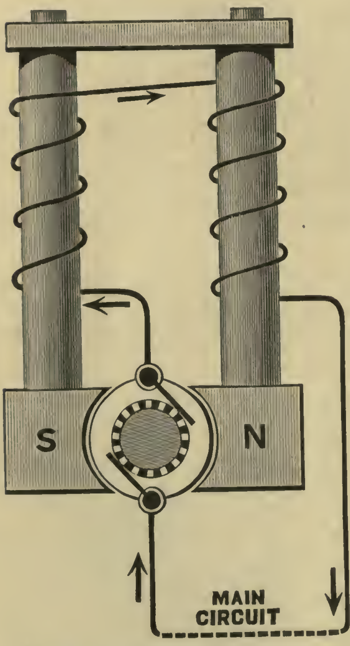


Fig. 422. — Diagram of a series dynamo.

If, instead of driving the shaft of this armature by means of some external power, and so producing a current, a current is sent through the armature from some other dynamo or battery, the armature will revolve and the shaft can be used to furnish power. This is, of course, the principle of the motor, as already mentioned.

In other types of dynamos, the armature consists of coils wound on a shuttle-shaped piece of iron, but the principle is the same as that of the ring dynamo just described. In order to avoid induced currents in the iron body of the armature, it is divided transversely into a great number of thin plates which are insulated from each other, but clamped together so as to form a solid frame.

The current to magnetize the field magnet is furnished in general by the dynamo itself, although in some cases a current is provided from a separate source. In some forms of machines the total current flows around the magnetizing coils, while in others the current is divided and only part is so used.

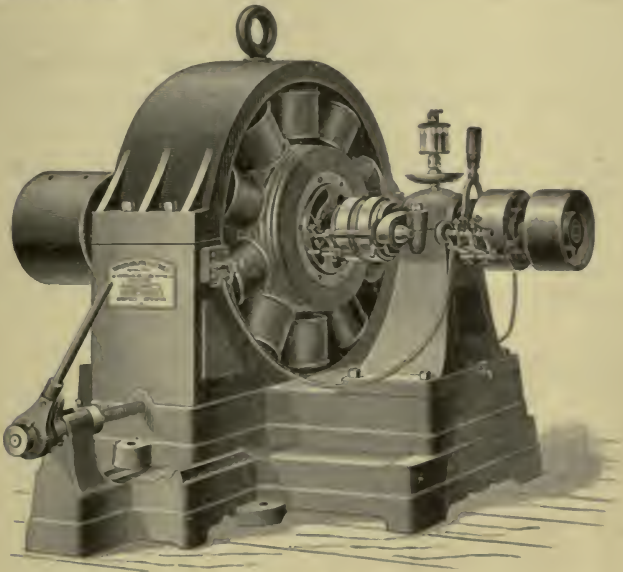


FIG. 423. — Alternating current dynamo.

The current produced by the dynamos just described is direct; but by simple changes the machines may be so arranged as to produce alternating currents. A common type of "alternating dynamo" is shown in the cut. It is seen that instead of having only two pole pieces, there are several arranged around a circle and so magnetized that consecutive ones are opposite in their nature, *i.e.* *N*, *S*, *N*, etc.; and that the armature consists of coils which are alternately wound in opposite

FIG. 423 *a*. — Diagram of winding in alternating current dynamo.

directions; so, as the armature is revolved, the induced electro-motive forces are in the same direction in them all, but they are reversed in direction as the coils pass from one magnetic field into the next. If there are n pole pieces, and if the armature makes m revolutions per minute, the E. M. F. will be reversed nm times each minute. On the armature shaft there are two conducting *rings*, which are insulated from each other, and to which are joined the terminals of the wire wound on the armature. On them rest the two brushes joined to the external circuit; so an alternating E. M. F. is applied to it. The pole pieces of the dynamo are magnetized by a separate direct-current dynamo.

Oscillatory Discharge of a Condenser. — It has been stated in the description of the discharge of a condenser that under certain conditions it is oscillatory. The reasons for this may now be discussed more fully. When the condenser is charged, the surrounding medium has a certain amount of electrostatic energy. If its two plates are joined by a conductor, a current will flow in it, and thus, if the resistance is small, so that the energy is not in the main spent in heating the conductor, a considerable amount of it will be consumed in producing a magnetic field. When all the electrostatic energy is thus exhausted, the electro-magnetic energy will flow back into the conductor, continuing the current in the same direction, and thus charging the condenser plates again, but in the opposite manner to their original charges. Finally, all the energy which has not gone into heating the conductor or to producing waves in the ether is again in the form of electrostatic energy; and the process is repeated in the opposite direction; etc. After a certain number of oscillations the energy is all exhausted, and everything comes to rest. It is evident that the greater the self-induction of the conductor joining the condenser plates, so much the more energy goes into the magnetic field for a given current.

A useful mechanical analogy is furnished by the vibrations of a spiral spring carrying a heavy body. When the spring is first stretched, the energy is entirely potential; then when the body passes through its position of equilibrium, the energy is entirely kinetic; as the body moves up, the spring is compressed, the strain being opposite to what it was originally, and the energy is again potential; etc. Further, the greater the mass of the moving body, so much the greater is the kinetic energy for a given velocity. As the vibrations continue, the amplitudes decrease gradually, owing to losses of energy by friction, and finally the motion ceases. The period of the motion is evidently increased if the mass is increased, and is decreased if the stiffness of the spring is increased.

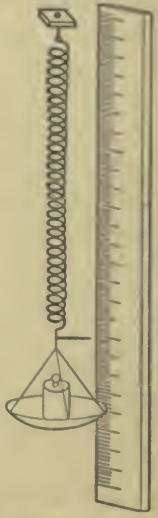


FIG. 424. — Vibrating spiral spring.

In the electrical oscillation, then, electrostatic energy corresponds to potential energy, and electro-magnetic to kinetic. The potential energy of the charged condenser is

$$W = \tfrac{1}{2}e(V_2 - V_1) = \tfrac{1}{2}\frac{e^2}{C};$$

and for a given charge is greater if C is small; so $\frac{1}{C}$ corresponds to the stiffness of the spring. The self-induction of the electrical conductor corresponds to the inertia of the vibrating matter. By means of the infinitesimal calculus it may be shown that if the resistance of the conductor is small, the period of oscillation of the discharge is given by the formula $T = 2\pi\sqrt{LC}$; so, if L is large, the period is great; and, if $\frac{1}{C}$ is large, the period is small, other things being equal.

Electrical Waves along Wires. — This mechanical analogy is illustrated still further by the phenomena of electrical waves along wires. If the potential of one end of a *long* conductor is first raised, then lowered, raised again, etc., or, in other words, if it varies periodically, it is found that the potential at all points along the conductor is not the same at any one instant, and that there is a train of waves of electrical potential passing down it. This is the process by which messages are transmitted across submarine cables, and by which telephone messages are sent. The velocity of these waves can be measured, and the wave length also. Experience and theory both agree in showing that the velocity of short waves along wires is the same as the velocity of light in the free ether, viz., 3×10^{10} cm. per second.

These waves may be compared with ordinary mechanical waves along a stretched string or rope. The self-induction of a unit length of the conductor corresponds to the mass per unit length of the string; the reciprocal of the capacity per unit length of the conductor, to the stiffness of the string; the resistance of the conductor, to the internal friction of the string.

Waves of all lengths decrease in amplitude as they advance along the conductor; but the long waves decrease least; so as a complex train of waves advances, it becomes more and more simple, because its shorter components vanish. If the self-induction of the conductor is increased sufficiently, not alone is the attenuation of all waves decreased, but it is the same for waves of all wave lengths; thus there is no distortion of the waves. In Pupin's system of constructing telephone lines, this condition is secured by inserting in the line, at regular intervals of every few miles, a helix whose self-induction is large. By this means it is possible to telephone with distinctness over intervals of a thousand miles or more.

Electrical Waves in the Ether. — Electrical oscillations produce waves in the surrounding ether, as has been already stated. These waves are identical in their properties with those which produce the sensation of light; and there are many ways by which they may be detected. They will produce oscillations in other conductors; and these may be made evident by sparks or by the heating effects. Again, if a number of small metallic particles are put loosely together, *e.g.* in a glass tube, they offer a great resistance to a current; but, when these long ether waves fall upon them, they cohere in such a manner as to have comparatively small resistance. If the particles are jarred slightly, after the waves pass, their resistance again increases. Such an apparatus is called a "coherer." It is evident that by introducing wires into the two ends of a coherer, and putting it in circuit with a battery and a galvanoscope, or a telegraph sounder, one can observe the passage of these "electric waves." There are many more methods, descriptions of which may be found in advanced text-books.

The papers of Henry and Faraday, on the subject of induced currents, have been reprinted in the Scientific Memoirs Series, Vols. XI and XII, New York, 1900.

CHAPTER XLIX

OTHER ELECTRICAL PHENOMENA

THERE are several phenomena, showing the connection between electricity and light, which should be mentioned.

Faraday Effect. — The fact discovered by Faraday that, if a beam of plane polarized light is passed through a strong magnetic field, parallel to the lines of force, the plane of polarization is rotated has already been discussed. (See page 564.) The direction of rotation follows the right-handed-screw law; so that, if the field of force is inside a solenoid, the direction of rotation is that of the current flowing in the coils. This rotation leads one to believe that associated with a line of magnetic force there is a rotational motion in the ether; so that any vibration in the ether at right angles to the line of force will have its direction changed. The energy of the magnetic field of force should then be considered as due to this motion; and it is thus seen why it is kinetic.

Hall Effect. — This rotational action of a magnetic field is shown also in what is called the "Hall effect," a phenomenon discovered in 1879 by E. H. Hall, now of Harvard University. If an electric current flows through a thin metallic film, it will so distribute itself that corresponding to any point on one edge of the film there is another on the opposite edge which has the same potential. Let A and B, in the cut, be two corresponding points; if they are joined by a wire which includes a galvanometer, no current will flow in it. If now this film is placed between the poles of a magnet, so that a magnetic field is produced perpendicular to the current sheet, a current will flow from A to B through the galvanometer, showing that they are no longer at the same potential. This proves that the lines of flow of the current

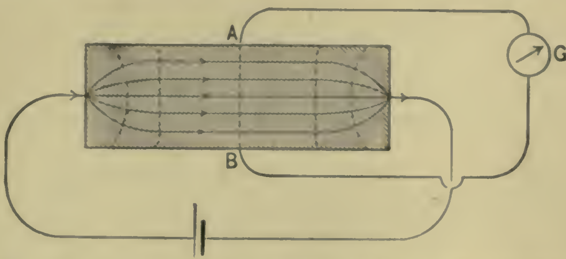


FIG. 425. — Lines of flow in a thin metal strip. Dotted lines are lines of constant potential.

have been rotated by the magnetic field. By moving one terminal of the wire slightly along the edge of the film, another point may be found for which there is again no current.

Before the magnetic field is produced, the lines of flow of the current are as shown in the cut; and lines of constant potential, at right angles to these, are also shown by dotted lines. After the field is produced, these lines are all rotated, as shown. The direction of the rotation is different in different films, and is reversed if that of the field or the current is reversed.

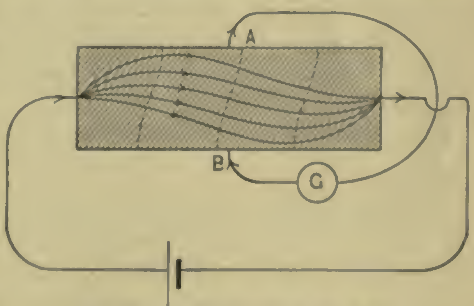


FIG. 426. — Diagram showing Hall effect.

This rotation may be described in a different way. Before the film is introduced into the magnetic field, the line of the electric force may be represented by $\overline{PQ}$; after the field is produced it is given by $\overline{PQ_1}$, if the rotation is in the direction indicated. This is equivalent to saying that the effect of the magnetic field is to cause a *transverse* E. M. F., $\overline{QQ_1}$.

Certain metal films, notably bismuth, have their resistance apparently increased when placed perpendicular to a magnetic field. This may be described by saying that the field causes a *longitudinal* E. M. F. opposite to that existing before; because, as shown by Ohm's law, $i = \frac{E}{R}$, a decrease in E will produce a decrease in i, exactly as if R were increased. This apparent increase in resistance may be explained if it is remembered that a current consists of the motion of charged particles; and such a particle in motion will have its direction of motion changed by a magnetic field, exactly as cathode rays are affected by a magnet. Therefore, owing to this deflection of the moving particles, the strength of the current *lengthwise* of the film is decreased. (It should be noted also that the flow of heat through a body is affected by a magnetic field in a manner exactly similar to this just described for a flow of electricity; and the reason may be explained as a consequence of the existence of the Thomson electro-motive forces.)

Fig. 427. — Transverse E. M. F.

Zeeman Effect. — A most interesting phenomenon was discovered by Zeeman in the year 1897, which has an important bearing upon theories of light. He observed that if a gas which was rendered luminous in any way was placed in a magnetic field, the spectrum "lines" were all affected; any single line was changed into two, three, or even more lines. These effects may be explained only by assuming that the ether waves emitted by the gas are due to the vibrations of charged particles. These are the electrons inside the atoms. By measuring the effect of a given magnetic field upon any particular radiation, it is possible to calculate the ratio of the charge of an electron to its mass; and it is found that this value is the same as for the negative particles emitted by radio-active substances and as for the cathode rays.

It is thus seen how intimately connected are the phenomena of electrical charges and radiations in the ether. Further, since a moving charge has an inertia of its own, quite apart from that of the material particle carrying it, it is not impossible that all the phenomena of matter may be explained as due to the motions of electrons.

The original memoirs of Faraday, Kerr, and Zeeman are given in Lewis, *The Effects of a Magnetic Field on Radiation*, Scientific Memoirs Series, New York, 1900.

Historical Sketch of Electricity

The first to use the word "electricity" or "electrical" was Gilbert in his treatise on Magnetism published in the year 1600. Many electrical phenomena, however, were known to the ancients. Thales, of Miletus, who lived about 600 B.C., is reported to have described the power which was produced in amber by friction and which enables it to attract bits of straw and other light bodies. The earliest description, however, of this property which is extant is that given by Theophrastus, 321 B.C.; and from the writings of Aristotle, Pliny, and others we learn that the ancient naturalists were aware of the electrical phenomena in the shocks of the torpedo. Gilbert was the first to make a sharp distinction between magnetic and electrical forces, and from his day on almost every natural philosopher performed many experiments in this most interesting field of science. The discovery of the fact that electrified bodies could repel each other is attributed to Von Guericke (about 1670). It was he also who invented the first frictional electrical machine, which was afterward improved by Ramsden, Canton, and others.

The discovery of the power of certain substances to conduct electricity is attributed to Stephen Gray, who died in 1736.

The fact that there are two kinds of electricity, positive and negative, was discovered by Du Fay in 1733; and the idea that in every process of electrification equal quantities of opposite kinds are produced is due to Symmer, 1759.

The phenomena of electrostatic induction and of charging by induction were first investigated by Canton, 1753, and Æpinus, 1759. We owe many important ideas to Benjamin Franklin. Among other things, he established the identity of atmospheric electricity as manifested in lightning, etc., with electricity as obtained by ordinary means. Many important phenomena were discovered by Cavendish in the last few years of the eighteenth century, but unfortunately his researches were not published for nearly one hundred years. Cavendish was the first to prove the law of the inverse square for electrical forces; to study the capacity of condensers; to discover the effect of introducing dielectrics other than air; to propose a law for an electric current, which is practically equivalent to Ohm's law; and to measure roughly the electrical resistances of many conductors.

The most important work done within recent years has been that of Michael Faraday, who was the first to recognize the importance of the surrounding medium in all questions dealing with electrical forces.

The varied phenomena dealing with the properties of electric currents and the names of those scientists associated with their discovery are given in the preceding pages.

BOOKS OF REFERENCE

Perkins. Outlines of Electricity and Magnetism. New York. 1896.
An elementary treatise, in which the phenomena are all explained in terms of tubes of induction, either electrostatic or magnetic.

Watson. A Text-book of Physics. London. 1899.
The sections on Electricity are particularly good.

J. J. Thomson. Elements of Electricity and Magnetism. London. 1895.
This is a text-book which treats the subject from a more mathematical standpoint than most other elementary books; but it may be consulted with advantage by any student.

INDEX