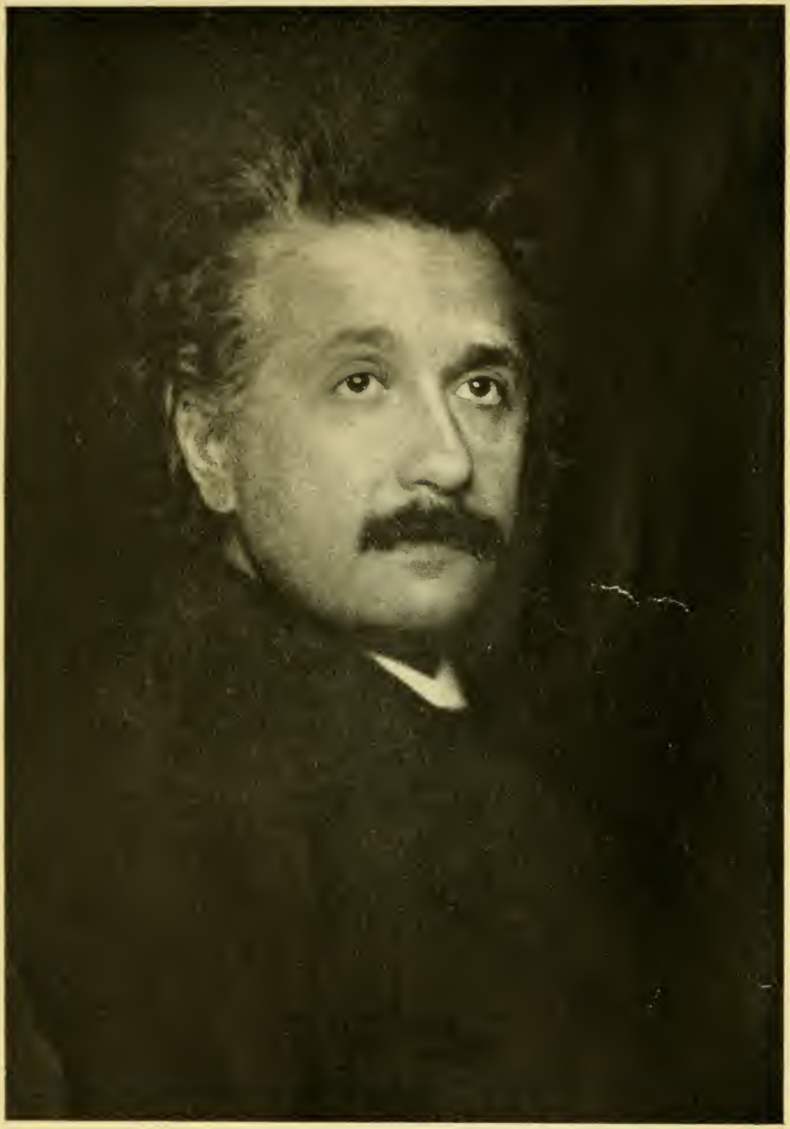
EINSTEIN'S THEORY OF RELATIVITY

MAX BORN

EINSTEIN'S THEORY OF RELATIVITY



A. Einstein.

EINSTEIN'S THEORY OF RELATIVITY

BY

MAX BORN

PROFESSOR OF THEORETICAL PHYSICS IN THE UNIVERSITY OF GÖTTINGEN

TRANSLATED BY

HENRY L. BROSE, M.A.

CHRIST CHURCH, OXFORD

WITH 135 DIAGRAMS AND A PORTRAIT

NEW YORK

E. P. DUTTON AND COMPANY

PUBLISHERS

Printed in Great Britain

FROM THE PREFACE TO THE FIRST EDITION

THIS book is an elaboration of certain lectures which were given last winter to a somewhat considerable audience. The difficulty which persons not conversant with mathematics and physics experience in understanding the theory of relativity seems to me to be due for the most part to the circumstance that they are not familiar with the fundamental conceptions and facts of physics, in particular of mechanics. During the lectures I therefore showed some quite simple qualitative experiments to serve as an introduction to such conceptions as velocity, acceleration, mass, force, intensity of field, and so forth. In my endeavour to find a similar means, adapted to book purposes, the semi-historical method of representation here chosen occurred to me, and I hope I have succeeded in avoiding the uninspiring method of the elementary text books of physics. But it must be emphasised that the historical arrangement has been selected only as a cloak which is to bring into stronger relief the outline of the main theme, the logical relationship. Having once started this process I found myself compelled to continue, and in this way my undertaking increased to the dimensions of this book.

The reader is assumed to have but little mathematical knowledge. I have attempted to avoid not only the higher mathematics but even the use of elementary functions, such as logarithms, trigonometrical functions, and so forth. Nevertheless, proportions, linear equations, and occasionally squares and square roots had to be introduced. I advise the reader who is troubled with the formulæ to pass them by on the first reading and to seek to arrive at an understanding of the mathematical symbols from the text itself. I have made abundant use of figures and graphical representations. Even those who are unpractised in the use of co-ordinates will learn to read the curves easily.

The philosophical questions to which the theory of relativity gives rise will only be touched on in this book. Nevertheless a definite logical point of view is maintained throughout. I believe I am right in asserting that this view agrees in the main with Einstein's own opinion. Moritz Schlick takes up a similar view in his valuable work "Allgemeine Erkenntislehre" (The General Theory of Knowledge).

Of the other books which I have used I should like to quote, above all, Ernst Mach's classical "Mechanics" (which has appeared in English), and then the very lucidly written volume by E. T. Whittaker, "A History of the Theories of Aether and Electricity" (London, Longmans, Green & Co., 1910), and the comprehensive account of the Theory of Relativity given by Hermann Weyl in his "Space, Time, Matter" (English translation published by Messrs. Methuen & Co., Ltd., 1922). Anyone who wishes to penetrate further into Einstein's doctrines must study the latter work. It is impossible to enumerate the countless books and essays from which I have drawn more or less directly. In conformity with the character of the book I have refrained from giving references.

MAX BORN

FRANKFURT ON THE MAIN
June, 1920

PREFACE
TO THE THIRD EDITION

APART from a number of minor alterations, this edition differs from its two predecessors in that the chapter on Einsteinian dynamics has been revised. Previously, in forming the acceleration, we did not distinguish sharply between time and proper time, and we used Minkowski's covariant force-vector in place of ordinary force; this of course increased the difficulty of understanding a chapter which was, from the outset, not easy. Dr. W. Pauli, jun., called my attention to a method of deriving the relativistic formula of mass proposed by Lewis and Tolman, which fitted in admirably with the scheme of this book, as it linked up with the conception of momentum in the same way as the account of mechanics here chosen. The chapter on Einsteinian dynamics was revised in conformity with this point of view; this also entailed some alterations in the manner of presenting ordinary mechanics. It is hoped that these changes will simplify the reading.

I should not like to lose this opportunity of thanking Dr. W. Pauli for his advice. His great work on the theory of relativity which has appeared as Article 19 in the fifth volume of the "Enzyklopädie der mathematischen Wissenschaften," which appeared recently, has been of great service to me. It is to be recommended foremost of all to those who wish to become intimately acquainted with the theory of relativity.

MAX BORN

GÖTTINGEN
6th March, 1922

CONTENTS

CHAPTER I

Geometry and Cosmology

CHAPTER II

The Fundamental Laws of Classical Mechanics

CHAPTER III

The Newtonian World-System

CHAPTER IV

The Fundamental Laws of Optics

CHAPTER V

The Fundamental Laws of Electrodynamics

CHAPTER VI

Einstein's Special Principle of Relativity

CHAPTER VII

Einstein's General Theory of Relativity

EINSTEIN'S THEORY OF RELATIVITY

INTRODUCTION

> Das schönste Glück des denkenden Menschen ist, das Erforschliche erforscht zu haben und das Unerforschliche ruhig zu verehren.
> —GOETHE.

THE world is not presented to the reflective mind as a finished product. The mind has to form its picture from innumerable sensations, experiences, communications, memories, perceptions. Hence there are probably not two thinking people whose picture of the world coincides in every respect.

When an idea in its main lines becomes the common property of large numbers of people, the movements of spirit that are called religious creeds, philosophic schools, and scientific systems arise ; they present the aspect of a chaos of opinions, of articles of faith, of convictions, that resist all efforts to disentangle them. It seems a sheer impossibility to find a thread that will guide us along a definite path through these widely ramified doctrines that branch off perchance to recombine at other points.

What place are we to assign to *Einstein's theory of relativity,* of which this book seeks to give an account ? Is it only a special part of physics or astronomy, interesting in itself but of no great importance for the development of the human spirit ? Or is it at least a symbol of a particular trend of thought characteristic of our times ? Or does it itself, indeed, signify a "world-view" (*Weltanschauung*) ? We shall be able to answer these questions with confidence only when we have become acquainted with the content of Einstein's doctrine. But we may be allowed to present here a point of view which, even if only roughly, classifies the totality of all world-views and ascribes to Einstein's theory a definite position within a uniform view of the world as a whole.

The world is composed of the ego and the non-ego, the inner world and the outer world. The relations of these two poles are the object of every religion, of every philosophy. But the part that each doctrine assigns to the ego in the world is different. *The importance of the ego* in the world-picture seems to me a measure according to which we may order confessions of faith, philosophic systems, world-views rooted in art or science, like pearls on a string. However enticing it may be to pursue this idea through the history of thought, we must not diverge too far from our theme, and we shall apply it only to that special realm of human thought to which Einstein's theory belongs—to natural science.

Natural science is situated at the end of this series, at the point where the ego, the subject, plays only an insignificant part; every advance in the mouldings of the conceptions of physics, astronomy, and chemistry denotes a further step towards the goal of excluding the ego. This does not, of course, deal with the act of knowing, which is bound to the subject, but with the finished picture of Nature, the basis of which is the idea that the ordinary world exists independently of and uninfluenced by the process of knowing.

The doors through which Nature imposes her presence on us are the senses. Their properties determine the extent of what is accessible to sensation or to intuitive perception. The further we go back in the history of the sciences, the more we find the natural picture of the world determined by the qualities of sense. Older physics was subdivided into mechanics, acoustics, optics, and theory of heat. We see the connexions with the organs of sense, the perceptions of motion, impressions of sound, light, and heat. Here the qualities of the subject are still decisive for the formation of conceptions. The development of the exact sciences leads along a definite path from this state to a goal which, even if far from being attained, yet lies clearly exposed before us: it is that of creating a picture of nature which, confined within no limits of possible perception or intuition, represents a pure structure of conception, conceived for the purpose of depicting the sum of all experiences uniformly and without inconsistencies.

Nowadays mechanical force is an abstraction which has only its name in common with the subjective feeling of force. Mechanical mass is no longer an attribute of tangible bodies but is also possessed by empty spaces filled only by ether radiation. The realm of audible tones has become a small province in the world of inaudible vibrations, distinguishable physically from these solely by the accidental property of the human ear which makes it react only to a definite interval of frequency numbers. Modern optics is a special chapter out of the theory of electricity and magnetism, and it treats of the electro-magnetic vibrations of all wave-lengths, passing from the shortest γ-rays of radioactive substances (having a wave-length of one hundred millionth of a millimetre) over the Röntgen rays, the ultraviolet, visible light, the infra-red, to the longest wireless (Hertzian) waves (which have a wave-length of many kilometres). In the flood of invisible light that is accessible to the mental eye of the physicist, the material eye is almost blind, so small is the interval of vibrations which it converts into sensations. The theory of heat, too, is but a special part of mechanics and electro-dynamics. Its fundamental conceptions of absolute temperature, of energy, and of entropy belong to the most subtle logical configurations of exact science, and, again, only their name still carries a memory of the subjective impression of heat or cold.

Inaudible tones, invisible light, imperceptible heat, these constitute the world of physics, cold and dead for him who wishes to experience living Nature, to grasp its relationships as a harmony, to marvel at her greatness in reverential awe. Goethe abhorred this motionless world. His bitter polemic against Newton, whom he regarded as the personification of a hostile view of Nature, proves that it was not merely a question of an isolated struggle between two investigators about individual questions of the theory of colour. Goethe is the representative of a world-view which is situated somewhere near the opposite end of the scale suggested above (constructed according to the relative importance of the ego), that is, the end opposite to that occupied by the world-picture of the exact sciences. The essence of poetry is inspiration, intuition, the visionary comprehension of the world of sense in symbolic forms. But the source of poetic power is experience, whether it be the clearly conscious perception of a sense-stimulus, or the powerfully represented idea of a relationship or connexion. What is logically formal and rational plays no part in the world-picture of such a type of gifted or indeed heaven-blessed spirit. The world as the sum of abstractions that are connected only indirectly with experience is a province that is foreign to it. Only what is directly presented to the ego, only what can be felt or at least represented as a possible experience is real to it and has significance for it. Thus to later readers, who survey the development of exact methods during the century after Goethe's time and who measure the power and significance of Goethe's works on the history of natural science by their fruits, these works appear as documents of a visionary mind, as the expression of a marvellous sense of one-ness with (*Einfühlung*) the natural relationships, but his *physical* assertions will seem to such a reader as misunderstandings and fruitless rebellions against a greater power, whose victory was assured even at that time.

Now in what does this power consist, what is its aim and device ?

It both takes and renounces. The exact sciences presume to aim at making *objective* statements, but they surrender their *absolute* validity. This formula is to bring out the following contrast.

All direct experiences lead to statements which must be allowed a certain degree of absolute validity. If I see a red flower, if I experience pleasure or pain, I experience events which it is meaningless to doubt. They are indubitably valid, but only for me. They are absolute, but they are subjective. All seekers after human knowledge aim at taking us out of the narrow circle of the ego, out of the still narrower circle of the ego that is bound to a moment of time, and at establishing common ground with other thinking creatures. It first establishes a link with the ego as it is at another moment, and then with other human beings or gods. All religions, philosophies, and sciences have been evolved for the purpose of expanding the ego to the wider community that " we " represent. But the ways of doing this are different. We are again confronted by the chaos of contradictory doctrines and opinions. Yet we no longer feel consternation, but order them according to the importance that is given to the subject in the mode of comprehension aimed at. This brings us back to our initial principle, for the completed process of comprehension is *the* world-picture. Here again the opposite poles appear.

The minds of one group do not wish to deny or to sacrifice the absolute, and they therefore remain clinging to the ego. They create a world-picture that can be produced by no systematic process, but by the unfathomable action of religious, artistic, or poetic means of expression in other souls. Here faith, pious ardour, love of brotherly communion, but often also fanaticism, intolerance, intellectual suppression hold sway.

The minds of the opposite group sacrifice the absolute. They discover—often with feelings of terror—the fact that inner experiences cannot be communicated. They no longer fight for what cannot be attained, and they resign themselves. But they wish to reach agreement at least in the sphere of the attainable. They therefore seek to discover what is common in their ego and in that of the other egos ; and the best that was there found was not the experiences of the soul itself, not sensations, ideas, or feelings, but abstract conceptions of the simplest kind—numbers, logical forms ; in short, the means of expression of the exact sciences. Here we are no longer concerned with what is absolute. The height of a cathedral does not, in the special sphere of the scientist, inspire reverence, but is measured in metres and centimetres. The course of life is no longer experienced as the running out of the sands of time, but is counted in years and days. Relative measures take the place of absolute impressions. And we get a world, narrow, one-sided, with sharp edges, bare of all sensual attraction, of all colours and tones. But in one respect it is superior to other world-pictures: the fact that it establishes a bridge from mind to mind cannot be doubted. It *is* possible to agree as to whether iron has a specific gravity greater than wood, whether water freezes more readily than mercury, whether Sirius is a planet or a star. There may be dissensions, it may sometimes seem as if a new doctrine upsets all the old "facts," yet he who has not shrunk from the effort of penetrating into the interior of this world will feel that the regions known with certainty are growing, and this feeling relieves the pain which arises from solitude of the spirit, and the bridge to kindred spirits becomes built.

We have endeavoured in this way to express the nature of scientific research, and now we can assign Einstein's theory of relativity to its category.

In the first place, it is a pure product of the striving after the liberation of the ego, after the release from sensation and perception. We spoke of the inaudible tones, of the invisible light, of physics. We find similar conditions in related sciences, in chemistry, which asserts the existence of certain (radioactive) substances, of which no one has ever perceived the smallest trace with any sense directly—or in astronomy, to which we refer below. These "extensions of the world," as we might call them, essentially concern sense-qualities. But everything takes place in the space and the time which was presented to mechanics by its founder, Newton. Now, Einstein's discovery is that this space and this time are still entirely embedded in the ego, and that the world-picture of natural science becomes more beautiful and grander if these fundamental conceptions are also subjected to relativization. Whereas, before, space was closely associated with the subjective, absolute sensation of extension, and time with that of the course of life, they are now purely conceptual schemes, just as far removed from direct perception as entities, as the whole region of wave-lengths of present-day optics is inaccessible to the sensation of light except for a very small interval. But just as in the latter case, the space and time of perception allow themselves to be ordered without giving rise to difficulties, into the system of physical conceptions. Thus an objectivation is attained, which has manifested its power by predicting natural phenomena in a truly wonderful way. We shall have to speak of this in detail in the sequel.

Thus the achievement of Einstein's theory is the relativization and objectivation of the conceptions of space and time. At the present day it is the final picture of the world as presented by science.

CHAPTER I

GEOMETRY AND COSMOLOGY

1. The Origin of the Art of Measuring Space and Time

THE physical problem presented by space and time is nothing more than the familiar task of fixing numerically a place and a point of time for every physical event, thus enabling us to single it out, as it were, from the chaos of the co-existence and succession of things.

The first problem of Man was to find his way about on the earth. Hence the art of measuring the earth (geodesy) became the source of the doctrine of space, which derived its name "geometry" from the Greek word for earth. From the very outset, however, the measure of time arose from the regular change of night and day, of the phases of the moon and of the seasons. These phenomena forced themselves on Man's attention and first moved him to direct his gaze to the stars, which were the source of the doctrine of the universe, *cosmology*. Astronomic science applied the teachings of geometry that had been tested on the earth to the heavenly regions, allowing distances and orbits to be defined. For this purpose it gave the inhabitants of the earth the celestial (astronomic) measure of time which taught Man to distinguish between Past, Present, and Future, and to assign to each thing its place in the realm of Time.

2. Units of Length and Time

The foundation of every space- and time-measurement is laid by fixing the unit. A datum of length, "so and so many metres," denotes the ratio of the length to be measured to the length of a metre. A time-datum of "so and so many seconds" denotes the ratio of the time to be measured to the duration of a second. Thus we are always dealing with ratio-numbers, relative data concerning the units. The latter themselves are to a high degree arbitrary, and are chosen for reasons of their being capable of easy reproduction, of being easily transportable, durable, and so forth.

In physics the measure of length is the *centimetre* (cm.), the hundredth part of a metre rod that is preserved in Paris. This was originally intended to bear a simple ratio to the circumference of the earth, namely, to be the ten-millionth part of a quadrant, but more recent measurements have disclosed that this is not accurately true.

The unit of time in physics is the *second* (sec.), which bears the well-known relation to the time of rotation of the earth on its axis.

3. Origin and Co-ordinate System

But if we wish not only to determine lengths and periods of time, but also to designate places and points of time, further conventions must be made. In the case of time, which we regard as a one-dimensional configuration, it is sufficient to specify an *origin* (or zero-point). Historians reckon dates by counting the years from the birth of Christ. Astronomers choose other origins or initial points, according to the objects of their researches ; these they call epochs. If the unit and the origin are fixed, every event may be singled out by assigning a number-datum to it.

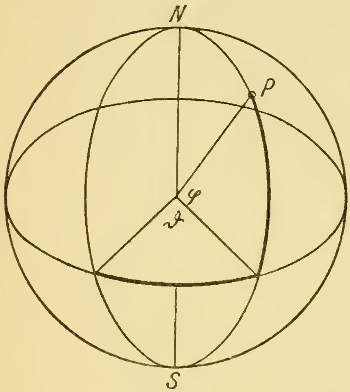


Fig. 1.

In geometry in the narrower sense, the determination of position on the earth, two data must be given to fix a point. To say " My house is in Baker Street," is not sufficient to fix it. The number of the house must also be given. In many American towns the streets themselves are numbered. The address No. 25, 13th Street, thus consists of two number-data. It is exactly what mathematicians call a " co-ordinate determination." The earth's surface is covered with a network of intersecting lines, which are numbered, or whose position is determined by a number, distance, or angle (made with respect to a fixed initial or zero-line).

Geographers generally use geographic longitude (east of Greenwich) and latitude (north or south) (Fig. 1). These determinations at the same time fix the zero-lines from which the co-ordinates are to be counted, namely, for geographical longitude the meridian of Greenwich, and for the latitude the equator. In investigations of plane geometry we generally use *rectilinear* (*Cartesian*) *co-ordinates* (Fig. 2), x, y, which signify the distances from two mutually perpendicular *co-ordinate axes;* or, occasionally, we also use *oblique co-ordinates* (Fig. 3), *polar co-ordinates* (Fig. 4), and others. When the

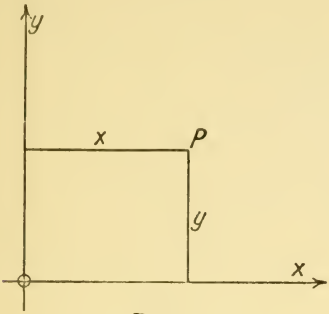


FIG. 2.

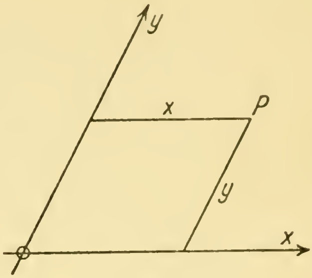


FIG. 3.

co-ordinate system has been specified, we can seek out each point or place if two numbers are given.

In precisely the same way we require three co-ordinates to fix points in space. It is simplest to choose mutually perpendicular rectilinear co-ordinates again ; we denote them by x, y, z (Fig. 5).

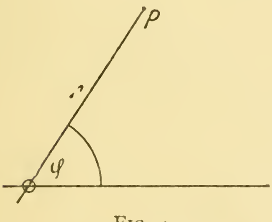


FIG. 4.

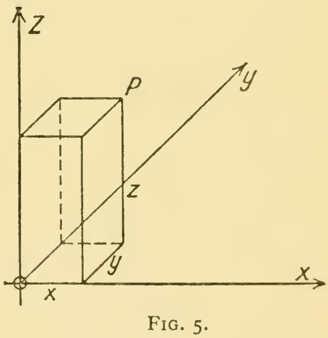


FIG. 5.

4. The Axioms of Geometry

Ancient geometry, regarded as a science, was less concerned with the question of determining positions on the earth's surface, than with determining the size and form of areas, figures in space, and the laws governing these questions. We see traces of the origin of this geometry in the art of surveying and of architecture. That is also the reason why it managed without the conception of co-ordinates. First and foremost, geometric theorems assert properties of things that are called points, straight lines, planes. In the classic canon of Greek geometry, the work of Euclid (300 B.C.), these things are not defined further but are only denominated or described. Thus we here recognize an appeal to intuition. You must already know what is a straight line if you wish to take up the study of geometry. Picture the edge of a house, or the stretched cable of your surveying instruments, form an abstraction from what is material and you will get your straight line. Next, laws are set up that are to hold between these configurations of abstraction, and it is to the credit of the Greeks to have made the great discovery that we need assume only a small number of these theorems to make all others come out of them correctly with logical inevitableness. These theorems, which are used as the foundation, are the *axioms*. Their correctness cannot be proved. They do not arise from logic but from other sources of knowledge. What these sources are has formed the subject of the theories of all the philosophies of the succeeding centuries. Scientific geometry itself, up to the end of the 18th century, accepted these axioms as given, and built up its purely deductive system of theorems on them.

We shall not be able to avoid discussing in detail the question of the meaning of the elementary configurations called point, straight line, and so forth, and the grounds of our knowledge of the geometric axioms. For the present, however, we shall adopt the standpoint that we are clear about these things, and we shall thus operate with the geometric conceptions in the way we learned (or should have learned) at school, and in the way numberless generations of people have done, without scruples. The intuitional truth of numerous geometric theorems, and the utility of the whole system in giving us bearings in our ordinary real world is to suffice for the present as our justification for using them.

5. The Ptolemaic System

To the eye the heavens appear as a more or less flat dome to which the stars are attached. But in the course of a day the whole dome turns about an axis whose position in the heavens is denoted by the pole-star. So long as this visual appearance was regarded as reality an application of geometry from the earth to astronomic space was superfluous, and was, as a matter of fact, not carried out. For lengths and distances measurable with earthly units were not present. To denote the positions of the stars only the apparent angle that the line of vision from the observer to the star formed with the horizon and with another appropriately chosen plane had to be known. At this stage of knowledge the earth's surface was considered at rest and was the eternal basis of the universe. The words " above " and " below " had an absolute meaning and when poetic fancy or philosophic speculation undertook to estimate the height of the heavens or the depth of Tartarus, the meaning of these terms required no word of elucidation. At this stage scientific concepts were still being drawn from the abundance of subjective data. The world-system called after Ptolemy (150 A.D.) is the scientific formulation of this mental attitude. It was already aware of a number of detailed facts concerning the motion of the sun, the moon, and the planets, and it had a considerable theoretical grasp of them, but it retained the notion that the earth is at rest and that the stars are revolving about it at immeasurable distances. Its orbits were determined as circles and epi-cycles according to the laws of earthly geometry, yet astronomic space was not actually through this subjected to geometry. For the orbits were fastened like rings to the crystal shells, which, arranged in strata, signified the heavens.

6. The Copernican System

It is known that Greek thinkers had already discovered the spherical shape of the earth and ventured to take the first steps from the geometric world-systems of Ptolemy to higher abstractions. But only long after Greek civilization and culture had died, did the peoples of other countries accept the *spherical* shape of the earth as a physical reality. This is the first truly great departure from the evidence of our eyes, and at the same time the first truly great step towards relativization. Again centuries have passed since that first turning-point, and what was at that time an unprecedented discovery has now become a platitude for school-children. This makes it difficult to convey an impression of what it signified to thinkers to see the conceptions " above " and " below " lose their absolute meaning, and to recognize the right of the inhabitants of the antipodes to call " above " in their regions what we call " below " in ours. But after the earth had once been circumnavigated all dissentient voices became silent. For this reason, too, the discovery of the sphericity of the earth offered no reason for strife between the objective and the subjective view of the world, between scientific research and the church. This strife broke out only after Copernicus (1543) displaced the earth from its central position in the universe and created the *helio-centric world-system.*

In itself the process of relativization was hardly advanced by this, but the importance of the discovery for the development of the human spirit consisted in the fact that the earth, mankind, the individual ego, became dethroned. The earth became a satellite of the sun and carried around in space the peoples swarming on it. Similar planets of equal importance accompany it in describing orbits about the sun. Man is no longer important in astronomy, except for himself. But still more, none of these amazing facts arise from ordinary observation (such as is the case with a circumnavigation of the globe), but from observations which were, for the time in question, very delicate and subtle, from different calculations of planetary orbits. The evidence was at any rate such as was neither accessible to all men nor of importance for everyday life. Ocular evidence, intuitive perceptions, sacred and pagan tradition alike speak against the new doctrine. In place of the visible disc of the sun it puts a ball of fire, gigantic beyond imagination; in place of the friendly lights of the heavens, similar balls of fire at inconceivable distances, or spheres like the earth, that reflect light from other sources; and all visible measures are to be regarded as deception, whereas immeasurable distances and incredible velocities are to represent the true state of affairs. Yet this new doctrine was destined to be victorious. For it drew its power from the burning wish of all thinking minds to comprehend all things of the material world, be they ever so unimportant for human existence, as a co-ordinate unity—to make them a permanent possession of the intellect and communicable to others. In this process, which constitutes the essence of scientific research, the human spirit neither hesitates nor fears to doubt the most striking facts of visual perception, and to declare them to be illusions, but prefers to resort to the most extreme abstractions rather than exclude from the scientific description of Nature one established fact, be it ever so insignificant. That, too, is why the church, at that time the carrier of the subjective world-view then dominant, had to persecute the followers of the Copernican doctrine, and that is why Galilei had to be brought before the inquisitorial tribunal as a heretic. It was not so much the contradictions to traditional dogmas as the changed attitude towards spiritual events that called this struggle into being. If the experience of the soul, the direct perception of things, was no longer to have significance in Nature, then religious experience might also one day be subjected to doubt. However far even the boldest thinkers of those times were removed from feelings of religious scepticism, the church scented the enemy.

The great relativizing achievement of Copernicus was the root of all the innumerable similar but lesser relativizations of growing natural science until the time when Einstein's discovery ranged itself as a worthy result alongside that of its great predecessor.

But now we must sketch in a few words the cosmos as mapped out by Copernicus.

We have first to remark that the conceptions and laws of earthly geometry can be directly applied to astronomic space. In place of the cycles of the Ptolemaic world, which were supposed to occur on surfaces, we now have real orbits in space, the planes of which may have different positions. The centre of the world-system is the sun. The planets describe their circles about it, and one of them is the earth, which rotates about its own axis, and the moon in its turn revolves in its orbit about the earth. But beyond, at enormous distances, the fixed stars are suns like our own, at rest in space. Copernicus' constructive achievement consists in the fact that with this assumption the heavens must exhibit all these phenomena which the traditional world-system was able to explain only by means of complicated and artificial hypotheses. The alternation of day and night, the seasons, the phenomena of the moon's phases, the winding planetary orbits, all these things become at one stroke clear, intelligible, and accessible to simple calculations.

7. The Elaboration of the Copernican Doctrine

The circular orbits of Copernicus soon no longer sufficed to account for the observations. The real orbits were evidently considerably more complicated. Now, an important point for the new view of the world was whether artificial constructions, such as the epicycles of the Ptolemaic system or an improvement in the calculations of the orbits could be successfully carried out without introducing complications. It was the immortal achievement of Kepler (1618) to discover the simple and striking laws of the planetary orbits, and hence to save the Copernican system at a critical period. The orbits are not, indeed, circles about the sun, but curves closely related to circles, namely, ellipses, in one focus of which the sun is situated. Just as this law determines the form of the orbits in a very simple manner, so the other two laws of Kepler determine the sizes of the orbits and the velocities with which they are traversed.

Kepler's contemporary, Galilei (1610), directed a telescope, which had just then been invented, at the heavens and discovered the moons of Jupiter. In them he recognized a microscopic model of the planetary system and saw Copernicus' ideas as optical realities. But it is Galilei's greater merit to have developed the principles of mechanics, the application of which to planetary orbits by Newton (1867) brought about the completion of the Copernican world-system.

Copernicus' circles and Kepler's ellipses are what modern science calls a *kinematic* or *phoronomic description* of the orbits, namely, a mathematical formulation of the motions which does not contain the causes and relationships that bring about these same motions. The causal expression of the laws of motion is the content of *dynamics* or *kinetics*, founded by Galilei. Newton has applied this doctrine to the motions of the heavenly bodies, and by interpreting Kepler's laws in a very ingenious way he introduced the causal conception of *mechanical force* into astronomy. Newton's law of gravitation proved its superiority over the older theories by accounting for all the deviations from Kepler's laws, the so-called perturbations of orbits, which refinements in the methods of observation had in the meantime brought to light.

This dynamical view of the phenomena of motion in astronomic space, however, at the same time demanded a more precise formulation of the assumptions concerning *space* and *time*. These axioms occur in Newton's work for the first time as explicit definitions. It is therefore justifiable to regard the theorems that held up to the advent of Einstein's theory as expressions of Newton's doctrine of space and time. To understand them it is absolutely necessary to have a clear survey of the fundamental laws of mechanics, and that, indeed, from a point of view which places the question of relativity in the foreground, a standpoint that is usually neglected in the elementary text-books. We shall therefore next have to discuss the simplest facts, definitions, and laws of mechanics.

CHAPTER II

THE FUNDAMENTAL LAWS OF CLASSICAL MECHANICS

1. Equilibrium and the Conception of Force

HISTORICALLY, mechanics took its start from the *doctrine of equilibrium* or *statics*; logically, too, the development from this point is the most natural one.

The fundamental conception of statics is *force*. It is derived from the subjective feeling of exertion experienced when we perform work with our bodies. Of two men he is the stronger who can lift the heavier stone or stretch the stiffer bow. This measure of force, with which Ulysses established his right among the suitors, and which, indeed, plays a great part in the stories of ancient heroes, already contains the germ of the objectivation of the subjective feeling of exertion. The next step was the choice of a unit of force and the measurement of all forces in terms of their ratios to the unit of force, that is, the relativization of the conception of force. *Weight*, being the most evident manifestation of force, and making all things tend downwards, offered the unit of force in a convenient form, namely, a piece of metal which was chosen as the unit of weight through some decree of the state or of the church. Nowadays it is an international congress that fixes the units. The unit of weight in technical matters is the weight of a definite piece of platinum in Paris. This unit, called the *gramme* (grm.) will be used in the sequel till otherwise stated. The instrument used to compare the weights of different bodies is the *balance*.

Two bodies have the same weight, or are equally heavy, when, on being placed in the two scales of the balance, they do not disturb its equilibrium. If we place two bodies found to be equally heavy in this manner in one pan of the balance, but, in the other, a body such that the equilibrium is again not disturbed, then this new body has twice the weight of either of the other two. Continuing in this way we get, starting from the unit of weight, a set of weights with the help of which the weight of every body may be conveniently determined.

It is not our task here to show how these means enabled man to find and interpret the simple laws of the statics of rigid bodies, such as the laws of levers. We here introduce only just those conceptions that are indispensable for an understanding of the theory of relativity.

Besides the forces that occur in man's body or in that of his domestic pets he encounters others, above all in the events that we nowadays call *elastic*. The force necessary to stretch a cross-bow or any other bow belongs to this category. Now, these can easily be compared with weights. If, for example, we wish to measure the force that is necessary to stretch a spiral spring a certain distance (Fig. 6), then we find by trial what weight must be suspended from it to effect equilibrium for just this extension. Then the force of the spring is equal to that of the weight, except that the former exerts a pull upwards but the latter downwards. The principle that action and reaction are equal and opposite in the condition of equilibrium has tacitly been applied.

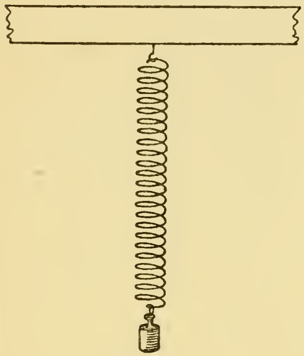


Fig. 6.

If such a state of equilibrium be disturbed by weakening or removing one of the forces, *motion* occurs. The raised weight falls when it is released by the hand supporting it and thus furnishing the reacting force. The arrow shoots forth when the archer releases the string of the stretched bow. Force tends to produce motion. This is the starting-point of *dynamics*, which seeks to discover the laws of this process.

2. The Study of Motions—Rectilinear Motion

It is first necessary to subject the conception of motion itself to analysis. The exact mathematical description of the motion of a point consists in specifying at what place relative to the previously selected co-ordinate system the point is situated from moment to moment. Mathematicians use formulæ to express this. We shall as much as possible avoid this method of representing laws and relationships, which is not familiar to everyone, and shall instead make use of a graphical method of representation. Let us illustrate this for the simplest case, the motion of a point in a straight line. Let the unit of length be the centimetre, as usual in physics, and let the moving point be at the distance $x = 1$ cm. from the zero point or origin at the moment at which we start our considerations and which we call the moment $t = 0$. In the course of 1 sec. suppose the point to have moved a distance of $\frac{1}{2}$ cm. to the right, so that for $t = 1$ the distance from the origin amounts to 1·5 cms. In the next second let it move by the same amount to $x = 2$ cms., and so forth. The following small table gives the distances x corresponding to the times t.

t	0	1	2	3	4	5	6	7	8 . . .
x	1	1·5	2	2·5	3	3·5	4	4·5	5 . . .

We see the same relationship pictured in the successive lines of Fig. 7, in which the moving point is indicated as a small circle on the scale of distances. Now, instead of drawing a number of small diagrams, one above the other, we may also

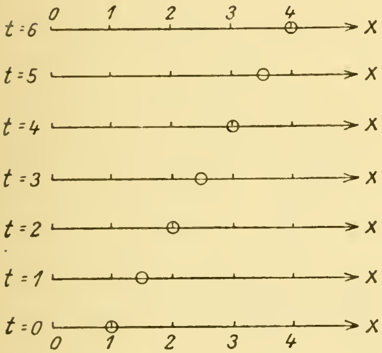


FIG. 7.

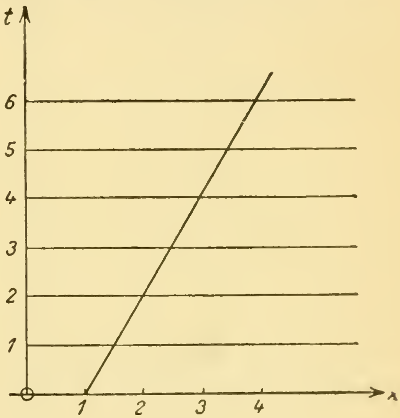


FIG. 8.

draw a single figure in which the x's and the t's occur as co-ordinates (Fig. 8). In addition, this has the advantage of allowing the place of the point to be depicted not only at the beginning of each full second but also at all intermediate times, We need only connect the positions marked in Fig. 7 by a continuous curve. In our case this is obviously a straight line. For the point advances equal distances in equal times; the co-ordinates x, t thus change in the same ratio (or proportionally), and it is evident that the graph of this law is a straight line. Such a motion is called *uniform*. The name *velocity* v of the motion designates the ratio of the path traversed to the time required in doing so, or in symbols :

$$v = \frac{x}{t} \quad . \quad . \quad . \quad . \quad . \quad (1)$$

In our example the point traverses $\frac{1}{2}$ cm. of path in each second. The velocity remains the same throughout and amounts to $\frac{1}{2}$ cm. per sec.

The unit of velocity is already fixed by this definition; it is the velocity which the point would have if it traversed 1 cm. per sec. It is said to be a *derived* unit, and, without introducing a new value, we call it cm. per sec. or cm./sec. To express that the measurement of velocities may be referred back to measurements of lengths and times in accordance with formula (1) we also say that velocity has the *dimensions* length divided by time, written thus: $[v] = \left[\frac{L}{T}\right]$ or $[L.T^{-1}]$. In the same way we assign definite dimensions to every quantity that allows itself to be built up of the fundamental quantities,

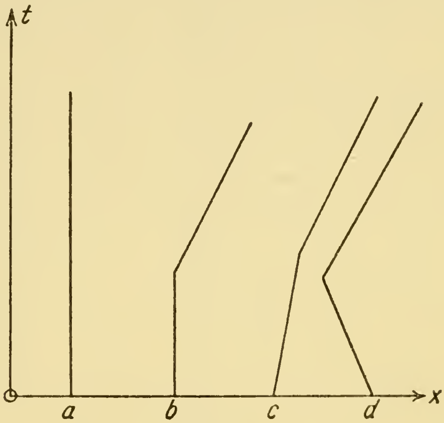


Fig. 9.

length l, time t, and weight G. When the latter are known the unit of the quantity may at once be expressed by means of those of length, time, and weight, say, cm., sec. and grm.

In the case of great velocities the path x traversed in the time t is great, thus the graph line has only a small inclination to the x-axis: the smaller the velocity, the steeper the graph. A point that is at rest has zero velocity and is represented in our diagram by a straight line parallel to the t-axis, for the points of this straight line have the same value of x for all times t (Fig. 9 *a*).

If a point is firstly at rest and then at a certain moment suddenly acquires a velocity and moves on with this velocity, we get as the graph a straight line one part of which is bent, the other being vertical (Fig. 9 *b*). Similarly broken lines represent the cases when a point that is initially moving uniformly for a while to the right or to the left suddenly changes its velocity (Figs. 9 *c* and 9 *d*).

If the velocity before the sudden change is v_1 (say, 3 cms. per sec.), and afterwards v_2 (say, 5 cms. per sec.), then the increase of velocity is $v_2 - v_1$ (that is, $5 - 3 = 2$ cms. per sec., added in each sec.). If v_2 is less than v_1 (say, $v_1 = 1$ cm. per sec.), then $v_2 - v_1$ is negative (namely, $1 - 3 = -2$ cms. per sec.), and this clearly denotes that the moving point is suddenly retarded.

If a point experiences a series of sudden changes of velocity then the graph of its motion is a succession of straight lines joined together (polygon) as in Fig. 10.

If the changes of velocity occur more and more frequently

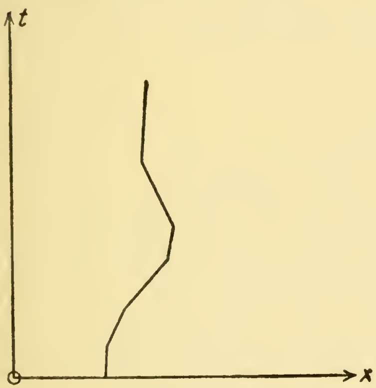


FIG. 10.

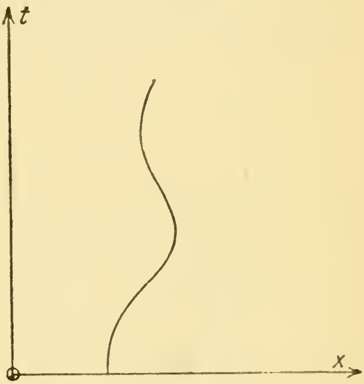


FIG. 11.

and are sufficiently small, the polygon will no longer be distinguishable from a curved line. It then represents a motion whose velocity is continually changing, that is, one which is non-uniform, accelerated or retarded (Fig. 11).

An exact measure of the velocity and its change, acceleration, can be obtained in this case only with the aid of the methods of infinitesimal geometry. It suffices for us to imagine the continuous curve replaced by a polygon whose straight sides represent uniform motions with definite velocities. The bends of the polygon, that is, the sudden changes of velocity, may be supposed to succeed each other at equal intervals of time, say, $t = \frac{1}{n}$ secs.

If, in addition these changes are equally great, the motion is said to be "uniformly accelerated." Let each such change of velocity have the value w, then if there are n per sec. the total change of velocity per sec. is

$$nw = \frac{w}{t} = b \quad . \quad . \quad . \quad . \quad (2)$$

Cf. Fig. 12.

Here $$x = \frac{1}{20}, \quad t = \frac{1}{10},$$

$$v_1 = \frac{1}{2}, \; v_2 = \frac{3}{2}, \; v_3 = \frac{5}{2}, \; . \; . \; ., \; w = 1,$$

$$b = \frac{w}{t} = 10.$$

This quantity b is the measure of the *acceleration*. Its dimensions are clearly $[b] = \left[\frac{V}{T}\right] = \left[\frac{L}{T^2}\right]$, and its unit is that

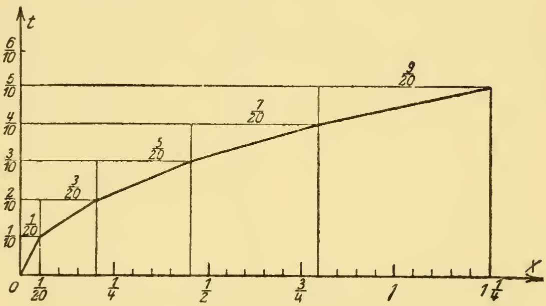


FIG. 12.

acceleration which causes unit velocity to increase by one unit in the unit of time, that is, referred to the physical system of measure cm./sec.

If we wish to know how far a movable point moves forward during a uniformly accelerated motion in any time t, we imagine the time t divided into n equal parts,* and suppose the point to receive a sudden increase of velocity w at the end of each small interval of time $\frac{t}{n}$. This little increase is connected with the acceleration b by the formula (2), if we replace the small interval of time t in it by $\frac{t}{n}$, thus : $w = b\frac{t}{n}$.

* In this case any *arbitrary* length of time t, and not as before, the unit of time, 1 sec., is divided into n parts.

Then the velocity
after the first interval of time is : $v_1 = w$,
,, ,, second ,, ,, $v_2 = v_1 + w = 2w$
,, ,, third ,, ,, $v_3 = v_2 + w = 3w$,
and so forth.

The point advances
after the first interval of time to : $x_1 = v_1 \frac{t}{n}$,
,, second ,, ,, $x_2 = x_1 + v_2 \frac{t}{n} = (v_1 + v_2) \frac{t}{n}$,
,, third ,, ,, $x_3 = x_2 + v_3 \frac{t}{n} = (v_1 + v_2 + v_3) \frac{t}{n}$,
and so forth.

After the nth interval of time, that is, at the end of the time t, the point will have arrived at

$$x = (v_1 + v_2 + \ldots \; v_n) \frac{t}{n}.$$

But $v_1 + v_2 + \ldots \; v_n = 1w + 2w + 3w + \ldots \; nw$
$= (1 + 2 + 3 + \ldots \; n)w.$

The sum of the numbers from 1 to n can be calculated quite simply by adding the first and the last ; the second and the second to last ; and so forth ; in each case we get for the sum of the two numbers $n + 1$, and altogether we have $\frac{n}{2}$ of such sums or pairs. Thus we get $1 + 2 + \ldots \; n = \frac{n}{2}(n + 1)$. If, further, we replace w by $b \frac{t}{n}$, we get

$$v_1 + v_2 + \ldots \; v_n = \frac{n}{2}(n + 1) \frac{bt}{n} = \frac{bt}{2}(n + 1),$$

thus $$x = \frac{bt}{2}(n + 1) \frac{t}{n} = \frac{bt^2}{2}\left(1 + \frac{1}{n}\right).$$

Here we may choose n to be as great as we please. Then $\frac{1}{n}$ becomes arbitrarily small and we get

$$x = \frac{1}{2} bt^2.$$

This signifies that in equal times the paths traversed are proportional to the squares of the times. If, for example, the acceleration $b = 10$ metres per sec., then the point traverses 5 metres in the first sec., $5 \cdot 2^2 = 5 \cdot 4 = 20$ metres in the second sec., $5 \cdot 3^2 = 45$ in the third sec., and so forth. This relationship is represented by a curved line, called a parabola, in the *xt* plane (Fig. 13). If we compare the figure with Fig. 12 we see how the polygon approximately represents the continuously curved parabola. In both figures the acceleration $b = 10$ has been chosen, and this determines the appearance of the curves, whereas the units of length and time are unessential.

We may also apply the conception of acceleration to non-uniformly accelerated motions, by using instead of 1 sec. a time of observation which is so small that, during it, the motion may be regarded as uniformly accelerated. The acceleration itself then becomes continuously variable.

All these definitions become rigorous and at the same time convenient to handle if the process of sub-division into small

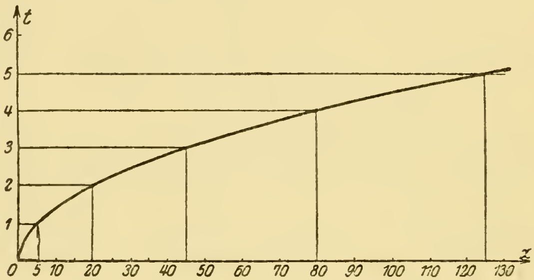


FIG. 13.

intervals, for which the quantity under consideration may be regarded as constant, is carefully studied. This leads us to the conception of the limiting value which forms the starting-point of the differential calculus. Historically, the doctrine of motion was actually the problem for the solution of which Newton invented the differential calculus and its converse, the integral calculus.

The theory of motion (kinematics, phoronomy) is the fore-runner of the proper mechanics of forces, or dynamics. It it evidently a sort of geometry of motion. As a matter of fact, in our graphical representation each motion is represented by a geometric configuration in the plane, with the co-ordinates *x*, *t*. In this we are concerned with more than a mere analogy. It is just the principle of relativity that attaches fundamental importance to the introduction of time as a co-ordinate in conjunction with the spatial dimensions.

3. Motion in a Plane

If we wish to study the motion of a point in a plane, our method of representation at once allows itself to be extended to this case. We take in the plane an *xy*-co-ordinate system and erect a *t*-axis perpendicular to it (Fig. 14). Then a straight line in the *xyt*-space corresponds to a rectilinear and uniform motion

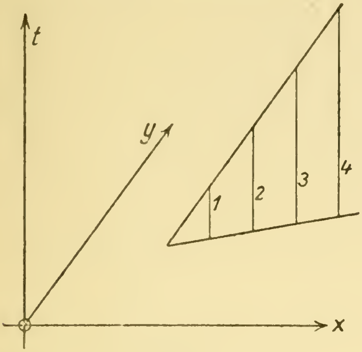


Fig. 14.

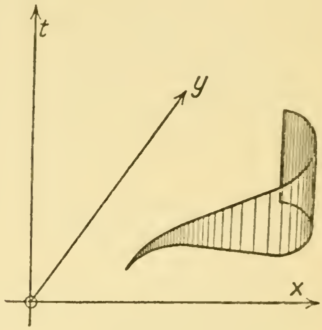


Fig. 15.

in the *xy*-plane. For if we project the points of the straight line that correspond to the points of time $t = 0, 1, 2, 3, \ldots$ on to the *xy*-plane, we see that the positional displacement takes place along a straight line and at equal intervals.

Every non-rectilinear but uniform motion is said to be *accelerated* even if, for example, a *curved* path is traversed with *constant* velocity. For in this case the *direction* of the velocity changes although its numerical value remains constant. An accelerated motion is represented in the *xyt*-plane (Fig. 15) by an arbitrary curve. The projection of this curve into the *xy*-plane is the orbit in the plane (or plane-orbit). The velocity and the acceleration are again calculated by supposing the curve replaced by a polygon closely wrapped round the curve. At each corner of this polygon not only the amount but also the direction of the velocity alters. A more exact analysis of the conception of acceleration would take us too far. It is sufficient to mention that it is best to project the graph of the moving point on to the co-ordinate axes *x*, *y*, and to follow out the rectilinear motion of these two points, or what is the same, the change

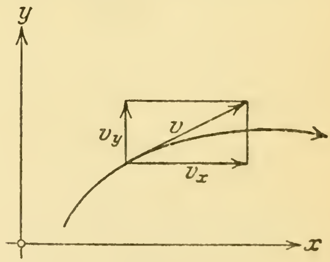


Fig. 16.

in time of the co-ordinates x, y. The conceptions defined for rectilinear motions as given above may now be applied to these projected motions. We thus get two *components of velocity* v_x, v_y, and two *components of acceleration* b_x, b_y, that together fix the velocity or the acceleration of the moving point at a given instant.

In the case of a plane motion (and also in one that occurs in space) velocity and acceleration are thus *directed magnitudes* (vectors). They have a definite *direction* and a definite *magnitude*. The latter can be calculated from the components. For example, we get the direction and magnitude of the velocity from the diagonal of the rectangle with the sides v_x and v_y (Fig. 16). Thus, by Pythagoras' theorem, its magnitude is

$$v = \sqrt{v_x^2 + v_y^2} \quad . \quad . \quad . \quad . \quad (3)$$

An exactly corresponding result holds for the acceleration.

4. Circular Motion

There is only one case which we wish to consider in greater detail, namely, the motion of a point in a circular orbit with

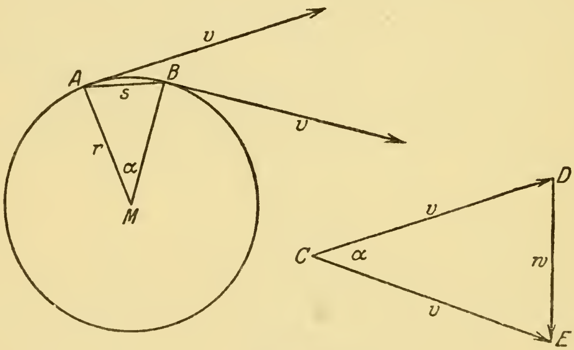


Fig. 17.

constant speed (Fig. 17). According to what was said above, it is an accelerated motion, since the direction of the velocity constantly alters. If the motion were unaccelerated the moving point would move forward from A in a straight line with the uniform velocity v. But in reality the point is to remain on the circle, and hence it must have a supplementary velocity or acceleration that is directed to the central point M. This is called the *centripetal acceleration*. It causes the velocity at a neighbouring point B, which is reached after a short interval t, to have a direction different from that at the point A. From a point c we next draw the velocities at A and B in a separate diagram (Fig. 17), paying due regard to their magnitude and direction. Their magnitude will be the same, namely v, since the circle is to be traversed with constant speed, but their direction is different. If we connect the end-points D and E of the two velocity lines, then the connecting line is clearly the supplementary velocity w, which transforms the first velocity state into the second. We thus get an isosceles triangle CED, having the base w and the sides v, and we at once see that the angle α at the vertex is equal to the angle subtended by the arc AB, which the point traverses, at the centre of the circle. For the velocities at A and B are perpendicular to the radii MA and MB, and hence include the same angle. Consequently the two isosceles triangles MAB and CDE are similar, and we get the proportion

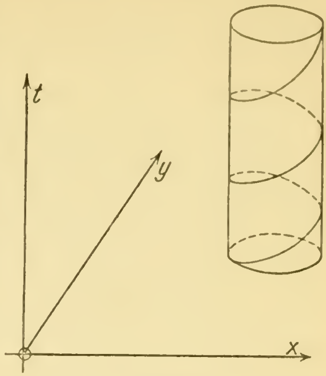


FIG. 18.

$$\frac{\mathrm{DE}}{\mathrm{CD}} = \frac{\mathrm{AB}}{\mathrm{MA}}$$

Now DE $= w$, CD $= v$, and further, MA is equal to the radius r of the circle, and AB is equal to the arc s except for a small error that can be made as small as we please by choosing the time-interval t sufficiently small.

Hence we have

$$\frac{w}{v} = \frac{s}{r} \text{ or } w = \frac{sv}{r}.$$

We now divide by t and notice that $\frac{s}{t} = v, \frac{w}{t} = b$. Hence the acceleration

$$b = \frac{v^2}{r}, \qquad (4)$$

that is, *the centripetal acceleration is equal to the square of the velocity in the circle divided by the radius.*

This theorem, as we shall see, is the basis of one of the first and most important empirical proofs of Newton's theory of gravitation.

Perhaps it is not superfluous to have a clear idea of what this uniform circular motion looks like in the graphical representation in the *xyt*-space. This is obviously produced by allowing the moving point to move upwards regularly parallel to the *t*-axis during the circular motion. We thus get a *helix* (screw line), which now represents the orbit *and* the course of the motion in time completely. In Fig. 18 it is drawn on the surface of a cylinder that has its base on the *xy*-plane.

5. Motion in Space

Our graphical method of representation fails for motions in space, for in this case we have three space co-ordinates *x*, *y*, *z*, and time has to be added as a fourth co-ordinate. But unfortunately our visual powers are confined to three-dimensional space. The symbolic language of mathematics must now lend us a helping hand. For the methods of *analytical geometry* allow us to treat the properties and relationships of spatial configurations as pure matters of calculation without requiring us to use our visual power or to sketch figures. Indeed, this process is much more powerful than geometric construction. Above all, it is not bound to the dimensional number three but is immediately applicable to spaces of four or more dimensions. In the language of mathematics the conception of a space of more than three dimensions is not at all mystical but is simply an abbreviated expression of the fact that we are dealing with things that allow themselves to be fully determined by more than three number data. Thus the position of a point at a given moment of time can be fixed only by specifying four number data, the three space-co-ordinates *x*, *y*, *z* and the time *t*. After we have learned to deal with the *xyt*-space as a means of depicting plane motion it will not be difficult also to regard the motions in three-dimensional space in the light of curves in the *xyzt*-space. This view of kinematics as geometry in a four-dimensional *xyzt*-space has the advantage of allowing us to apply the well-known laws of geometry to the study of motions. But it has a still deeper significance that will become clearly apparent in Einstein's theory. It will be shown that the conceptions space and time, which are contents of experience of quite different kinds, cannot be sharply differentiated at all as objects of physical measurement. If physics is to retain its maxim of recognizing as real only what is physically observable it must combine the conceptions space and time to a higher unity, namely, the four-dimensional *xyzt*-space. Minkowski called this the "*world*" (1908), by which he wished to express that the element of all order of real things is not place nor point of time but the "*event*" or the "*world-point*," that is, a place at a definite time. He called the graphical picture of a moving point "world-line," an expression that we shall continue to use in the sequel. Rectilinear uniform motion thus corresponds to a straight world-line, accelerated motion to one that is curved.

6. Dynamics—The Law of Inertia

After these preliminaries we revert to the question with which we started, namely, as to how forces generate motions.

The simplest case is that in which no forces are present at all. A body at rest will then certainly not be set into motion. The ancients had already made this discovery, but, above this, they also believed the converse to be true, namely, that wherever there is motion there must be forces that maintain them. This view at once leads to difficulties if we reflect on why a stone or a spear that has been thrown continues to move when it has been released from the hand. It is clearly the latter that has set it into motion, but its influence is at an end so soon as the motion has actually begun. Ancient thinkers were much troubled in trying to discover what forces actually maintain the motion of the thrown stone. Galilei was the first to find the right point of view. He observed that it is a prejudiced idea to assume that wherever there is motion there must always be force. Rather it must be asked what quantitative property of motion has a regular relationship with force, whether it be the place of the moving body, its velocity, its acceleration, or some composite quantity dependent on all of these. No amount of reflection will allow us to evolve an answer to these questions by philosophy. We must address ourselves directly to nature. The question which she gives is, firstly, that force has an influence in effecting *changes* of velocity. No force is necessary to maintain a motion in which the magnitude and the direction of the velocity remain unaltered. And conversely, where there are no forces, the magnitude and direction of the velocity remain unaltered ; thus a body which is at rest remains at rest, and one that is moving uniformly and rectilinearly continues to move uniformly and rectilinearly.

This *law of inertia* (or of persistence) is by no means so obvious as its simple expression might lead us to surmise. For in our experience we do not know of bodies that are really withdrawn from all influences from without, and if we use our imaginations to picture how they travel on in their solitary rectilinear paths with constant velocity throughout astronomic space, we are at once confronted with the problem of the absolutely straight path in space absolutely at rest, with which we shall have to deal in detail later on. For the present, then, we shall interpret the law of inertia in the restricted sense in which Galilei meant it.

Let us picture to ourselves a smooth exactly horizontal table on which a smooth sphere is resting. This is kept pressed against the table by its own weight, but we ascertain that it requires no appreciable force to move the sphere quite slowly on the table. Evidently there is no force acting in a horizontal direction on the sphere, otherwise it would not itself remain at rest at *any* point on the table.

But if we now give the sphere a velocity it will continue to move in a straight line and will lose only very little of its speed. This retardation was called a secondary effect by Galilei, and it is to be ascribed to the friction of the table and the air, even if the frictional forces cannot be proved to be present by the statical methods with which we started. It is just this depth of vision, which correctly differentiates what is essential in an occurrence from disturbing subsidiary effects, that characterizes the great investigator.

The law of inertia is at any rate confirmed for motion on the table. It has been established that in the absence of forces the velocity remains constant in direction and magnitude.

Consequently the forces will be associated with the change of velocity, the acceleration. In what way they are associated can again be decided only by experiment.

7. Impulses

We have presented the acceleration of a non-uniform motion as a limiting case of sudden changes of velocity of brief uniform motions. Hence we shall first have to enquire how a single sudden change of velocity is produced by the application of a force. For this a force must act for only a short time; it is then what we call an *impulse* or a *blow*. The result of such a blow depends not only on the magnitude of the force but also on the duration of the action, even if this is very short. We therefore define the intensity of a blow or impulse as follows:

n impulses J, each of which consists of the force K acting during the time $t = \frac{1}{n}$ secs., will, if they follow each other without appreciable pauses, have exactly the same effect as if the force K were to continue to act throughout the whole second. Thus we should have

$$nJ = \frac{1}{t}J = K,$$

or,

$$J = \frac{1}{n}K = tK \quad . \quad . \quad . \quad . \quad (5)$$

To visualize this, let us imagine a weight placed on one side of a lever having equal arms (such as a balance), and suppose a hammer to tap very quickly and evenly on the other side with blows just powerful enough to preserve equilibrium except for inappreciable fluctuations (Fig. 19). It is clear that we may tap more weakly but more often, or more strongly and less often, so long as the intensity J of the blow multiplied by the number of blows n, or divided by the time t required by each blow, always remains exactly equal to the weight K. This " Impulse Balance " enables us to measure the intensity of blows even when we cannot ascertain the duration and the force of each one singly. We need only find the force K that keeps equilibrium with n such equal blows per second (disregarding the inappreciable trembling of the arms), then the magnitude of each blow is the nth part of K.

The dimensions of impulse are $[J] = [T . G]$, where G denotes weight.

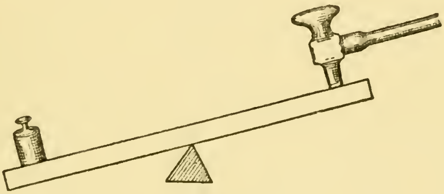


FIG. 19.

8. The Law of Impulses

We again consider the sphere on the table and study the action of impulses on it. To do this we require a hammer that may be swung, say, about a horizontal axis. Firstly, we calibrate the power of the blows of our hammer for each length of drop by means of our " impulse balance." Then we allow it to impinge against the sphere resting on the table and observe the velocity that it acquires through the blow by measuring how many cms. it rolls in 1 sec. (Fig. 20). The result is very simple.

The more powerful the blow the greater the velocity, the relation being such that twice the blow imparts twice the velocity, three times the blow three times the velocity, and so forth, that is, the velocity and the blow bear a constant ratio to each other (they are proportional).

This is the fundamental law of dynamics, the so-called *law of impulse* (*or momentum*) for the simple case when a body is set into motion from rest. If the sphere already has a velocity initially, the blow will increase or decrease it according as it strikes the sphere in the rear or in the front. By a strong counter-blow it is possible to reverse the direction of motion of the sphere.

The *law of impulse* then states that *the sudden changes of velocity of the body are in the ratio of the impulses or blows that produce them.* The velocities are here considered as positive or negative according to their direction.

9. Mass

Hitherto we have dealt with a single sphere. We shall now perform the same impulse experiment with spheres of different kinds, say, of different size or of different material, some being solid and others hollow. Suppose all these spheres to be set into motion by exactly equal blows or impulses. Experiment shows that they then acquire quite different

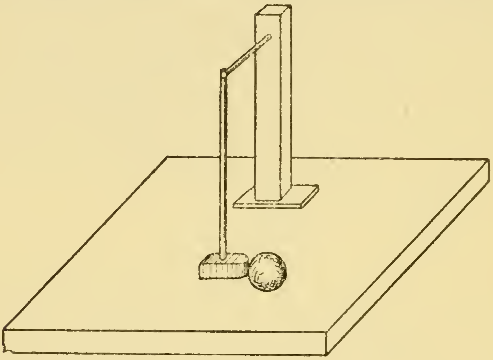


Fig. 20.

velocities, and, indeed, it is at once observed that light spheres are made to travel at great speed, but heavy ones roll away only slowly. Thus we find a relationship with weight, into which we shall enter into detail later, for it is one of the empirical foundations of the general theory of relativity. But here, on the contrary, we wish to bring out clearly and prominently that from the abstract point of view the fact that various spheres acquire various velocities after equally strong impacts has nothing to do with weight. Weight acts downwards and produces the pressure of the sphere on the table, but exerts no horizontal force. We now find that *one* sphere opposes greater resistance to the blow than *another;* if the former is at the same time the heavier, then this is a new fact of experience, but does not from the point of view here adopted allow itself to be deduced from the conception of weight. What we establish is a difference of resistance of the spheres to impacts. We call it *inertial resistance*, and measure it as the ratio of the impulse or impact J to the velocity v generated. The name *mass* has been chosen for this ratio, and it is denoted by the symbol m. Thus we set

$$m = \frac{J}{v} \quad . \quad . \quad . \quad . \quad . \quad (6)$$

This formula states that for one and the same body an increase of the impulse J calls up a greater velocity v in such a way that their ratio has always the same value m. When mass has been defined in this way its unit can no longer be chosen at pleasure, because the units of velocity and of impulse have already been fixed. Rather, mass has the dimensions

$$[M] = \left[\frac{T^2G}{L}\right]$$

and its unit in the ordinary system of measures is sec.^{2}grm./cm.

In ordinary language the word mass denotes something like amount of substance or quantity of matter, these conceptions themselves being no further defined. The concept of substance, as a category of the understanding, is counted among those things that are directly given, i.e. are immediate data. In physics, however—as we must very strongly emphasize—the word mass has *no* meaning other than that given by formula (6). It is the measure of the resistance to changes of velocity.

We may write the law of impulses more generally thus :

$$mw = J \quad . \quad . \quad . \quad . \quad . \quad (7)$$

It determines the *change* of velocity w that a body in motion experiences as the result of an impulse J.

The formula is often interpreted too as follows :—

The given impulsive force J of the hammer is transferred to the movable sphere. The hammer "loses" the impulse J, and this impulse reappears in the motion of the sphere to the same extent mw. This impulsive force carries the sphere along, and when the latter itself impinges on another body, it, in its turn, gives the latter a blow or impulse, and thereby loses just as much impulse as the other body gains. For example, if the bodies of mass m_1 and m_2 impinge against each other rectilinearly (that is, whilst moving in the same straight line), then the impulsive forces which they exert on each other are always equal and opposite, that is, $J_1 = - J_2$, or their sum is zero :

$$J_1 + J_2 = m_1w_1 + m_2w_2 = 0 \quad . \quad . \quad . \quad (8)$$

From this it follows that

$$w_2 = -\frac{m_1}{m_2} w_1,$$

that is, when one sphere loses velocity (w_1 negative), the other gains velocity (w_2 positive), and *vice versa.*

If we introduce the velocities of the two spheres before and after the impact, namely, v_1, v_1' for the first sphere, and v_2, v_2' for the second, then the changes of velocity are

$$w_1 = v_1' - v_1 \qquad w_2 = v_2' - v_2$$

and we may also write the equation (8) thus :

$$m_1(v_1' - v_1) + m_2(v_2' - v_2) = 0$$

If we then collect all the quantities referring to the motion *before* the impact on the one side, and all those referring to the motion *after* the impact on the other, we get

$$m_1v_1 + m_2v_2 = m_1v_1' + m_2v_2' \quad . \quad . \quad . \quad (9)$$

and this equation may be interpreted as follows :

To bring a body of mass m from a state of rest into one in which it has the velocity v we require the impulse mv ; it then carries this impulse along with it. Thus the total impulse carried along by the two spheres before the impact is $m_1v_1 + m_2v_2$. The equation (9) then states that this total impulse is not changed as a result of the impact. This is the *law of conservation of impulse or momentum.*

10. Force and Acceleration

Before pursuing further the striking parallelism between mass and weight we shall apply the laws so far established to the case of forces that act continuously. Unfortunately, again, the theorems can be set up rigorously only with the aid of the methods of the infinitesimal calculus, yet the following considerations may serve to give an approximate idea of the relationships involved.

A force that acts continuously generates a motion whose velocity alters continuously. We now suppose the force replaced by a rapid succession of blows or impulses. Then at each blow the velocity will suffer a sudden change and a world-line that is bent many times, as in Fig. 10, will result, and which will fold closely around the true, uniformly curved, world-line and will be able to be used in place of the latter in the calculations. Now if w blows per sec. replace the force K, then by (5) each of them has the value $J = \frac{1}{n} K$ or $= tK$, where t is the short interval occupied by each blow. At each impulse a change of velocity w occurs which, according to (7), is determined by $mw = J = tK$, or $m\frac{w}{t} = K$. But, by (2), $\frac{w}{t} = b$, thus we get

$$mb = K \quad . \quad . \quad . \quad . \quad . \quad (10)$$

This is *the law of motion of dynamics* for forces that act continuously. It states in words that *a force produces an acceleration that is proportioned to it; the constant ratio* K : *b is the mass.*

We may give this law still a different form which is advantageous for many purposes, in particular for the generalization that is necessary in the dynamics of Einstein (see VI, 7, p. 221). For if the velocity v alters by the amount w, then the impulse carried along by the moving body, namely, $J = mv$, alters by mw. Thus we have $mb = \frac{mw}{t}$, the change of the impulse carried along in the time t required to effect it. Accordingly we may express the fundamental law expressed in formula (10) thus:

If a force K acts on a body, then the impulse $J = mv$ carried along by the body changes in such a way that its change per unit of time is equal to the force K.

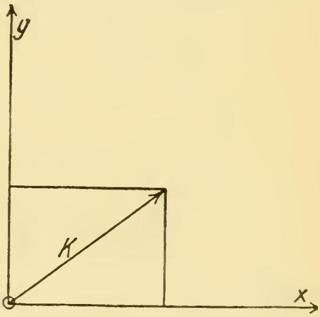


FIG. 21.

Expressed in this form the law holds only for motions which take place in a straight line and in which the force acts in the same straight line. If this is not the case, that is, if the force acts obliquely to the momentary direction of motion the law must be generalized somewhat. Let us suppose the force drawn as an arrow which is then projected on to three mutually perpendicular directions, say, the co-ordinate axes. In Fig. 21 the case is represented in which the force acts in the xy-plane, and its projections on the x- and the y-axis have been drawn. Let us imagine the moving point projected on the axes in the same way. Then each of the points of projection executes a motion on its axis of projection. The law of motions then states that the accelerations of these motions of projection bear the relation $mb = K$ to the corresponding components of force. But we shall not enter more closely into these mathematical generalizations, which involve no new conceptions.

11. Example—Elastic Vibrations

As an example of the relation between force, mass, and acceleration we consider a body that can execute vibrations under the action of elastic forces. We take, say, a straight broad steel spring and fasten it at one end so that it lies horizontally in its position of rest (and does *not* hang downwards). It bears a sphere at the other end (Fig. 22). The sphere can then swing to and fro in the horizontal plane (that of the page). Gravity has no influence on its motion, which depends only on the elastic force of the spring. When the displacements are small the sphere moves almost in a straight line. Let its direction of motion be the x-axis.

If we set the sphere into motion, it executes a periodic vibration, the nature of which we can make clear to ourselves as follows: If we displace the sphere slightly out of the position of equilibrium with our hands, we experience the restoring force of the spring. If we let the sphere go, this force imparts to it an acceleration, which causes it to return to the mean position with increasing velocity. In this process the restoring force, and hence also the acceleration, continuously decreases, and becomes zero when passing through the mean position itself, for here the sphere is in equilibrium and no accelerative force acts on it. At the place, therefore, at which the velocity is greatest, the acceleration is least. In consequence of its inertia the sphere passes rapidly through the position of equilibrium, and then the force of the spring begins to retard it and applies a brake, as it were, to the motion. When the original deflection has been attained on the other side the velocity has decreased to zero and the force has reached its highest value. At the same time the acceleration has reached its greatest value in reversing the direction of the velocity at this moment. From this point onwards it repeats the process in the reverse sense.

If we next replace the sphere by another of different mass we see that the character of the motion remains the same but the time of a vibration is changed. When the mass is greater the motion is retarded, and the acceleration becomes less; a decrease of mass increases the number of vibrations per sec.

In many cases the restoring force K may be assumed to be exactly proportional to the deflection x. The course of the motion may then be represented geometrically as follows: Consider a movable point P on the circumference of a circle of radius a, which is being traversed uniformly ν times per sec. by P. It then traverses the circumference, which is $2\pi a$

(where $\pi = 3{\cdot}14159 \ldots$), in the time $T = \frac{1}{\nu}$ secs., thus its velocity is $\frac{s}{t} = \frac{2\pi a}{T} = 2\pi a\nu$.

Let us now take the centre O of the circle as the origin of a rectangular set of co-ordinates in which P has the co-ordinates x, y. Then the point of projection A of the point P on the x-axis will move to and fro during the motion just like the mass fastened to the spring. *This point* A *is to represent the vibrating mass.* If P moves forward along a small arc s, then A moves along the x-axis a small distance ξ, and we have $v = \frac{\xi}{t}$ as the velocity of A. Fig. 23 now shows that the displacements ξ

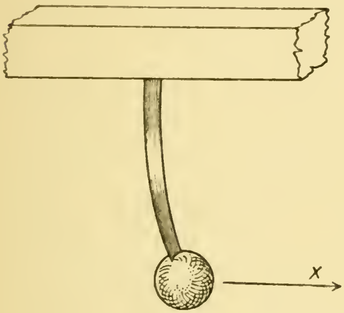


FIG. 22.

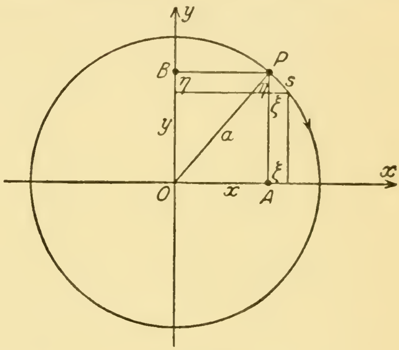


FIG. 23.

and s are the side and the hypotenuse of a small right-angled triangle, which is clearly similar to the large right-angled triangle OAP. Hence we have the proportion

$$\frac{\xi}{s} = \frac{y}{a} \text{ or } \xi = \frac{sy}{a}.$$

Hence the velocity of A becomes

$$v = \frac{\xi}{t} = \frac{s}{t} \cdot \frac{y}{a} = 2\pi\nu y.$$

Now, the point of projection B of the point P executes exactly the same pendulum motion on the y-axis. During the small displacement s of P the point B moves forward a distance η, and just as for ξ, we have

$$\frac{\eta}{s} = \frac{x}{a} \text{ or } \eta = s\frac{x}{a}.$$

This change η of y corresponds to a change in the velocity $v = 2\pi\nu y$ of the point A which is given by

$$w = 2\pi\nu\eta = 2\pi\nu s\frac{x}{a},$$

and hence to an acceleration of A,

$$b = \frac{w}{t} = 2\pi\nu\frac{s}{t}\cdot\frac{x}{a} = (2\pi\nu)^2 x.$$

The acceleration in this vibrational motion of the point A is thus actually at every moment proportional to the deflection x. We get for the force

$$\mathrm{K} = mb = m(2\pi\nu)^2 x \quad . \quad . \quad . \quad . \quad (11)$$

By measuring the force corresponding to a deflection x and by counting the vibrations we can thus determine the mass m of the spring pendulum.

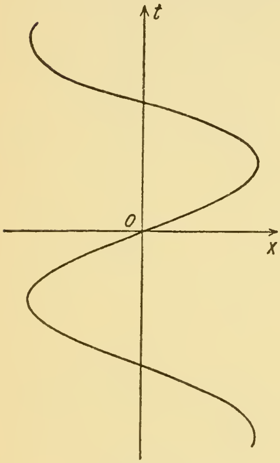


Fig. 24.

The picture of the world-line of such a vibration is clearly a wave-line in the xt-plane, if x is the direction of vibration (Fig. 24). In the figure it has been assumed that at the time $t = 0$ the sphere is moving through the middle position $x = 0$ towards the right. We see that whenever the sphere passes through the t-axis, that is, for $x = 0$, the direction of the curve is most inclined to the x-axis, and this indicates the greatest velocity. Hence the curve is not curved at this point, and the change of velocity or the acceleration is zero. The opposite is true of those points that correspond to the extreme deflections.

12. Weight and Mass

At the beginning when we introduced the conception of mass, we observed immediately that mass and weight exhibit a remarkable parallelism. Heavy bodies offer a stronger resistance to an accelerating force than light bodies. Is this, then, an exact law? As a matter of fact, it is. To have the facts quite clear, let us again consider the experiment of setting into motion spheres on a smooth horizontal table by means of impacts or impulses. We take two spheres A and B, of which B is twice as heavy as A, that is, on the impulse balance B exactly counterpoises two bodies each exactly like A. We next apply equal blows to A and B on the table and observe the velocity attained. We find that A rolls away twice as quickly as B.

Thus the sphere B, which is twice as heavy as A, opposes a change of velocity exactly twice as strongly as A. We may also express this as follows: Bodies having twice the mass have twice the weight; or, more generally, the masses m are in the ratio of the weights G. The ratio of the weight to the mass is a perfectly definite number. It is denoted by g, and we write

$$\frac{G}{m} = g \text{ or } G = mg \qquad . \quad . \quad . \quad (12)$$

Of course, the experiment used to illustrate the law is very rough.* But there are many other phenomena that prove the same fact; above all, there is the observed phenomenon that all bodies fall equally fast. It is hereby assumed, of course, that no forces other than gravity exert an influence on the motion. This means that the experiment must be carried out *in vacuo* so that the resistance of the air may be eliminated. For purposes of demonstration an inclined plane (Fig. 25) is found suitable, on which two spheres, similar in appearance but of different weight, are allowed to roll down. It is observed that they reach the bottom exactly simultaneously.

The weight is the driving force; the mass determines the resistance. If they are proportional to each other, then a heavy body will indeed be driven forward more strongly than a lighter one, but to balance this it resists the impelling force more strongly, and the result is that the heavy and the light body roll or fall down equally fast. We also see this from our formulæ. For if in (10) we replace the force by the weight G, and assume the latter, by (12), proportional to the mass, we get

$$mb = G = mg,$$

that is,

$$b = g \quad . \quad . \quad . \quad . \quad . \quad (13)$$

Thus all bodies have one and the same acceleration vertically downwards, if they move under the influence of gravity alone, whether they fall freely or are thrown. The quantity g, the acceleration due to gravity, has the value

$$g = 981 \text{ cm./sec.}^2 (\text{or } 32 \text{ ft./sec.}^2).$$

* For example, we have neglected the circumstance that in producing the *rotation* of the rolling sphere a resistance must also be overcome which depends on the distribution of mass in the interior of the sphere (the moment of inertia).

The most searching experiments for testing this law may be carried out successfully with the aid of simple pendulums with very fine threads. Newton even in his time noticed that the times of swing are always the same for the same length of pendulum, whatever the composition of the sphere of the pendulum. The process of vibration is exactly the same as that described above for the elastic pendulum, except that now it is not a steel spring but gravity that pulls back the sphere. We must imagine the force of gravity acting on the sphere to be resolved into two components, one acting in the direction of the continuation of the thread, and keeping it stretched, the other acting in the direction of motion and being the driving force that acts on the sphere or bob.

Fig. 26 exhibits the bob at the deflection x. We see at once

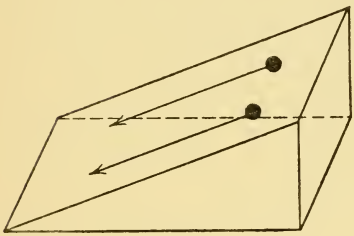


FIG. 25.

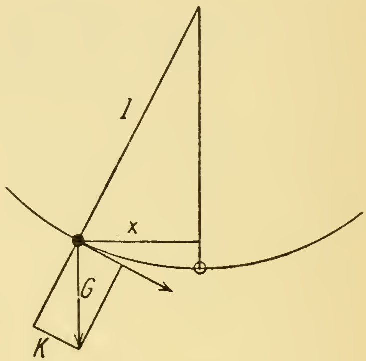


FIG. 26.

the two similar right-angled triangles, the sides of which are in the same proportion :

$$\frac{K}{x} = \frac{G}{l}.$$

Accordingly, formula (11) gives for two pendulums, the bobs of which are G_1 and G_2, respectively :

$$(2\pi\nu)^2 m_1 = \frac{G_1}{l}, \qquad (2\pi\nu)^2 m_2 = \frac{G_2}{l},$$

thus

$$\frac{G_1}{m_1} = \frac{G_2}{m_2} = (2\pi\nu)^2 l,$$

that is, the ratio of the weight to the mass is the same for both pendulums. We called this ratio g in formula (12). Hence we get the equation

$$g = (2\pi\nu)^2 l, \quad . \quad . \quad . \quad . \quad (14)$$

from which we see that g may be determined by measuring the length l of the pendulum and the vibration number ν.

The law of the proportionality of weight to mass is often expressed as follows:

gravitational and inertial mass are equal.

Here gravitational mass simply signifies the weight divided by g, and the proper mass is distinguished by prefixing the word "inertial."

The fact that this law holds very exactly was already known to Newton. Nowadays it has been confirmed by the most delicate measurements known in physics, which were carried out by Eötvös (1890). Hence we are completely justified in using the balance to compare not only weights but also masses.

One might now imagine that such a law is firmly embedded in the foundations of mechanics. Yet this is by no means the case, as is shown by our account, which follows fairly closely the ideas contained in classical mechanics. Rather, it is attached, as a sort of curiosity, somewhat loosely to the fabric of the other laws. Probably it has been a source of wonder to many, but no one suspected or sought any deeper relationship that might be wrapt in it. For there are many kinds of forces that can act on a mass. Why should there not be one that is exactly proportional to the mass? A question to which no answer is *expected* will receive none. And so the matter rested for centuries. This was possible only because the successes of the mechanics of Galilei and Newton were overwhelming It controlled not only the motional events on the earth but also those of the stars, and showed itself to be a trustworthy foundation for the whole realm of the exact sciences. For in the middle of the nineteenth century it was looked on as the object of research to interpret all physical events as mechanical events in the sense of the Newtonian doctrine. And thus in building up their stately edifice physicists forgot to ascertain whether the basis was strong enough to support the whole. Einstein was the first to recognize the importance of the law of equality of inertial and gravitational mass for the foundations of the physical sciences.

13. Analytical Mechanics

The problem of analytical mechanics is to find from the law of motion

$$mb = \mathrm{K}$$

the motion when the forces K are given. The formula itself gives us only the acceleration, that is, the change of velocity.

To get from the latter the velocity, and from this again the varying position of the moving point, is a problem of the integral calculus that may be very difficult if the force alters in a complicated way with the place and the time. An idea of the nature of the problem is given by our derivation of the change of position in a uniformly accelerated motion along a straight line (p. 20). The motion is already more complicated when it is in a plane and due to the action of a constant force of definite direction, as in the case of a motion due to falling or to a throw. Here, too, we may substitute as an approximation for the continuous course of the motion one consisting of a series of uniform motions, each of which is transformed into the next by means of impulses. We again call to mind our table and agree that the sphere rolling on it is to receive a blow of the same size and direction after the same short interval t (Fig. 27). Now, if the sphere starts off from the point O with

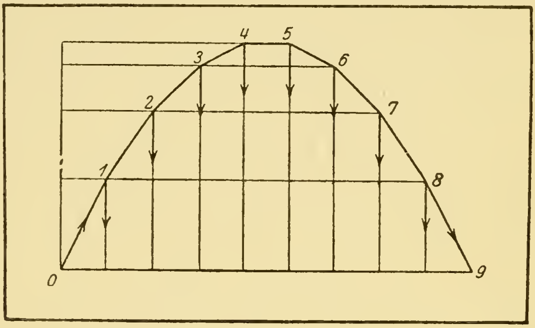


FIG. 27.

an arbitrary initial velocity it arrives after t secs. at a point 1 where the first blow strikes it. From this point it pursues its course in another direction with a different velocity for t secs. until at a point 2 it is struck by the second blow, which again deflects it, and so forth. Each individual deflection may be determined from the law of impulses. Accordingly, we may draw the whole motion, and we see that the initial point, the initial direction, and the initial velocity completely determine the subsequent course of the motion. This jerky motion gives us a rough picture of the motion of a sphere on an inclined plane. The graph coincides the more closely with the event, which is continuous in reality, the smaller we choose the time interval between the blows.

What is here achieved by direct construction is usually done in the case of forces acting continuously by means of the integral calculus. In this case, too, the point of departure and the initial velocity remain quite arbitrary as regards magnitude and direction. But if these are given, the further course of the motion is fully determined. Thus one and the same law of force may produce an infinity of motions according to the choice of the initial conditions. Thus the enormous number of motions due to falling or to throws depends on the same law of force, of gravity that acts vertically downwards.

In mechanical problems we are usually concerned with the motion not of *one* body but of several that exert forces on one another. The forces are then not themselves given but depend for their part on the unknown motion. It is easy to understand that the problem of determining the motions of several bodies by calculation becomes highly complicated.

14. The Law of Energy

But there is a law which makes these problems much simpler and affords a survey of the motion. It is the *law of the conservation of energy*, which has become of very great impor-

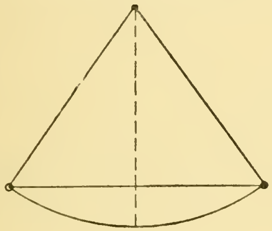


Fig. 28.

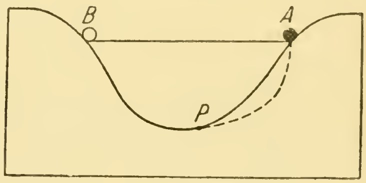


Fig. 29.

tance for the development of the physical sciences. We cannot, of course, enunciate it generally here nor prove it. We shall only seek to know its content from simple examples.

A pendulum which is released after the bob has been raised to a certain point rises on the opposite side of the mean position to the same height—except for a small error caused by friction and the resistance of the air (Fig. 28). If we replace the circular orbit by some other by allowing the sphere to run on rails as in a "toy railway" (Fig. 29), then the same result holds: the sphere always rises to the same height as that from which it started.

From this it easily follows that the velocity that the sphere has at any point P of its path depends only on the depth of this point P below the initial point A. To see this we imagine the piece AP of the orbit changed, the rest PB remaining unaltered. Now, if the sphere were to arrive at P along the one orbit from A with a velocity different from that with which it arrives along the other, then in its further course from P to B it would not in each case exactly reach its goal B. For, to achieve this, a uniquely determinate velocity is clearly necessary at P. Consequently the velocity at P does not depend on the form of the piece of orbit traversed, and since P is an arbitrary point, this result holds generally. Hence the velocity v must be determined by the height of fall h alone. The truth of the law depends on the circumstance that the path (the rails) as such opposes no resistance to the motion, that is, exerts no force on the sphere in its direction of motion, but receives only its perpendicular pressure. If the rails are not present, we have the case of a body falling freely or of one that has been thrown, and the same result holds: the velocity at each point depends only on the height of fall.

This fact may not only be established experimentally but may also be derived from our laws of motion. We hereby also get the form of the law that regulates the dependence of the velocity on the height. We assert that it states the following:

Let x be the path fallen through, measured upwards (Fig 30), v the velocity, m the mass, and G the weight of the body. Then the quantity

$$E = \frac{m}{2}v^2 + Gx \quad . \quad . \quad . \quad . \quad (15)$$

has the same value during the whole process of falling.

To prove this we first suppose E to stand for any arbitrary quantity that depends on the motion and hence alters from moment to moment. Let E alter by the amount e in a small interval of time t, then we shall call the ratio $\frac{e}{t}$ the rate of change of E, and, exactly as before in defining the orbital velocity v and the acceleration b, we suppose that the time interval t may be taken as small as we please. If the quantity E does not change in the course of time, then its rate of change is, of course, zero, and *vice versa*. We next form the change of the above expression E in the time t. During this time the height of fall x decreases by vt, and the velocity v increases by $w = bt$. Hence after the time t the value of E becomes

$$E' = \frac{m}{2}(v + w)^2 + G\,(x - vt).$$

Now, $$(v + w)^2 = v^2 + w^2 + 2vw.$$

This states that the square erected over v and w, joined together in the same straight line, may be resolved into a square having the side v, one having the side w, and two equal rectangles having the sides v and w (Fig. 31).

Hence we get

$$E' = \frac{m}{2}\, v^2 + \frac{m}{2}\, w^2 + mvw + Gx - Gvt.$$

If we deduct the old value of E from this, we get as the change in value

$$e = E' - E = \frac{m}{2}\, w^2 + mvw - Gvt,$$

or, since $w = bt$,

$$e = \frac{m}{2}\, b^2t^2 + mvbt - Gvt.$$

Hence the rate of change becomes

$$\frac{e}{t} = \frac{m}{2}\, b^2t + mvb - Gv.$$

FIG. 30.

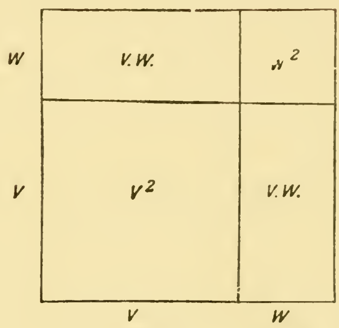


FIG. 31.

The term involving t may be neglected since it can be made vanishingly small by making the time interval vanishingly small. Hence we get finally for the rate of change of E,

$$\frac{e}{t} = v(mb - G).$$

But on the basis of the laws of mechanics this expression has the value zero, for by (13) we have $mb = mg = G$. Hence we have proved that the quantity E (15) remains unchanged with time. If the initial point and the initial velocity of the motion are given, that is, the values of x and v for $t = o$, then the expression E, according to (15), acquires a definite value. It then retains this value during the whole motion.

From this it follows that if the body rises, that is, if x increases, v must decrease, and *vice versa*. Either of the two terms of the expression E can increase only at the expense of the other. The first term is characteristic of the state of velocity of the body, the second, of the height that it has attained against the force of gravitation. We have special names for these terms.

$T = \frac{m}{2}v^2$ is called the *vis viva* or *kinetic energy*.

$U = Gx$ is called the *capacity for work* or the *potential energy*.

Their sum,
$$T + U = E, \quad . \quad . \quad . \quad . \quad (16)$$
is called simply the *mechanical energy* of the body; and the law which states that it remains constant during the motion of the body is called the *law of conservation of energy*.

The dimensions of energy are $[E] = [GL]$. Its unit is grm. cm.

The name capacity for doing work is of course derived from the work done by the human body in lifting a weight. According to the law of conservation of energy this work becomes transformed into kinetic energy in the process of falling. If, on the other hand, we give a body kinetic energy by throwing it upwards, this energy changes into potential energy or capacity for doing work.

Exactly the same as has been described for falling motions, holds in the widest sense for systems composed of any number of bodies, so long as two conditions are fulfilled, namely:

1. External influences must not be involved, that is, the system must be self-contained or isolated.
2. Phenomena must not occur in which mechanical energy is transformed into heat, electrical tension, chemical affinity, and such like.

If these are fulfilled the law that
$$E = T + U$$
always remains constant holds true, the kinetic energy depending on the velocities, the potential energy on the positions of the moving bodies.

In the mechanics of the heavenly bodies this ideal case is realized very perfectly. Here the ideal dynamics of which we have developed the principles is valid.

But on the earth this is by no means the case. Every motion is subject to friction, whereby its energy is transformed into heat. The machines by means of which we produce motion, transform thermal, chemical, electric, and magnetic forces into mechanical forces, and hence the law of energy in its narrow mechanical form does not apply. But it may always be maintained in an extended form. If we call the heat energy Q, the chemical energy C, the electro-magnetic energy W, and so forth, then the law that for closed systems the sum

$$E = T + U + Q + C + W \ldots \quad (17)$$

is always constant holds.

It would lead us too far to pursue the discovery and logical evolution of this fact by Robert Mayer, Joule (1842) and Helmholtz (1847), or to investigate how the non-mechanical forms of energy are determined quantitatively. But we shall use the conception of energy later when we speak of the intimate relationship that the theory of relativity has disclosed between mass and energy.

15. Dynamical Units of Force and Mass

The validity of the process by which we have derived the fundamental laws of mechanics is, in a certain sense, restricted to the surface of our table and its immediate neighbourhood. For we have abstracted our conceptions and laws from experiments in a very limited space, in the laboratory. The advantage of this is that we need not trouble our heads about the assumption concerning space and time. The rectilinear motions with which the law of inertia deals may be copied on the table with a ruler. Apparatus and clocks are assumed to be available for measuring the orbits and the motions.

Our next concern will be to step out of the narrow confines of our rooms into the wider world of astronomic space. The first stage will be a " voyage round the world " which idiomatic usage applies to the small globe of the earth. We shall pose the question : do all the laws of mechanics set up apply just as much in a laboratory in Buenos Aires or in Capetown as here ?

Yes, they do, with one exception, namely, the value of the gravitational acceleration g. We have seen that this can be measured exactly by observations of pendulums. It has been found that one and the same pendulum swings somewhat more slowly at the equator than in the more southerly or more northerly regions. Fewer vibrations occur in the course of a day, that is, in the course of one rotation of the earth. From this it follows that g has a minimum value at the equator and increases towards the north and the south. This increase is quite regular as far as the poles, where g has its greatest value. We shall see later to what this is due. Here we merely take note of the fact. For the system which we have hitherto used for measuring forces and masses this fact, however, has very awkward consequences.

So long as weights are compared with each other only by means of the scale balance, there are no difficulties. But let us imagine a spring balance here in the laboratory which has been calibrated with weights. If we then bring this spring balance into more southerly or more northerly regions, we shall find that when loaded with the same weights it will give different deflections. If, therefore, we identify weight with force as we have hitherto done, there is nothing left for us but to assert that the force of the spring has altered and that it depends on the geographical latitude. But this is obviously not the case. It is not the force of the spring that has altered but the gravitational force. It is, therefore, wrong to take the weight of one and the same piece of metal as the unit of force at all points of the earth. We may choose the weight of a definite body at a definite point on the earth as the unit of force, and this may be applied at other points if the acceleration g due to gravity is known by pendulum measurements at both points. This is, indeed, just what technical science actually does do. Its unit of force is the weight of a definite normal body in Paris, the gramme. Hitherto we have always used this without taking into account its variability with position. In exact measurements, however, the value must be reduced to that at the normal place (Paris).

Science has departed from this system of measures, at which one place on the earth is favoured, and has selected a system that is less arbitrary.

The fundamental law of mechanics itself offers a suitable method for doing this. Instead of referring the mass to the force, we establish the mass as the fundamental quantity of the independent dimensions [M] and choose its unit arbitrarily: let a definite piece of metal have the mass 1. As a matter of fact, the same piece of metal that served technical science as the unit of weight, the Paris gramme, is taken for this purpose, and this unit of mass is likewise called the gramme (grm.).

The fact that the same word is used in technical science to denote the unit of weight and in physics to denote the unit of mass may easily lead to error. In the sequel we use the *physical system of measure*, the fundamental units of which are: cm. for length, sec. for time, grm. for mass.

Force now has the derived dimensions

$$[K] = [MB] = \left[\frac{ML}{T^2}\right]$$

and the unit, called the *dyne*, is grm. cm./sec.^{2}

Weight is defined by $G = mg$; thus the unit of mass has the weight $G = g$ dynes. It changes with the geographical latitude. and in our own latitude it has the value $g = 981$ dynes. This is the technical unit of force. The weight given by a spring balance, expressed in dynes, is, of course, a constant ; for its power of accelerating a definite mass is independent of the geographical latitude.

The dimensions of impulse or momentum are now :

$$[J] = [TK] = \left[\frac{ML}{T}\right]$$

and its unit is grm. cm./sec. Finally, the dimensions of energy are

$$[E] = [MV^2] = \left[\frac{ML^2}{T^2}\right]$$

and its unit is grm. cm.^{2}/sec.^{2}

Now that we have cleansed the system of measures of all earthly impurities, we can proceed to the mechanics of the stars.

CHAPTER III

THE NEWTONIAN WORLD-SYSTEM

1. Absolute Space and Absolute Time

THE principles of mechanics, as here developed, were partly suggested to Newton by Galilei's works and were partly created by himself. To him we owe above all the expression of definitions and laws in such a generalized form that they appear detached from earthly experiments and allow themselves to be applied to events in astronomic space.

In the first place Newton had to preface the actual mechanical principles by making definite assertions about space and time. Without such determinations even the simplest law of mechanics, that of inertia, has no sense. According to this, a body on which no force is acting is to move uniformly in a straight line. Let us fix our thoughts on the table with which we first experimented in conjunction with the rolling sphere. If now the sphere rolls on the table in a straight line, an observer who follows and measures its path from another planet would have to assert that the path is not a straight line according to his point of view. For the earth itself is rotating, and it is clear that a motion that appears rectilinear to the observer travelling with the earth, because it leaves the trace of a straight line on his table, must appear curved to another observer who does not participate in the rotation of the earth. This may be roughly illustrated as follows :

A circular disc of white cardboard is mounted on an axis so that it can be turned by means of a handle. A ruler is fixed in front of the disc. Now turn the disc as uniformly as possible, and at the same time draw a pencil along the ruler with constant velocity, so that the pencil marks its course on the disc. This path will, of course, not be a straight line on the disc, but a curved line, which will even take the form of a loop if the rotary motion is sufficiently rapid. Thus, the same motion which an observer fixed to the ruler would call uniform and rectilinear, would be called curvilinear (and non-uniform) by an observer moving with the disc. This motion may be constructed point for point, as is illustrated in the drawing (Fig. 32), which explains itself.

This example shows clearly that the law of inertia has sense, indeed, only when the space, or rather, the system of reference in which the rectilinear character of the motion is to hold, is exactly specified.

It is in conformity with the Copernican world-picture, of course, not to regard the earth as the system of reference, for which the law of inertia holds, but one that is somehow fixed in astronomic space. In experiments on the earth, for example, rolling the sphere on the table, the path of the freely moving body is not then in reality straight but a little curved. The fact that this escapes our primitive type of observation is due only to the shortness of the paths used in the experiments compared with the dimensions of the earth. Here, as has often happened in science, the inaccuracy of observation

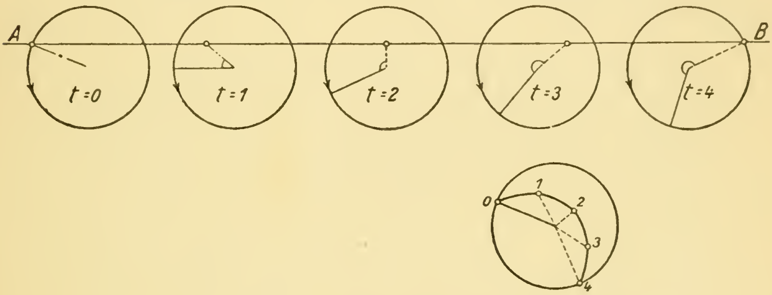


FIG. 32.

has led to the discovery of a great relationship. If Galilei had been able to make observations as refined as those of later centuries the confused mass of phenomena would have made the discovery of the laws much more difficult. Perhaps, too, Kepler would never have unravelled the motions of the planets, if the orbits had been known to him as accurately as at the present day. For Kepler's ellipses are only approximations from which the real orbits differ considerably in long periods of time. The position was similar, for example, in the case of modern physics with regard to the regularities of spectra; the discovery of simple relationships was rendered much more difficult and was considerably delayed by the abundance of very exact data of observation.

So Newton was confronted with the task of finding the system of reference in which the law of inertia and, further, all the other laws of mechanics were to hold. If he had chosen the sun, the question would not have been solved, but would only have been postponed, for the sun might one day be discovered also to be in motion, as has actually happened in the meantime.

Probably it was for such reasons that Newton gained the conviction that an empirical system of reference fixed by material bodies could, indeed, never be the foundation of a law involving the idea of inertia. But the law itself, through its close connection with Euclid's doctrine of space, the element of which is the straight line, appears as the natural starting-point of the dynamics of astronomic space. It is, indeed, in the law of inertia that Euclidean space manifests itself outside the narrow limits of the earth. Similar conditions obtain in the case of time, the flow of which receives expression in the uniform motion due to inertia.

In this way, possibly, Newton came to the conclusion that there is an *absolute space and an absolute time.* It will be best to give the substance of his own words. Concerning time he says:

I. "*Absolute, true* and *mathematical time* flows in itself and in virtue of its nature uniformly and without reference to any external object whatever. It is also called duration."

"Relative, apparent, and ordinary time is a perceptible and external, either exact or unequal, measure of duration, which we customarily use instead of true time, such as hour, day, month, year."

"Natural days, which are usually considered as equal measures of time are really unequal. This inequality is somewhat corrected by the astronomers who measure the motion of the heavenly bodies according to the correct time. It may be that there is uniform motion by which time may be measured accurately. All motions may be accelerated or retarded. Only the flow of absolute time cannot be changed. The same duration and the same persistence occurs in the existence of all things, whether the motions be rapid, slow, or zero."

Concerning space Newton expresses similar opinions. He says:

II. "*Absolute space,* in virtue of its nature and without reference to any external object whatsoever, always remains immutable and immovable."

"Relative space is a measure or a movable part of the absolute space. Our senses designate it by its position with respect to other bodies. It is usually mistaken for the immovable space."

"So in human matters we, not inappropriately, make use of *relative* places and motions instead of *absolute* places and motions. In natural science, however, we must abstract from the data of the senses. For it may be the case that no body that is really at rest exists, with reference to which we may refer the places and the motion."

The definite statement, both in the definition of absolute time as in that of absolute space, that these two quantities exist " without reference to any external object whatsoever " seems strange in an investigator of Newton's attitude of mind. For he often emphasizes that he wishes to investigate only what is actual, what is ascertainable by observation. " Hypotheses non fingo," is his brief and definite expression. But what exists " without reference to any external object whatsoever " is not ascertainable, and is not a fact. Here we have clearly a case in which the ideas of unanalysed consciousness are applied without reflection to the objective world. We shall investigate the question in detail later on.

Our next task is to describe how Newton interpreted the laws of the cosmos and in what the advance due to his doctrine consisted.

2. Newton's Law of Attraction

Newton's idea consisted in setting up a dynamical idea of planetary orbits, or, as we nowadays express it, in founding *Celestial Mechanics*. To do this it was necessary to apply Galilei's conception of force to the motions of the stars. Yet Newton did not find the law according to which the heavenly bodies act on one another by setting up bold hypotheses, but by pursuing the systematic and exact path of analysing the known facts of planetary motions. These facts were expressed in the three Kepler laws that compressed all the observations of that period of time into a wonderfully concise and vivid form. We must here state Kepler's laws in full. They are :

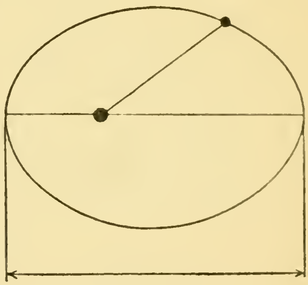


Fig. 33.

1. The planets move in ellipses with the sun at one of the foci (Fig. 33).
2. The radius vector drawn from the sun to a planet describes equal areas in equal times.
3. The cubes of the major axes of the ellipses are proportional to the squares of the periods of revolution.

Now the fundamental law of mechanics gives a relation between the acceleration b of the motion, and the force K that produces it. The acceleration b is completely determined by the course of the motion, and, if this is known, b can be calculated. Newton recognized that the orbit as defined by Kepler's laws just sufficed to allow a calculation of b. Then the law

$$K = mb$$

also allows the acting force to be calculated.

The ordinary mathematics of Newton's time would not have enabled him to carry out this calculation. He had first to invent the mathematical apparatus. Thus there was created in England the *Differential and Integral Calculus*, the root of the whole of modern mathematics, as a bye-product of astronomical researches, whereas Leibniz (1684) simultaneously invented the same method on the continent by starting from a totally different point of view.

Since we do not wish to use the infinitesimal calculus in this book we cannot pause to give a picture of the wonderful nature of Newton's inferences. Yet the fundamental idea may be illustrated by a simple example.

The orbits of the planets are slightly eccentric ellipses, that are almost of a circular shape. It will be permissible to assume approximately that the planets describe circles about the sun, as was, indeed, supposed by Copernicus. Since circles are special ellipses with the eccentricity zero, this assumption certainly fulfils Kepler's first law.

The second law next states that every planet traverses its circle with constant speed. Now, by II 4, we know all about the acceleration in such circular motions. It is directed towards the centre and, by formula (4), p. 25, it has the value

$$b = \frac{v^2}{r}$$

where $v =$ the speed in the orbit, and r is the radius of the circle.

If now T is the period of revolution, the velocity is determined as the ratio of the circumference $2\pi r (\pi = 3{\cdot}14159 \ldots)$ to the time T, thus

$$v = \frac{2\pi r}{T} \quad . \quad . \quad . \quad . \quad (18)$$

so that

$$b = \frac{4\pi^2 r^2}{r T^2} = \frac{4\pi^2 r}{T^2}$$

We next direct our attention to the third Kepler law which, in the case of a circular orbit, clearly states that the ratio of the cube of the radius, r^3, to the square of the time of revolution T^2, has the same value C for all planets :

$$\frac{r^3}{T^2} = C \quad \text{or} \quad \frac{r}{T^2} = \frac{C}{r^2} \qquad . \qquad . \qquad . \quad (19)$$

If we insert this in the value for b above, we get

$$b = \frac{4\pi^2 C}{r^2} \qquad . \qquad . \qquad . \qquad . \quad (20)$$

According to this the value of the centripetal acceleration depends only on the distance of the planet from the sun, being inversely proportional to the square of the distance, but it is quite independent of the properties of the planet, such as its mass. For the quantity C is, by Kepler's third law, the same for all planets, and can therefore involve at most the nature of the sun and not that of the planets.

Now, it is a remarkable circumstance that exactly the same law comes out for elliptic orbits—by a rather more laborious calculation, it is true. The acceleration is always directed towards the sun situated at a focus, and has the value given by formula (20).

3. General Gravitation

The law of acceleration thus found has an important property in common with the gravitational force on the earth (weight) : it is quite independent of the nature of the moving body. If we calculate the force from the acceleration, we find it likewise directed towards the sun. It is thus an attraction and has the value

$$K = mb = m\frac{4\pi^2 C}{r^2} \qquad . \qquad . \qquad . \quad (21)$$

It is proportional to the mass of the moving body, just like the weight

$$G = mg$$

of a body on the earth.

This fact suggests to us that both forces may have one and the same origin. Nowadays, this circumstance, having been handed down to us through the centuries, has become such a truism, as it were, that we can scarcely conceive how bold and how great was Newton's idea. What a prodigious imagination it required to conceive the motion of the planets about the sun or of the moon about the earth as a process of "falling" that takes place according to the same laws and under the action of the same force as the falling of a stone released by my hand. The fact that the planets or the moon do not actually rush into their central bodies of attraction is due to the law of inertia that here expresses itself as a centrifugal force. We shall have to deal with this again later.

Newton first tested this idea of *general weight or gravitation* in the case of the moon, the distance of which from the earth was known from angular measurements.

This test is so important that we shall repeat the very simple calculation here as evidence of the fact that all scientific ideas become valid and of worth only when calculated and measured numerical values agree.

The central body is now the earth; the moon takes the place of the planet. r denotes the radius of the moon's orbit, T the period of revolution of the moon. Let the radius of the earth be a. If the gravitational force on the earth is to have the same origin as the attraction that the moon experiences from the earth, then the acceleration g due to gravity must, by Newton's law (20), have the form

$$g = \frac{4\pi^2 C}{a^2}$$

where C has the same value as for the moon, namely, by (19),

$$C = \frac{r^3}{T^2}.$$

If we insert this value in that for g, we get

$$g = \frac{4\pi^2 r^3}{T^2 a^2} \quad . \quad . \quad . \quad . \quad (22)$$

Now, the "sidereal" period of revolution of the moon, that is, the time between two positions in which the line connecting the earth to the moon has the same direction with respect to the stars, is

$$\begin{aligned} T &= 27 \text{ days } 7 \text{ hours } 43 \text{ minutes } 12 \text{ seconds} \\ &= 2{,}360{,}592 \text{ seconds.} \end{aligned}$$

In physics it is customary to write down a number to only so many places as are required for further calculation. So we write here

$$T = 2{\cdot}36 \,.\, 10^6 \text{ secs.}$$

The distance of the moon from the centre of the earth is about 60 times the earth's radius, or, more exactly,

$$r = 60{\cdot}1a.$$

The earth's radius itself is easy to remember because the metric system of measures is simply related to it. For 1 metre = 100 cms. = one ten-millionth of the earth's quadrant, that is, the forty-millionth part or $(4 \cdot 10^7)^{th}$ part of the earth's circumference $2\pi a$

$$100 = \frac{2\pi a}{4 \cdot 10^7}, \quad \text{or} \quad a = 6{\cdot}37 \cdot 10^8 \text{ cms.} \qquad . \quad (23)$$

If we insert all these values in (22) we get

$$g = \frac{4\pi^2 \cdot 60{\cdot}1^3 \cdot 6{\cdot}37 \cdot 10^8}{2{\cdot}36^2 \cdot 10^{12}} = 981 \text{ cm./sec.}^2 \qquad . \quad (24)$$

This value agrees exactly with that found by pendulum observations on the earth (see II, 12, p. 37).

The great importance of this result is that it represents the *relativization of the force of weight.* To the ancients weight denoted a pull towards the absolute " below," which is experienced by all earthly bodies. The discovery of the spherical shape of the earth brought with it the relativization of the direction of earthly weight ; it became a pull towards the centre of the earth.

And now the identity of earthly weight with the force of attraction that keeps the moon in her orbit is proved, and since there can be no doubt that the latter is similar in nature to the force that keeps the earth and the other planets in their orbits round the sun, we get the idea that bodies are not simply " heavy " but are mutually heavy or *heavy relatively to each other.* The earth, being a planet, is attracted towards the sun, but it itself attracts the moon. Obviously this is only an approximate description of the true state of affairs, which consists in the sun, moon, and earth attracting each other. Certainly, so far as the orbit of the earth round the sun is concerned, the latter may, to a high degree of approximation, be regarded as at rest, because its enormous mass hinders the calling up of appreciable accelerations, and, conversely, the moon, on account of its size, does not come into account. But an exact theory will have to take into consideration these influences, called " perturbations."

Before we begin to consider more closely this view, which signifies the chief advance of Newton's theory, we shall give Newton's law its final form. We saw that a planet situated at a distance r from the sun experiences from it an attraction of the value (21)

$$K = m\frac{4\pi^2 C}{r^2},$$

where C is a constant depending only on the properties of the sun, not on those of the planet. Now, according to the new view of mutual or relative weight the planet must likewise attract the sun. If M is the mass of the sun, c a constant dependent only on the nature of the planet, then the force exerted on the sun by the planet must be expressed by

$$K' = M\frac{4\pi^2 c}{r^2}.$$

But earlier, in introducing the conception of force (II, 1, p. 16), we made use of the principle that the reaction equals the action, which is one of the simplest and most certain laws of mechanics. If we apply it here, we must set $K = K'$, or

$$m\frac{4\pi^2 C}{r^2} = M\frac{4\pi^2 c}{r^2}.$$

From this it follows that

$$mC = Mc,$$

or

$$\frac{C}{M} = \frac{c}{m},$$

that is, this ratio has the same value for both bodies (sun and planets), and hence also for any body whatsoever. If we call this value $\frac{k}{4\pi^2}$, then we may write

$$4\pi^2 C = kM \qquad 4\pi^2 c = km \quad . \quad . \quad . \quad (25)$$

The factor of proportionality k is called the *gravitational constant.*

The Newtonian law of general gravitation then assumes the symmetrical form

$$K = k\frac{mM}{r^2} \quad . \quad . \quad . \quad . \quad (26)$$

In words it states :

Two bodies attract each other with a force that is proportional to the mass of each body and is inversely proportional to the square of their distance apart.

4. Celestial Mechanics

It is only in this general form that the Newtonian law denotes a real advance in the calculation of the planetary orbits. For in the original form it was deduced from Kepler's laws by calculation and denoted no more than a very short and striking resumé of these laws. It is also possible to prove conversely that the motion of a body about a central body that is at rest and that attracts it according to Newton's law is necessarily a Kepler elliptic motion. A new feature arises only when, firstly, we now regard both bodies as moving and, secondly, add further bodies in the problem.

Then we get the *problem of three or more* bodies, which corresponds exactly to the actual conditions in the planetary system (Fig. 34). For not only are the planets attracted by the sun and the moons by their planets, but every body, be it sun, planet, moon, or comet, attracts every other body. Accordingly, the Kepler ellipses appear to be only approximately valid, and they are so only because the sun on account of its great mass overshadows by far the reciprocal action of all other bodies of the planetary system. But in long periods of time these reciprocal actions must also manifest themselves as deviations from the Kepler laws. We speak, as already remarked, of "perturbations."

In Newton's time such perturbations were already known, and in the succeeding centuries refinements in the methods of observation have accumulated an immense number of facts that had to be accounted for by Newton's theory. That it succeeded in doing so is one of the greatest triumphs of human genius.

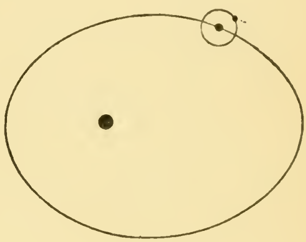


FIG. 34.

It is not our aim here to pursue the development of mechanics from Newton's time to the present day, and to describe the mathematical methods that were devised to calculate the "perturbed" orbits. The most ingenious mathematicians of all countries have played a part in setting up the "*theory of perturbations*," and even if no satisfactory solution has yet been found for the problem of three bodies, it is possible to calculate with certainty the motions for hundred thousands or millions of years ahead or back. So Newton's theory was tested in countless cases in new observations, and it has never failed—except in one case, of which we shall presently speak. Theoretical astronomy, as founded by Newton, was therefore long regarded as a model for the exact sciences. It achieved what had been the longing of mankind since earliest history. It lifts the veil that is spread over the future; it endows its followers with the gift of prophecy. Even if the subject-matter of astronomic predictions is unimportant or indifferent for human life, yet it became a symbol for the liberation of the spirit from the trammels of earthly bonds. We, too, follow the peoples of earlier times in gazing upwards with reverential awe at the stars, which reveal to us the law of the world.

But the world-law can tolerate no exception. Yet there is one case, as we have already mentioned, in which Newton's theory has failed. Although the error is small, it is not to be denied. It occurs in the case of the planet Mercury, the planet nearest the sun. The orbit of any planet may be regarded as a Kepler elliptic motion that is perturbed by the other planets, that is, the position of the orbital plane, the position of the major axis of the ellipse, its eccentricity, in short, all "elements of the orbit" undergo gradual changes. If we calculate these according to Newton's law and apply them to the observed orbit, it must become transformed into an exact Kepler orbit, that is, an ellipse in a definite plane at rest, with a major axis of definite direction and length, and so forth. This is so, indeed, for all planets, except that a little error remains in the case of Mercury. The direction of the major axis, that is, the line connecting the sun with the nearest focus, the *perihelion* (Fig. 35), does not remain fixed after all the above corrections have been applied, but executes a very slow motion

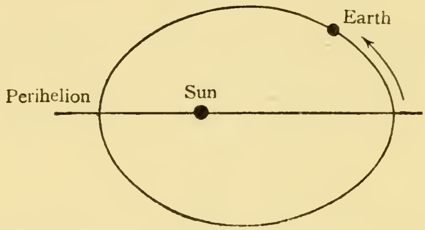


FIG. 35.

of rotation, advancing 43 secs. of arc every hundred years. The astronomer Leverrier (1845)—the same who predicted the existence of the planet Neptune from calculations based on the perturbations—first calculated this motion, and it is fully established. Yet it cannot be explained by the Newtonian attraction of the planetary bodies known to us. Hence recourse has been taken to hypothetical masses whose attraction was to bring about the motion of Mercury's perihelion. Thus, for example, the zodiacal light, which is supposed to emanate from thinly distributed nebulous matter in the neighbourhood of the sun, was brought into relation with the anomaly of Mercury. But this and numerous other hypotheses all suffer from the fault that they have been invented *ad hoc* and have been confirmed by no other observation.

The fact that the only quite definitely established deviation from Newton's law occurs in the case of Mercury, the planet nearest the sun, indicates that perhaps there is after all some fundamental defect in the law. For the force of attraction is greatest in the proximity of the sun, and hence deviations from the law of the inverse square will show themselves first there. Such changes have also been made, but as they are invented quite arbitrarily and can be tested by no other facts, their correctness is not proved by accounting for the motion of Mercury's perihelion. If Newton's theory really requires a refinement we must demand that it emanates, without the introduction of arbitrary constants, from a principle that is superior to the existing doctrine in generality and intrinsic probability.

Einstein was the first to succeed in doing this by making general relativity the most fundamental postulate of physical laws. We shall revert in the last chapter to his explanation of the motion of Mercury's perihelion.

5. The Relativity Principle of Classical Mechanics

In discussing the great problems of the cosmos we have almost forgotten the point of departure from the earth. The laws of dynamics found to hold on the earth were transplanted to the astronomical space through which the earth rushes in its orbit about the sun with stupendous speed. How is it, then, that we notice so little of this journey through space? How is it that Galilei succeeded in finding laws on the moving earth which, according to Newton, were to be rigorously valid only in space absolutely at rest? We have called attention to this question above when mentioning Newton's views about space and time. We stated there that the apparently straight path of a sphere rolling on the table would, in reality, owing to the rotation of the earth, be slightly curved, for the path is straight not with respect to the moving earth but with respect to absolute space. The fact that we do not notice this curvature is due to the shortness of the path and of the time of observation, during which the earth has turned only slightly. Even if we admit this, we are still left with the motion of revolution about the sun, which proceeds with the immense speed of 30 kms. per sec. Why do we notice nothing of this?

This motion, due to the revolution, is also, indeed, a rotation, and this must make itself remarked in earthly motions similarly to the rotation of the earth on its own axis, only much less, since the curvature of the earth's orbit is very small. But in our question we do not mean this rotatory motion but the forward motion, which, in the course of a day, is practically rectilinear and uniform.

Actually, all mechanical events on the earth occur as if this tremendous forward motion does not exist, and this law holds quite generally for every system of bodies that executes a uniform and rectilinear motion through Newton's absolute space. This is called the *relativity principle* of classical mechanics, and it may be formulated in various ways. For the present, we shall enunciate it as follows:

Relatively to a co-ordinate system moving rectilinearly and uniformly through absolute space the laws of mechanics have exactly the same expression as when referred to a co-ordinate system at rest in space.

To see the truth of this law we need only keep clear in our minds the fundamental law of mechanics, the law of impulses, and the conceptions that occur in it. We know that a blow produces a *change* of velocity. But such a change is quite independent of whether the velocities before and after the blow, v_1 and v_2, are referred to absolute space or to a system of reference which is itself moving with the constant velocity a. If the moving body is moving before the blow in space with the velocity $v_1 = 5$ cms. per sec., then an observer moving with the velocity $a = 2$ cms. per sec. in the same direction would measure only the relative velocity $v_1' = v_1 - a = 5 - 2 = 3$. If the body now experiences a blow in the direction of motion which magnifies its velocity to $v_2 = 7$ cms. per sec., then the moving observer would measure the final velocity as $v_2' = v_2 - a = 7 - 2 = 5$. Thus the change of velocity produced by the blow is $w = v_2 - v_1 = 7 - 5 = 2$ in absolute space. On the other hand the moving observer notes the increase of velocity as

$$w' = v_2' - v_1' = (v_2 - a) - (v_1 - a) = v_2 - v_1 = w = 5 - 3 = 2.$$

Both are of the same value.

Exactly the same holds for continuous forces and for the accelerations produced by them. For the acceleration b was defined as the ratio of the change of velocity w to the time t required in changing it, and since w is independent of whatever rectilinear uniform forward motion (motion of translation) the system of reference used for the measurement has, the same holds for b.

The root of this law is clearly the law of inertia, according to which a motion of translation occurs when no forces act. A system of bodies, all of which travel through space with the same constant velocity, is hence not only at rest as regards their geometric configuration, but also no actions of forces manifest themselves on the bodies of the system in consequence of the motion. But if the bodies of the system exert forces *on each other*, the motions thereby produced will occur relatively just as if the common motion of translation were not taking place. Thus, for an observer moving with the system, it would not be distinguishable from one at rest.

The experience, repeated daily and thousands of times, that we observe nothing of the translatory motion of the earth is a tangible proof of this law. But the same fact manifests itself in motions on the earth. For when a motion on the earth is rectilinear and uniform with respect to the earth, it is so also with respect to space, if we disregard the rotation in the earth's motion. Everyone knows that in a ship or a railway carriage moving uniformly mechanical events occur in the same way as on the earth (considered at rest). On the moving ship, too, for example, a stone falls vertically ; it falls along a vertical that is moving with the ship. If the ship were to move quite uniformly and without jerks of any kind the passengers would notice nothing of the motion so long as they did not observe the apparent movement of the surroundings.

6. Limited Absolute Space

The law of the relativity of mechanical events is the starting-point of all our later arguments. Its importance rests on the fact that it is intimately connected with Newton's views on the absolute space, and essentially limits the physical reality of this conception from the outset.

We gave as the reason that made it necessary to assume absolute space and absolute time that without it the law of inertia would be utterly meaningless. We must now enter into the question as to how far these conceptions deserve the terms " real " in the sense of physics. A conception has physical reality only when there is something ascertainable by measurement corresponding to it in the world of phenomena. This is not the place to enter into a discussion on the philosophic conception of reality. At any rate it is quite certain that the criterion of reality just given corresponds fully with the way reality is used in the physical sciences. Every conception that does not satisfy it has gradually been pushed out of the system of physics.

We see at once that in this sense a definite place in Newton's absolute space is nothing real. For it is fundamentally impossible to find the same place a second time in space.

This is clear at once from the principle of relativity. Given that we had somehow arrived at the assumption that a definite system of reference is at rest in space, then a system of reference moving uniformly and rectilinearly with respect to it may with equal right be regarded as at rest. The mechanical events in both occur quite similarly and neither system enjoys preference over the other. A definite body that seems at rest in the one system of rest performs a rectilinear and uniform motion, as seen from the other system, and if anyone were to assert that this body marks a spot in absolute space, another may with equal right challenge this and declare the body to be moving.

In this way the absolute space of Newton already loses a considerable part of its weird existence. A space in which there is no place that can be marked by any physical means whatsoever, is at any rate a very subtle configuration, and not simply a box into which material things are crammed.

We must now also alter the terms used in our definition of the principle of relativity, for in it we still spoke of a co-ordinate system at rest in absolute space, and this is clearly without sense physically. To arrive at a definite formulation the conception of *inertial system* (*inertia = laziness*) has been introduced, and it is taken to signify a co-ordinate system in which the law of inertia holds in its original form. There is not only the one system at rest as in Newton's absolute space, where this is the case, but an infinite number of others that are all equally justified, and since we cannot well speak of several "spaces" moving with respect to each other, we prefer to avoid the word "space" as much as possible. The *principle of relativity* then assumes the following form:

There are an infinite number of equally justifiable systems, inertial systems, executing a motion of translation with respect to each other, in which the laws of mechanics hold in their simple classical form.

We here see clearly how intimately the problem of space is connected with mechanics. It is not space that is there and that impresses its *form* on things, but the things and their physical laws determine space. We shall find later how this view gains more and more ground until it reaches its climax in the general theory of relativity of Einstein.

7. Galilei Transformations

Although the laws of mechanics are the same in all inertial systems, it does not of course follow that co-ordinates and velocities of bodies with respect to two inertial systems in relative motion are equal. If, for example, a body is at rest in a system S, then it has a constant velocity with respect to the other system S′, moving relatively to S. The general laws of mechanics contain only the accelerations, and these, as we saw, are the same for all inertial systems. This is not true of the co-ordinates and the velocities.

Hence the problem arises to find the position and the velocity of a body in an inertial system S′ when they are given for another inertial system S.

It is thus a question of passing from one co-ordinate system to another, which is moving relatively to the former. We must at this stage interpose a few remarks about equivalent (equally justified) co-ordinate systems in general and about the laws, the so-called *transformation equations*, that allow us to pass from one to the other by calculation.

In geometry co-ordinate systems are a means of fixing in a convenient manner the relative positions of one body with respect to another. For this we suppose the co-ordinate system to be rigidly fixed to the one body. Then the co-ordinates of the points of the other body fix the relative position

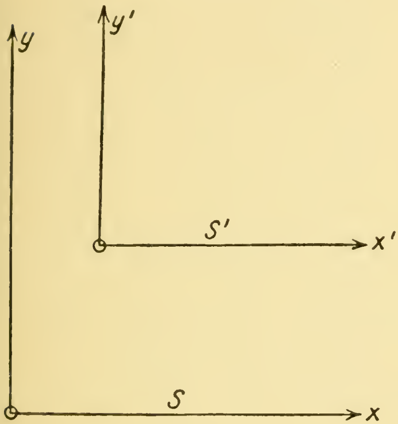


Fig. 36.

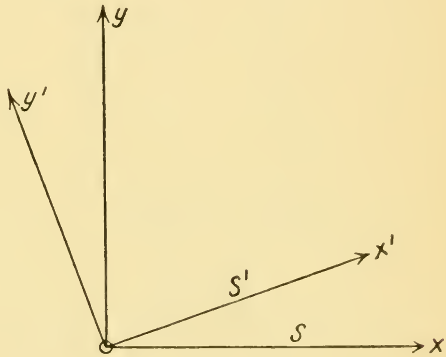


Fig. 37.

completely. It is, of course, immaterial whether the co-ordinate system is chosen as rectangular, oblique, polar or still more general. It is also immaterial how it is orientated with respect to the first body, except that either this orientation must be maintained, or, if it is changed, we must specify how the co-ordinate system alters its position with respect to the body. If, for example, we operate with rectangular co-ordinates in a plane, then in place of the system S first chosen we may select a second, S′, which is displaced (Fig. 36) or turned (Fig. 37) with respect to S. But we must specify exactly how great is the displacement and the turning. From these data we can then calculate what the co-ordinates of a point P that had the values x, y in the old system S are in the new system S′. If we call them x', y' we get formulæ that allow us to calculate x',y' from x, y. We shall do this for the simplest case, namely, that in which the system S′ arises from S as the result of a parallel displacement by the amount a in the x-direction (Fig. 38). Then clearly the new co-ordinate x' of a point P will be equal to its old x diminished by the displacement a, whereas the y-co-ordinate remains unaltered. Thus we have

$$x' = x - a, y' = y \quad . \quad . \quad . \quad (27)$$

Similar, but more complicated, transformation formulæ hold in the other case. We shall later have to discuss this more fully. It is important to recognize that every quantity that has a geometric meaning in itself must be independent of the choice of the co-ordinate system and must hence be expressed in similar co-ordinate systems in a similar way. Such a quantity is said to be *invariant* with respect to the co-ordinate-transformation concerned. Let us consider, as an example, the transformation (27) above, that expresses a displacement along the x-axis. It is clear that the difference of the x-co-ordinates of two points P and Q, namely, $x_2 - x_1$, does not change. As a matter of fact (Fig. 39),

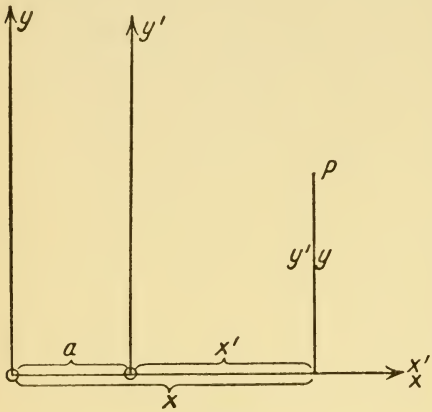


Fig. 38.

$$x'_2 - x'_1 = (x_2 - a) - (x_1 - a) = x_2 - x_1$$

If the two co-ordinate systems S and S′ are inclined to each other, then the distance s of any point P from the origin is an invariant (Fig. 40). It has the same expression in both systems, for, by Pythagoras' theorem, we have

$$s^2 = x^2 + y^2 = x'^2 + y'^2 \quad . \quad . \quad . \quad (28)$$

In the more general case, in which the co-ordinate system is simultaneously displaced and turned, the distance P, Q of two points becomes an invariant. The invariants are particularly important because they represent the geometrical relations in themselves without reference to the accidental choice of the co-ordinate system. They will play a considerable part in the sequel.

If we now return after this geometrical digression to our starting-point, we have to answer the question as to what are the transformation laws that allow us to pass from one inertial system to another.

We defined the inertial system as a co-ordinate system in which the law of inertia holds. Only the state of motion is important in this connexion, namely, the absence of accelerations with respect to the absolute space, whereas the nature and position of the co-ordinate system is unessential. If we choose it to be rectangular, as happens most often, its position still remains free. We may take a displaced or a rotated system, only it must have the same state of motion. In the foregoing we have always spoken of *system of reference* wherever we were concerned with the state of motion and not with the

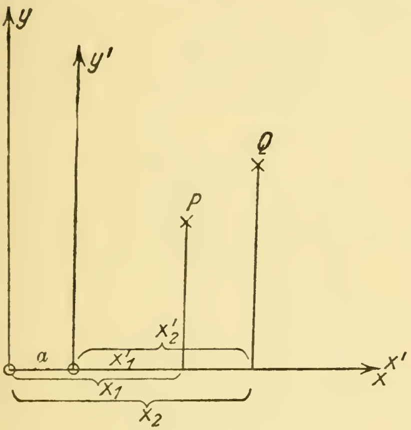


FIG. 39.

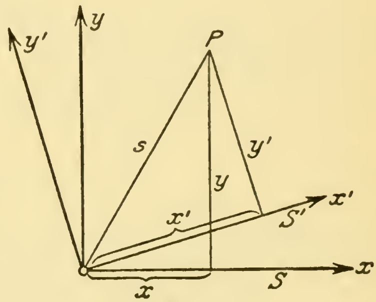


FIG. 40.

nature and position of the co-ordinate system, and we shall use the expression systematically from now onwards.

If an inertial system S′ is moving rectilinearly with respect to S with the velocity v, we may choose rectangular co-ordinates in both systems of reference such that the direction of motion becomes the x- and the x'-axis, respectively. Further, we may assume that at the time $t = o$ the origin of both systems coincides. Then, in the time t the origin of the S′-system will have been displaced by the amount $a = vt$ in the x-direction: thus at this moment the two systems are exactly in the position that was treated above purely geometrically. Hence the equations (27) hold, in which a is now to be set equal to vt. Consequently we get the transformation equations

$$x' = x - vt \qquad y' = y \qquad z' = z \qquad . \qquad . \quad (29)$$

in which we have added the unchanged z-co-ordinate. This law is called a Galilei transformation in honour of the founder of mechanics.

We may also enunciate the *principle of relativity* as follows :

The laws of mechanics are invariant with respect to Galilei transformations.

This is due to the fact that accelerations are invariant, as we have already seen above by considering the change of velocity of a moving body with respect to two inertial systems.

We showed earlier that the theory of motions or kinematics may be regarded as a geometry in four-dimensional *xyzt*-space, the " world " of Minkowski. In this connexion it is not without interest to consider what the inertial systems and the Galilei transformations signify in this four-dimensional geometry. This is by no means difficult, for the *y*- and the *z*-co-ordinate do not enter into the transformation at all. It is thus sufficient to operate in the *xt*-plane.

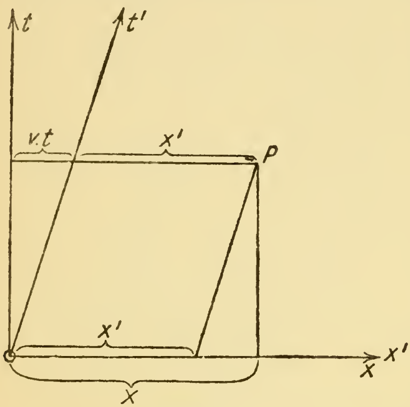


Fig. 41.

We represent our inertial system S by a rectangular *xt*-co-ordinate system (Fig. 41). A second inertial system S′ then corresponds to another co-ordinate system $x't'$, and the question is : what does the second look like and how is it situated relatively to the first ? First of all, the time-measure of the second system S′ is exactly the same as that of the first, namely, the one absolute time $t = t'$; thus the *x*-axis, on which $t = o$ lies, coincides with the x'-axis, $t' = o$. Consequently the system S′ can only be an oblique co-ordinate system. The t'-axis is the world-line of the point $x' = o$, that is, of the origin of the system S′. The *x*-co-ordinate of this point which moves with the velocity v relatively to the system S is equal to vt in this system at the time t. For any world-point P whatsoever the figure then at once gives the formula of the Galilei transformation $x' = x - vt$.

Corresponding to any other inertial system there is another oblique *xt*-co-ordinate system with the same *x*-axis, but a differently inclined *t*-axis. The rectangular system from which we started has no favoured position among all these oblique systems. The unit of time is cut off from all the *t*-axes of the various co-ordinate systems by the *same* parallel to the *x*-axis.

This is in a certain sense the " calibration curve " of the *xt*-plane with respect to the time.

We compress the result into the sentence :

In the xt-plane the choice of the direction of the t-axis is quite arbitrary ; in every xt-co-ordinate system having the same x-axis the fundamental laws of mechanics hold.

From the geometric point of view this manifold of equivalent co-ordinate systems is extremely singular and unusual. The fixed position or the invariance of the x-axis is particularly remarkable. When we operate in geometry with oblique co-ordinates there is usually no reason for keeping the position of one axis fixed. But this is required by Newton's fundamental law of absolute time. All events which occur simultaneously, that is for the same value of t, are represented by a parallel to the x-axis. Since, according to Newton, time flows " absolutely and without reference to any object whatsoever," simultaneous events must correspond to the same world-point in all allowable co-ordinate systems.

We shall see that this unsymmetrical behaviour of the world-co-ordinates x and t, here only mentioned as an error of style, is actually non-existent. Einstein has eliminated it through his relativization of the conception of time.

8. Inertial Forces

After having recognized that the individual points in Newton's absolute space have at any rate no physical reality, we enquire what remains of this conception at all. Well, it asserts itself quite clearly and emphatically, for the resistance of all bodies to accelerations must be interpreted in Newton's sense as the action of absolute space. The locomotive that sets the train in motion must overcome the inertial resistance. The shell that demolishes a wall draws its destructive power from inertia. Inertial actions arise wherever accelerations occur, and these are nothing more than changes of velocity in absolute space ; we may use the latter expression, for a change of velocity has the same value in all inertial systems. Systems of reference that are themselves accelerated with respect to inertial systems are thus *not* equivalent to the latter, or equivalent among themselves. We can, of course, also refer the laws of mechanics to them, but they then assume a new and more complicated form. Even the path of a body left to itself is no longer uniform and rectilinear in an accelerated system (see III, I, p. 48). This may also be expressed by saying that in an accelerated system *apparent forces, inertial forces,* act besides the true forces. A body on which no true forces act is yet subject to these inertial forces, and its motion is therefore in general neither uniform nor rectilinear. For example, a vehicle when being set into motion or stopped is such an accelerated system. Railway journeys have made everyone familiar with the jerk due to the train starting or stopping, and this is nothing other than the inertial force of which we have spoken.

We shall consider the phenomena individually for a system S moving rectilinearly, whose acceleration is to be equal to k. If we now measure the acceleration b of a body with respect to this moving system S, then the acceleration with respect to absolute space is obviously greater to the extent k. Hence the fundamental dynamical law with respect to space is

$$m(b + k) = \mathrm{K}.$$

If we write this in the form

$$mb = \mathrm{K} - mk,$$

we may say that in the accelerated system S a law of motion of Newtonian form, namely,

$$mb = \mathrm{K}'$$

again holds, except that now we must write for the force K′ the sum

$$\mathrm{K}' = \mathrm{K} - mk$$

where K is the true, and $- mk$ the apparent or inertial force.

Now, if there is no true force acting, that is, if $\mathrm{K} = o$, then the total force becomes equal to the force of inertia

$$\mathrm{K}' = - mk \quad . \quad . \quad . \quad . \quad (30)$$

Thus this force acts on a body left to itself. We may recognize its action from the following considerations. We know that the gravitation on the earth, the force of gravity, is determined by the formula $\mathrm{G} = mg$, where g is the constant acceleration due to gravity. The force of inertia $\mathrm{K}' = - mk$ thus acts exactly like weight or gravity; the minus sign denotes that the force of acceleration is in a direction opposite to the system of reference S used as a basis. The value of the apparent gravitational acceleration k is equal to the acceleration of the system of reference S. Thus the motion of a body left to itself in the system S is simply a motion such as that due to falling or being thrown.

This relationship between the inertial forces in accelerated systems and the force of gravity still appears quite fortuitous here. It actually remained unobserved for two hundred years.

But even at this stage we must state that it forms the basis of Einstein's general theory of relativity.

9. Centrifugal Forces and Absolute Space

In Newton's view the occurrence of inertial forces in accelerated systems proves the existence of absolute space or, rather, the favoured position of inertial systems. Inertial forces present themselves particularly clearly in rotating systems of reference in the form of *centrifugal forces*. It was from them that Newton drew his main support for his doctrine of absolute space. Let us give the substance of his own words:

"The effective causes which distinguish absolute and relative motion from each other are centrifugal forces, the forces tending to send bodies away from the axis of rotation. In the case of a motion that is only relatively circular these forces do not exist, but they are smaller or greater in proportion to the amount of the (absolute) motion."

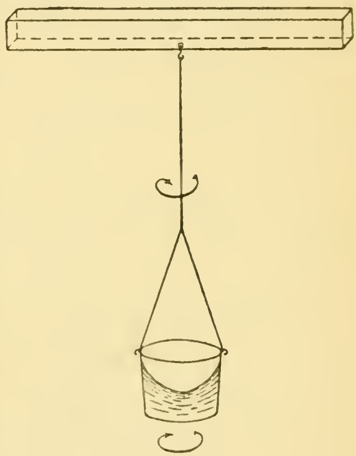


Fig. 42.

"Let us, for example, hang a vessel by a very long thread and turn it about its axis until the thread becomes very stiff through the torsion (Fig. 42). Then let us fill it with water and wait till both vessel and contents are completely at rest. If it is now made to rotate in the opposite direction by a force applied suddenly, and if this lasts for some time whilst the thread unwinds itself, the surface of the water will first be plane, just as before the vessel began to move, and then when the force gradually begins to act on the water, the vessel will make the water participate appreciably in the motion. It (the water) gradually moves away from the middle and mounts up the walls of the vessel, assuming a hollow shape (I have carried out this experiment personally)."

"At the beginning when the relative motion of the water in the vessel (with respect to the walls) was greatest, it displayed no tendency to move away from the axis. The water did not seek to approach the periphery by climbing up the walls, but remained plane, and thus the *true* circular motion had not yet begun. Later, however, as the relative motion of the water decreased, its ascent up the walls expressed the tendency to move away from the axis, and this tendency showed the continually increasing *true* circular motion of the water, until this finally reached a maximum, when the water itself was resting *relatively* to the vessel."

" Moreover, it is very difficult to recognize the *true* motions of individual bodies and to distinguish them from the *apparent* motions, because the parts of that immovable space in which the bodies are truly moving cannot be perceived by the senses."

" Yet the position is not quite hopeless. For the necessary auxiliary means are given partly by the apparent motions, which are the differences of the real ones, and partly by the forces on which the true motions are founded as working causes. If, for example, two spheres are connected at a given distance apart by means of a thread and thus turned about the usual centre of gravity (Fig. 43), we recognize in the tension of the thread the tendency of the spheres to move away from the axis of the motion, and from this we can get the magnitude of the circular motion . . . In this way we could find both the magnitude and the direction of this circular motion in every infinitely great space, even if there were nothing external and perceptible in it, with which the spheres could be compared."

These words express most clearly the meaning of absolute space. We have only a few words of explanation to add to them.

Concerning, firstly, the quantitative conditions in the case of the centrifugal forces we can at once get a survey of these if we call to mind the magnitude and the direction of the acceleration in the case of circular motions. It was directed towards the centre and, according to formula (4), p. 25, it had the value $b = \frac{v^2}{r}$, where r denotes the circular radius, and v the velocity.

Now, if we have a rotating system of reference S that rotates once in T secs., then the velocity of a point at the distance r from the axis (see formula (18), p. 52) is

$$v = \frac{2\pi r}{T},$$

hence the acceleration relative to the axis, which we denoted by k (see p. 68) is

$$k = \frac{4\pi^2 r}{T^2}.$$

Now, if a body has the acceleration b relatively to S, its absolute acceleration is $b + k$. Just as above in the case of rectilinear accelerated motion there then results an apparent force of the absolute value

$$mk = m\frac{4\pi^2 r}{T^2} \quad . \quad . \quad . \quad . \quad (31)$$

which is directed away from the axis. It is the *centrifugal force.*

It is well known that the centrifugal force also plays a part in proving that the earth rotates (Fig. 44). It drives the masses away from the axis of rotation and through this causes, firstly, the flattening of the earth at the poles, and, secondly, the decrease of gravity from the pole towards the equator. We became acquainted with the latter phenomenon above, when we were dealing with the choice of the unit of force (II, 15, p. 45), without going into its cause. According to Newton it is a proof of the earth's rotation. The centrifugal force, acting outwards, acts against gravity and reduces the weight. The

FIG. 43.

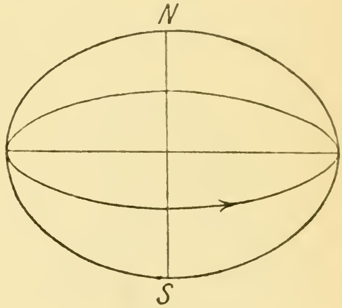


FIG. 44.

decrease of the acceleration g due to gravity has the value $\frac{4\pi^2 a}{T^2}$ at the equator, where a is the earth's radius. If we here insert for a the value given above (III, 3, (23), p. 55), $a = 6{\cdot}37 \,.\, 10$ cms., and for the time of rotation $T = 1$ day $= 24 \,.\, 60 \,.\, 60$ secs. $= 86{,}400$ secs., we get for the difference of the gravitational acceleration at the pole and at the equator the value $3{\cdot}37$ cm./sec.^{2}, which is relatively small compared with 981; this value has to be increased slightly, owing to the flattening of the earth.

According to Newton's doctrine of absolute space these phenomena are positively to be regarded not as due to motion relative to other masses, such as the fixed stars, but as due to absolute rotation in empty space. If the earth were at rest, and if, instead, the whole stellar system were to rotate in the opposite sense once around the earth's axis in 24 hours, then, according to Newton, the centrifugal forces would not occur. The earth would not be flattened and the gravitational force would be just as great at the equator as at the pole. The motion of the heavens, as viewed from the earth, would be exactly the same in both cases. And yet there is to be a definite difference between them ascertainable physically.

The position is brought out perhaps still more clearly in Foucault's pendulum experiment (1850). According to the laws of Newtonian dynamics a pendulum swinging in a plane must permanently maintain its plane of vibration in absolute space if all deflecting forces are excluded. If the pendulum is suspended at the North Pole, the earth rotates, as it were, below it (Fig. 45). Thus the observer on the earth sees a rotation of the plane of oscillation in the reverse sense. If the earth were at rest but the stellar system in rotation, then,

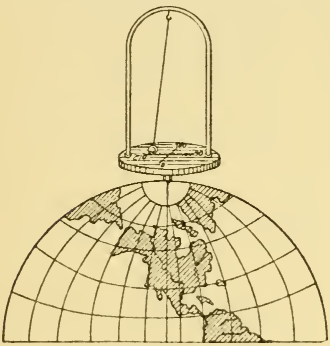


Fig. 45.

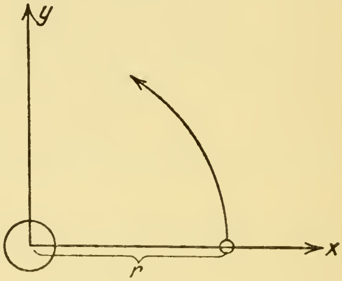


Fig. 46.

according to Newton, the position of the plane of oscillation should not alter with respect to the earth. The fact that it does so again appears to prove the *absolute* rotation of the earth.

We shall consider a further example—the motion of the moon about the earth (Fig. 46). According to Newton the moon would fall on to the earth if it had not an absolute rotation about the latter. Let us imagine a co-ordinate system, with its origin at the centre of the earth, and the xy-plane as that of the moon's orbit, the x-axis always passing through the moon. If this system were to be absolutely at rest, then the moon would be acted on only by the gravitational force towards the centre of the earth, which, by formula (26) on p. 56, has the value

$$K = k\frac{Mm}{r^2}.$$

Thus it would fall to the earth along the x-axis. The fact that it does not do so apparently proves the absolute rotation of the co-ordinate system xy. For this rotation produces a centrifugal force that keeps equilibrium with the force K, and we get

$$\frac{mv^2}{r} = k\frac{Mm}{r^2}.$$

This formula is, of course, nothing other than Kepler's third law. For if we cancel the mass m of the moon on both sides and express v by the period of revolution T, $v = \frac{2\pi r}{T}$, we get

$$\frac{4\pi^2 r}{T^2} = \frac{kM}{r^2}$$

or, by (25) on p. 56,

$$\frac{r^3}{T^2} = \frac{kM}{4\pi^2} = C.$$

An exactly corresponding result holds, of course, for the rotation of the planets about the sun.

These and many other examples show that Newton's doctrine of absolute space rests on very concrete facts. If we run through the sequence of arguments again, we see the following:

The example of the rotating glass of water shows that the relative rotation of the water with respect to the glass is not responsible for the occurrence of centrifugal forces. It might be that greater masses in the neighbourhood, say the whole earth, are the cause. The flattening of the earth, the decrease of gravity at the equator, Foucault's pendulum experiment show that the cause is to be sought outside the earth. But the orbits of all moons and planets likewise exist only through the centrifugal force that maintains equilibrium with gravitation. Finally, we notice the same phenomena in the case of the farthermost double stars, the light from which takes thousands of years to reach us. Thus it seems as if the occurrence of centrifugal forces is universal and cannot be due to inter-actions. Hence nothing remains for us but to assume absolute space as their cause.

Such modes of conclusion have been generally current and regarded as valid since the time of Newton. Only few thinkers have opposed them. We must name among these few above all *Ernst Mach.* In his critical account of mechanics he has analysed the Newtonian conceptions and tested their logical bases. He starts out from the view that mechanical experience can never teach us anything about absolute space. Relative positions and relative motions alone may be ascertained and are hence alone physically real. Hence Newton's proofs of the existence of absolute space must be illusory. As a matter of fact, everything depends on whether it is admitted that if the whole stellar system were to rotate about the earth no flattening, no decrease of gravity at the equator, and so forth, would occur. Mach asserts rightly that such statements go far beyond possible experience. He reproaches Newton very energetically with having become untrue to his principle of allowing only facts to be considered valid. Mach himself has sought to free mechanics from this grievous blemish. He was of the opinion that the inertial forces would have to be regarded as actions of the whole mass of the universe, and sketched the outlines of an altered system of dynamics in which only relative quantities occurred. Yet his attempt could not succeed. In the first place the importance of the relation between inertia and gravitation that expresses itself in the proportionality of weight to mass escaped him. In the second place he was unacquainted with the relativity theory of optical and electro-magnetic phenomena which eliminated the prejudice in favour of absolute time. A knowledge of both these facts was necessary to build up the new mechanics, and the discovery of both was the achievement of Einstein.

CHAPTER IV

THE FUNDAMENTAL LAWS OF OPTICS

1. The Ether

MECHANICS is both historically and logically the foundation of physics, but it is nevertheless only a part of it, and, indeed, a small part. Hitherto to solve the problem of space and time we have made use only of mechanical observations and theories. We must now enquire what the other branches of physical research teach us about it.

It is, above all, the realms of optics, of electricity, and of magnetism that are connected with the problem of space; this is due to the circumstance that light and the electric and magnetic forces traverse empty space. Vessels out of which the air has been pumped are completely transparent for light no matter how high the vacuum. Electric and magnetic forces, too, act across such a vacuum. The light of the sun and the stars reaches us after its passage through empty space. The relationships between the sun-spots and the polar light on the earth and magnetic storms show independently of all theory that electromagnetic actions take place through astronomic space.

The fact that certain physical events propagate themselves through astronomic space led long ago to the hypothesis that space is not empty but is filled with an extremely fine imponderable substance, the ether, which is the carrier or medium of these phenomena. So far as this conception of the ether is still used nowadays it is taken to mean nothing more than empty space associated with certain physical states or "fields." If we were to adopt this abstract conception from the very outset, the majority of the problems that are historically connected with the ether would remain unintelligible. The earlier ether was indeed regarded as a real substance, not only endowed with physical states, but also capable of executing motions.

We shall now describe the development, firstly, of the principles of optics, and, secondly, of those of electrodynamics. This will for the present make us digress a little from the problem of space and time, but will then help us to take it up again fortified with new facts and laws.

2. The Corpuscular and the Undulatory Theory

* I say then that pictures of things and thin shapes are emitted from things off their surfaces . . .

Therefore in like manner idols must be able to scour in a moment of time through space unspeakable . . .

But because we can see with the eyes alone, the consequence is that, to whatever point we turn our sight, there all the several things meet and strike it with their shape and colour . . .

That is what we read in the poem of Titus Lucretius Carus on the Nature of Things (Book 4), that poetic guide to Epicurean philosophy, which was written in the last century before the birth of Christ.

The lines quoted contain a sort of corpuscular theory of light which is elaborated by the imaginative power of the poet but at the same time developed in a true scientific spirit. Yet we can no more call this doctrine a scientific doctrine than we can other ancient speculations about light. There is no sign of an attempt to determine the phenomena quantitatively, the first characteristic of objective effort. Moreover it is particularly difficult to dissociate the subjective sensation of light from the physical phenomenon and to render it measurable.

The science of optics may be dated from the time of Descartes. His Dioptrics (1638) contains the fundamental laws of the propagation of light, the laws of reflection and refraction. The former was already known to the ancients, and the latter had been found experimentally shortly before by Snell (about 1618). Descartes evolved the idea of the ether as the carrier of light, and this was the precursor of *the undulatory theory*. It was already hinted at by Robert Hooke (1667), and was clearly formulated by Christian Huygens (1678). Their great contemporary, Newton, who was somewhat younger, is regarded as the author of the opposing doctrine, *the corpuscular theory*. Before entering on the struggle between these theories we shall explain the nature of each in rough outline.

The corpuscular theory asserts that luminescent bodies send out fine particles that move in accordance with the laws of mechanics and that produce the sensation of light when they strike the eye.

The undulatory theory sets up an analogy between the propagation of light and the motion of waves on the surface of water or sound-waves in air. For this purpose it has to assume the existence of a medium that permeates all trans-

* From Munro's prose translation, published by Deighton, Bell & Co.

parent bodies and that can execute vibrations; this is the *luminiferous ether.* In this process of vibration the individual particles of this substance move only with a pendulum-like motion about their positions of equilibrium. That which moves on as the light-wave is the *state of motion* of the particles and not the particles themselves. Fig. 47 illustrates the process for a series of points that can vibrate up and down. Each of the diagrams drawn vertically below one another corresponds to a moment of time, say, $t = 0, 1, 2, 3 \ldots$ Each individual point executes a vibration vertically. The points all taken together present the aspect of a wave that advances towards the right from moment to moment.

Now there is a significant objection to the undulatory theory.

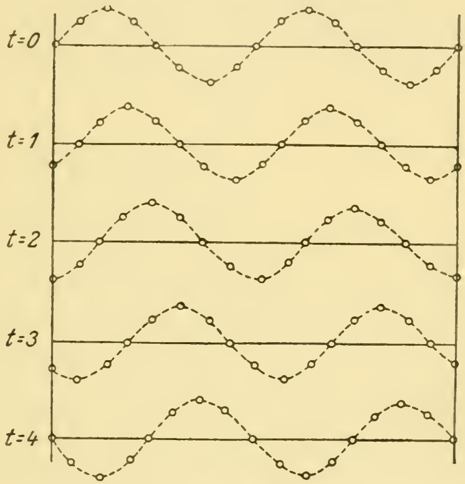


FIG. 47.

It is known that waves run around obstacles. It is easy to see this on every surface of water, and sound waves also "go around corners." On the other hand, a ray of light travels in a straight line. If we interpose a sharp-edged opaque body in its path we get a shadow with a definite outline.

This fact moved Newton to discard the undulatory theory. He did not himself decide in favour of a definite hypothesis but merely established that light is something that moves away from the luminescent body "like ejected particles." But his successors interpreted his opinion as being in favour of the emission theory, and the authority of his name gained the acceptance of this theory for a whole century. Yet, at that time Grimaldi had already discovered (the result was published posthumously in 1665) that light can also "bend round corners."

At the edges of sharp shadows a weak illumination in successive striæ are seen; this phenomenon is called the *diffraction* of light. It was this discovery in particular that made Huygens a zealous pioneer of the undulatory theory. He regarded as the first and most important argument in favour of it the fact that two rays of light cross each other without interfering with each other, just like two trains of water-waves, whereas bundles of emitted particles would necessarily collide or at least disturb each other. Huygens succeeded in explaining the reflection and the refraction of light on the basis of the undulatory theory. He made use of the principle, now called after his name, according to which every point on which the light impinges is to be regarded as the source of a new spherical wave of light. This resulted in a fundamental difference between the emission and the undulatory theory, a difference that later led to the final experimental decision in favour of the latter.

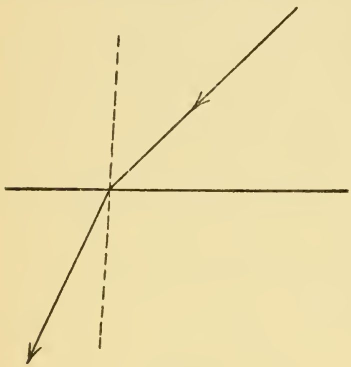


Fig. 48.

It is known that a ray of light which passes through the air and strikes the plane bounding surface of a denser body such as glass or water is bent or refracted so that it is more steeply inclined to the bounding surface (Fig. 48).

The emission theory accounts for this by assuming that the corpuscles of light experience an attraction from the denser medium at the moment they enter into it. Thus they are accelerated by an impulse perpendicular to the bounding surface and hence deflected towards the normal. It follows from this that they must move more rapidly in the denser than in the less dense medium. Huygen's construction on the wave theory depends on just the opposite assumption (Fig. 49). When the light wave strikes the bounding surface it excites elementary waves at every point. If these become transmitted more slowly in the second, denser, medium, then the plane that touches all these spherical waves and that represents the refracted wave according to Huygens, is deflected in the right sense.

Huygens also interpreted the *double refraction* of Iceland spar, discovered by Erasmus Bartholinus in 1669, on the basis of the wave-theory, by assuming that light can propagate itself in the crystal with two different velocities in such a way that the one elementary wave is a sphere, the other a spheroid.

He discovered the remarkable phenomenon that the two rays of light that emerge out of such a piece of fluor spar behave quite differently from other light towards a second piece of fluor spar. If the second crystal is turned about a ray that comes out of the first, then two rays arise out of it which are of varying intensity according to the position of the crystal, and it is possible to make one or other of these rays vanish

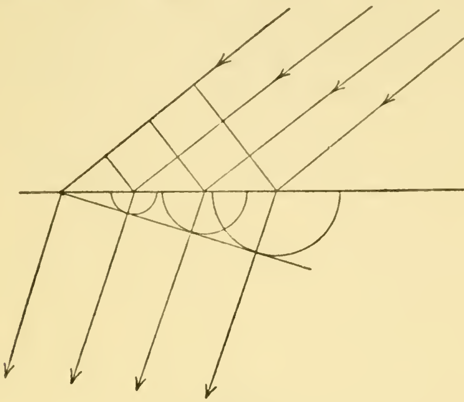


FIG. 49.

entirely (Fig. 50). Newton remarked (1717) that it is to be concluded from this that a ray of light corresponds in symmetry not to a prism with a circular but rather to one with a square cross-section. He interpreted this as evidence against the undulatory theory, for at that time, analogously with sound-waves, only waves of compression and rarefaction were thought of, in which the particles swing "longitudinally" in the direction

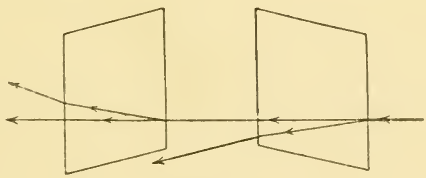


FIG. 50.

of propagation of the wave (Fig. 51), and it is clear that these must have rotatory symmetry about the direction of propagation.

3. THE VELOCITY OF LIGHT

The first determinations of the most important property of light, that which will form the nucleus of our following reflections, namely, the *velocity of light*, were made independently of the controversy between the two hypotheses about the nature of light. The fact that it was enormously great was clear from all observations about the propagation of light. Galilei had endeavoured (1607) to measure it with the aid of lantern signals but without success, for light traverses earthly distances in extremely short fractions of time. Hence the measurement succeeded only when the enormous distances between the heavenly bodies in astronomic space were used.

Olaf Römer observed (1676) that the regular eclipses of Jupiter's satellites occur earlier or later according as the earth

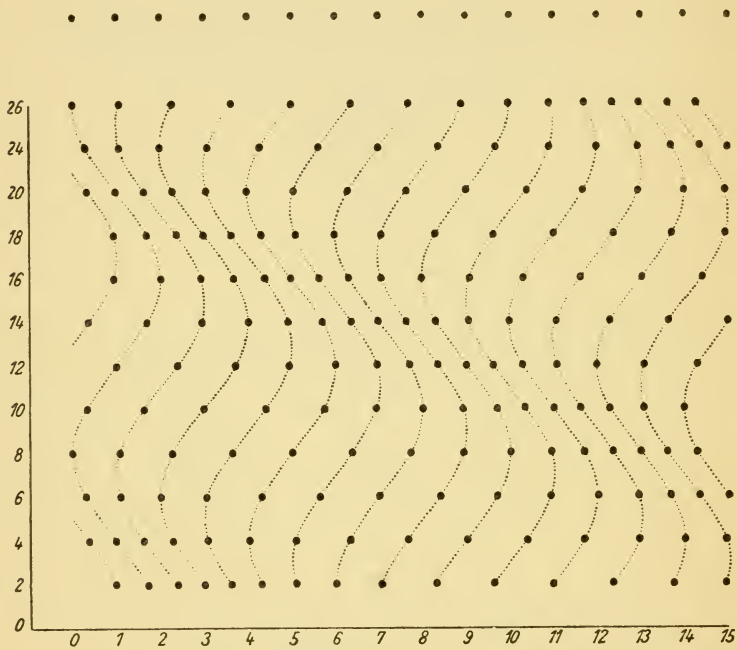


Fig. 51.

is nearer to or farther away from Jupiter (Fig. 52). He interpreted this phenomenon as being caused by the difference of time used by the light to traverse the paths of different lengths, and he calculated the velocity of light on this basis. We shall in future call this velocity c. Its exact value, to which Römer approximated very closely, is

$$c = 300{,}000 \text{ km./sec.} = 3 \cdot 10^{10} \text{ cms. per sec.} \quad (32)$$

James Bradley discovered (1727) another effect of the finite velocity of light, namely, that all fixed stars appear to execute a common annual motion that is evidently a counterpart to the rotation of the earth around the sun. It is very easy to understand how this effect comes about from the point of view of the emission theory. We shall give this interpretation here, but we must remark that it is just this phenomenon that raises certain difficulties for the wave-theory, about which we shall yet have much to say. We know (see III, 7, p. 64) that a motion which is rectilinear and uniform in our system of reference S is so also in another system S′, if the latter executes a motion of translation with respect to S. But the magnitude and the direction of the velocity is different in the two systems. It follows from this that a stream of light corpuscles which, coming from a fixed

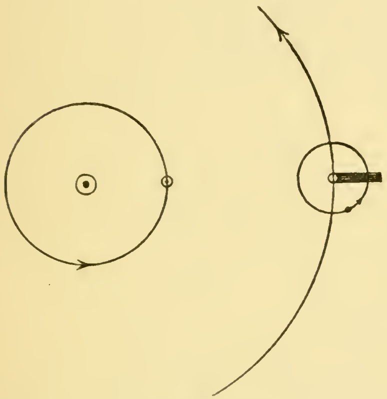


FIG. 52.

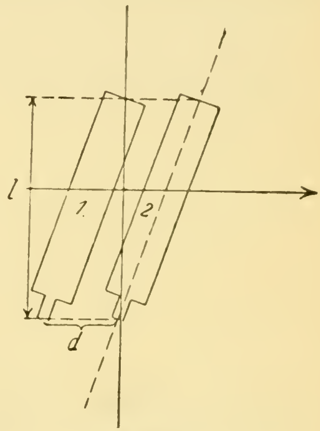


FIG. 53.

star, strike the earth, appear to come from another direction. We shall consider this deflection or *aberration* for the particular case when the light impinges perpendicularly to the motion of the earth (Fig. 53). Let a telescope, on the objective of which a light corpuscle strikes, be in the position 1. Now, whilst the light traverses the length l of the telescope, the earth, and with it the telescope, moves into the position 2 by an amount d. Thus the ray strikes the centre of the eye-piece only when it comes, not from the direction of the telescopic axis, but from a direction lying somewhat behind the earth's motion. Hence the direction in which the telescope aims does not point to the true position of the star, but to a point of the heavens that is displaced forward. The angle of deflection is determined by the ratio $d : l$, and is evidently independent of the length l of the telescope. For if the latter be increased, so also is the time that the light requires to traverse it, and hence also the displacement d of the earth is increased in the same ratio. The two paths l and d, traversed in equal times by the light and the earth, must be in the ratio of the corresponding velocities :

$$\frac{d}{l} = \frac{v}{c}.$$

This ratio, also called the *aberration constant*, will in future be denoted by β :

$$\beta = \frac{v}{c} \quad . \quad . \quad . \quad . \quad (33)$$

It has a very small numerical value, for the velocity of the earth in its orbit about the sun amounts to about $v = 30$ km./sec., whereas the velocity of light, as already mentioned, amounts to 300,000 km./sec. Hence β is of the order 1 : 10,000.

The apparent positions of all the fixed stars are thus always a little displaced in the direction of the earth's motion at that

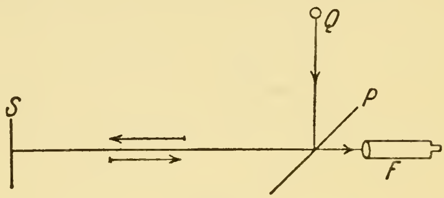


FIG. 54.

moment, and hence describe a small elliptical figure during the annual revolution of the earth around the sun. By measuring this ellipse the ratio β may be found, and since the velocity v of the earth in its orbit is known from astronomic data, the velocity of light c may be determined from it. The result is in good agreement with Römer's measurement.

We shall next anticipate the historical course of events and shall give a note on the earthly measurements of the velocity of light. All that was essential for this was a technical device that allowed the extremely short times required by light to traverse earthly distances of a few kilometres or even only a few metres, to be measured with certainty. Fizeau (1849) and Foucault (1865) used two different methods to carry out these measurements, and confirmed the numerical value of c found by the astronomic method. The details of the process need not be discussed here, particularly as they are to be found in every elementary textbook of physics. We call attention to only one point : in both processes the ray of light is projected from the source Q on to a distant mirror S, where it is reflected and returns to its starting-point (Fig. 54). It traverses the same path twice, and hence it is only the mean velocity during the motion there and back that is measured. The following result, which is important for later considerations, arises from this circumstance: if we suppose that the velocity of light is not the same in both directions, because the earth itself is in motion—we shall discuss this point later (IV, 9, p. 110)—then this influence will be wholly or partially cancelled in the motion to and fro. Therefore, in view of the smallness of the velocity of the earth in comparison with that of light, we need take no account of the earth's motion in these measurements in practice.

The measurements of the velocity of light were later repeated with improved apparatus, and a considerable degree of accuracy was obtained. Nowadays they can be carried out in a room of moderate length. The result is the value (32) given above. Foucault's method also allowed the velocity of light to be measured in water. It was found to be *smaller* than that obtained for air. This gave a definite decision on one of the most important points under dispute between the emission and the undulatory theory in favour of the latter. This occurred, indeed, at a time when the triumph of the wave-theory had already long been assured on other grounds.

4. Fundamental Conceptions of the Wave Theory Interference

Newton's greatest achievement in optics was the resolution of white light into its coloured constituents by means of a prism and the exact examination of the spectrum, which led him to the conviction that the individual spectral colours were the indivisible constituents of light. He is the founder of the theory of colour, the physical content of which is still fully valid to-day—in spite of Goethe's attacks. The power of Newton's discoveries paralysed the free thought of the succeeding generations. His refusal to accept the undulatory theory blocked the road to its acceptance for well nigh a century. Nevertheless it found isolated supporters such as, for example, the great mathematician Leonhard Euler in the 18th century.

The revival of the wave-theory is due to the works of Thomas Young (1802), who adduced the principle of *interference* to explain the coloured rings and fringes which even Newton had observed in thin layers of transparent substances. We shall at this stage deal somewhat in detail with the phenomenon of interference because it plays a decisive part in all finer optical measurements, particularly in researches that constitute the foundations of the theory of relativity.

We explained the nature of waves above: it consists in the individual particles of a body executing periodic oscillations about their positions of equilibrium, whereby the momentary position or the phase of the motion is different for neighbouring particles and moves forward with constant velocity. The time that a definite particle requires for one vibration, to and fro, is called the *time of vibration* or the *period*, and is denoted by T. The *number of vibrations* in one second or the *frequency* is designated by ν. Since the time of a vibration multiplied by their number per second must give exactly one second, we must have $\nu T = 1$, thus

$$\nu = \frac{1}{T} \quad \text{or} \quad T = \frac{1}{\nu} \qquad . \quad . \quad . \quad (34)$$

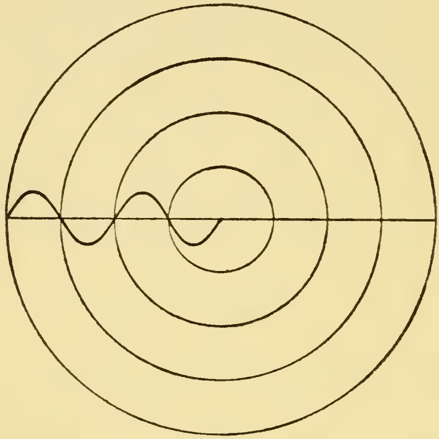


FIG. 55.

Instead of vibration number or frequency we often say "colour," because a light wave of a definite frequency produces a definite sensation of colour in the eye. We shall not enter into the complicated question as to how "physical colours," as we may call the great manifold of psychological impressions of colour, come about through the conjoined action of simple periodic vibrations. The waves that start out from a small source of light have the form of spheres. This means that particles on a sphere drawn around the source as centre are always in the same state of vibration or they are of equal "phase" (Fig. 55). By means of refraction or other influences a part of such a spherical wave may be deformed so that the surfaces of equal phase or the wave surfaces have some other form. The simplest wave-surface is evidently the plane, and it is clear that a sufficiently small piece of any arbitrary wave surface, hence even of a spherical surface, may always be regarded in approximations as plane. Hence we consider in particular the propagation of plane waves (Fig. 56). The direction that is perpendicular to the planes of the waves, that is, the normal to the waves, is at the same time the direction of propagation. It is clearly sufficient to consider the state of vibration along a straight line parallel to this direction.

Whether the vibration of the individual particle occurs parallel or perpendicularly to the direction of propagation, whether it is longitudinal or transverse, will be left quite open at this stage. In the figures we shall always draw wave lines and call the greatest displacements upwards and downwards crests and hollows.

The distance from one crest to the next is called *wave-length* and is designated by λ. The distance between the successive or any two consecutive planes in the same phase is obviously exactly the same amount.

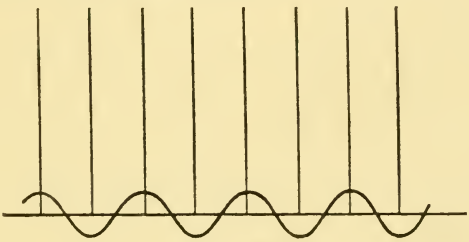


FIG. 56.

During a vibration of a definite particle to and fro, the duration of which is T, the whole wave moves forward exactly a wave-length λ (Fig. 47, p. 77). Since the velocity in every motion is equal to the ratio of the path traversed to the time required to do so, the wave velocity c is equal to the ratio of the wave-length to the time of vibration:

$$c = \frac{\lambda}{T} \text{ or } c = \lambda\nu \quad . \quad . \quad . \quad (35)$$

If a wave enters from one medium into another, say, from air into glass, the time rhythm of the vibrations is, of course, carried over the bounding surface, that is, T (or ν) remains the same. On the other hand, the velocity c and hence, on account of formula (35), also the wave-length λ changes. Thus all methods of measuring λ may serve to compare the velocity of light in various substances or under various circumstances. We shall make use of this fact later.

We are now in a position to understand the nature of interference phenomena, the discovery of which helped the wave-theory to prevail. Interference may be described by the paradoxical words: light added to light does not necessarily give intensified light, but may become extinguished.

The reason for this is that, according to the wave-theory, light is no stream of material particles but a state of motion. Two vibration impulses that occur together may, however, destroy the motion just like two people who wish to do contrary things impede each other and produce nothing. Let us imagine two trains of waves that intersect. This phenomenon can be conveniently observed if we look from a hillock down into a lake in which the waves caused by two ships meet (Fig. 57). These two wave-systems interpenetrate without disturbing each other. In the region at which both exist simultaneously a complicated motion arises, but so soon as the one

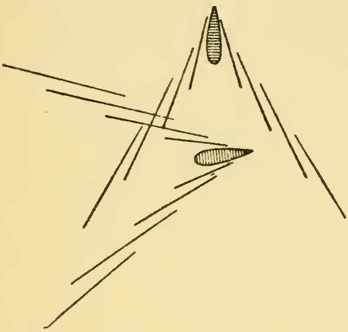


Fig. 57.

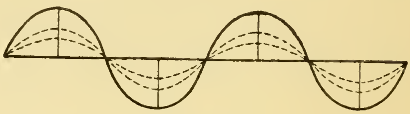


Fig. 58.

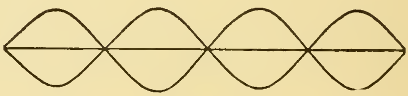


Fig. 59.

wave has passed through the other, it continues further as if nothing had happened to it. If we fix our attention on one particular vibrating particle, we see that it experiences independent impulses from both waves. Hence its displacement at any point is simply the sum of the displacements that it would have under the influence of the individual waves. Two wave-motions are said to superpose without disturbing each other. From this it follows that at points where crest and crest and also at points where hollow and hollow meet, where two equal waves encounter each other, the elevations and the depressions are twice as great (Fig. 58). But at points where crest and hollow meet the impulses destroy each other and no displacement occurs at all (Fig. 59).

If we wish to observe interference of light it will not do simply to take two sources of light and to allow the trains of waves emerging from them to interpenetrate. No observable interference phenomenon occurs through this, because the actual light waves are no absolutely regular waves. Rather, the state of vibration suddenly alters in a striking way after a series of regular vibrations have occurred, corresponding to the accidental phenomena that occur during the emission of light in the source. These irregular changes effect a corresponding fluctuation of the interference phenomena, which occurs much too quickly for the eye to follow it, and hence we see only uniform light.

To obtain observable interferences we must resolve a ray of light by artificial means, by reflection or refraction, into two rays, and afterwards make them come together again. Then the irregularities of the vibrations in both rays occur in exactly the same time rhythm, and hence it follows that the interference phenomena do not fluctuate in space, but remain fixed. Wherever the waves strengthen or extinguish each other at a certain moment, they do so at every moment. If we direct the eye, armed with a magnifying glass or a telescope, at such a point we see fringes or rings, provided we use light of one colour (monochromatic light), such as is approximately emitted by a Bunsen flame coloured yellow by common salt. In ordinary light which is composed of many colours the interference spots corresponding to the various wave-lengths do not exactly coincide. At one point red is intensified, say, and blue is extinguished, at other points other colours occur, and hence spots and fringes arise with wonderful colourings. It would, however, take us away from our path of enquiry to pursue these interesting phenomena further.

The simplest arrangements for producing interferences were given by Fresnel (1822), an investigator whose works have furnished the foundation for the theory of light which has remained unattacked up to the present day. We shall often meet with his name in the sequel. That time, the first decades of the nineteenth century, must in many respects have resembled our own. Just as nowadays through the discovery of radioactivity and the associated phenomena of radiation, through the enunciation of the physical principle of relativity and of the doctrine of quanta, our knowledge of physical nature is undergoing a stupendous process of deepening and enlargement, which seems to the beholders a complete revolution of all conceptions, so, a hundred years ago, the thousands of individual observations, theoretical experiments, physical or metaphycisal speculations coalesced for the first time into complete and uniform ideas and theories, the application of which at once suggested an undreamed-of abundance of new observations and experiments. At that time Lagrange's

" Analytical Mechanics " and Laplace's " Celestial Mechanics " appeared, the two works that brought Newton's ideas to their conclusion. From them there was developed on the one hand, by Navier, Poisson, Cauchy, and Green, the mechanics of deformable bodies and the theory of fluids and elastic substances; on the other hand, by the works of Young, Fresnel, Arago, Malus, and Brewster, the theory of light. At the same time began the era of electromagnetic discoveries, of which we shall speak later. At that time physical research was almost entirely in the hands of the French, Italians, and English. Nowadays *all* educated nations participate, and the authors of the great revolutionary theories of relativity and quanta, Einstein and Planck, are German.

Fresnel allowed a ray of light to be reflected at two mirrors,

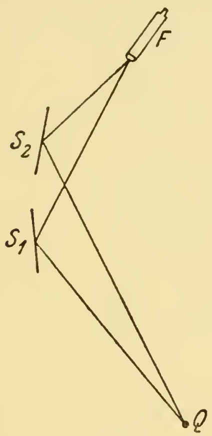


FIG. 60.

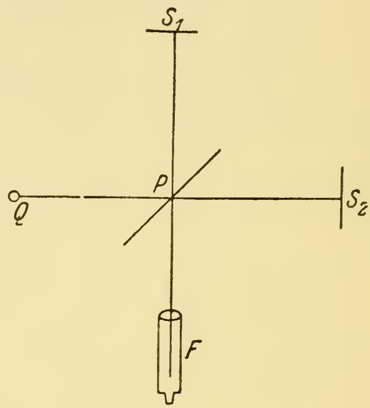


FIG. 61.

S_1 and S_2 (Fig. 60), slightly inclined to each other. At the points where they meet, these two reflected rays give interference fringes that can be seen with a magnifying glass. Similar arrangements of apparatus have been given in great numbers. We shall here enter only into a field which is important for our purpose, namely, that of the experimental methods of measuring minute changes in the velocity of light. The apparatus used is called an *interferometer*. It depends on the fact that the wave-length alters proportionately with the velocity of the light, and hence the interferences are displaced. An example of an apparatus of this kind is the interferometer of Michelson. It consists in the main (Fig. 61) of a glass plate P that is slightly silvered so as to allow one half of the light from the source Q to pass through while the other half is reflected.

These two component rays travel on to two mirrors S_1 and S_2, where they are reflected and again encounter the semi-transparent glass plate P, which again resolves them, sending one half of each ray into the observing telescope F. If the two paths PS_1 and PS_2 are exactly equal the two component rays arrive at the telescope in the same phase of vibration and recombine to form the original light again. But if the path of the first ray be lengthened by displacing the mirror S_1, then the crests and hollows of the two trains of waves no longer coincide when the rays are recombined at F, but are displaced with respect to each other and weaken each other more or less. If the mirror S_1 is moved slowly we see alternate patches of light and darkness in the telescope F. The distance of the positions of S_1 for two successive dark fields is exactly equal to the wave-length of the light. In this way Michelson has made measurements of wave-length that exceed almost all other physical measurements in accuracy. This is done by counting the changes of light and darkness during a considerable shift of the mirror S_1, which comprises many thousands of wave-lengths. The error of observation of an individual wave-length then becomes just as many thousand times smaller.

We have here to give several numerical data. By the above method it is found that the wave-length of the yellow light that is sent out by a Bunsen flame coloured with common salt (NaCl), and the source of which is sodium atoms, is about $\frac{6}{10,000}$ mm. $= 6 \cdot 10^{-5}$ cms. *in vacuo*. All visible light lies within the small region of wave-lengths stretching from about 4.10^{-5} (violet) to 8.10^{-5} cms. (red). Thus in the language of acoustics this comprises one octave; that is, it is the region between one wave and another that is twice as long. From formula (35) there then follows for the vibration number of yellow sodium light the stupendous number $\nu = \frac{c}{\lambda} = \frac{3 \cdot 10^{10}}{6 \cdot 10^{-5}} = 5 \cdot 10^{14}$ or 500 billion vibrations per second. The most rapid acoustic vibrations that are still audible vibrate only about 50,000 times per second.

The astonishing accuracy of optical methods of measurement rests on the multiplication of the individual wave-lengths used in interferometric measurements. For example, it allows us to ascertain that the velocity of light in a gas likewise alters if there is a very small change of pressure or temperature (due, say, to the apparatus being touched by the hand). To show this the gas is passed into a cylinder between the glass plate P and the mirror S_1. It is then seen that for even the slightest increase of pressure the interference changes, light fields being converted into darkness and *vice versa*.

For the rest, we must remark that in the interferometer we do not simply see a light on a dark field in the telescope, but a system of light and dark rings. This is due to the fact that the two rays are not exactly parallel and the waves are not exactly plane. The separate parts of the two rays have thus to traverse paths of different length. We shall not, however, enter into the geometric details, but mention this circumstance only because it is customary to speak of interference bands or fringes.

We shall meet with Michelson's interferometer again when we have to decide the question as to whether the earth's motion influences the velocity of light.

5. Polarisation and Transversality of Light Waves

Although interference phenomena allow scarcely any interpretation other than that of the wave-theory, its general recognition was impeded by two difficulties which, as we saw above, were regarded by Newton as being decisive contradictions to it: firstly, the general rectilinear propagation of light (that is, except for trifling diffraction phenomena); secondly, the explanation of *polarization phenomena*. The first difficulty became removed when the wave theory itself was worked out more exactly; for it was found that waves do, indeed, "bend round corners," but only in regions that are of the order of magnitude of the wave-length. As this is very small in the case of light, our ordinary unrefined vision receives the impression of sharp shadows and rectilinearly bounded rays. Only minute observation is able to detect the interference fringes of diffracted light along the edges of the shadow. The merit of elaborating the theory of diffraction is due to Fresnel, later Kirchhoff, (1882), and, more recently, Sommerfeld (1895). They have deduced the finer phenomena mathematically and have defined the limits within which the conception *ray of light* may be applied.

The second difficulty concerned the phenomena due to the polarization of light.

When we earlier spoke of waves we always had in mind longitudinal waves such as are known in the case of sound. For a sound wave consists of rhythmical condensations and rarefactions, during which the individual particles of air move to and fro in the direction of propagation of the wave. Transversal waves were, indeed, also known; for example, the waves on a surface of water, or the vibrations of a stretched string, in which the particles vibrate perpendicularly to the direction of propagation of the wave. But in this case we are dealing not with waves that advance in the *interior* of a substance but in part with phenomena on the upper surface (water waves) and in part with motions of whole configurations (vibration of strings). Observations or theories about the propagation of waves in elastic solid bodies were not yet known. This accounts for the circumstance, which appears strange to us, that it was so long before optical waves were recognized as transverse vibrations. In fact, the remarkable instance occurred that the impulse for the development of the mechanics of coarse-grained solid elastic bodies was given by experiments and conceptions derived from the dynamics of the imponderable and intangible ether.

We explained above (p. 79) what constitutes the nature of polarization. The two rays that emerge out of a doubly refracting crystal of calcite do not behave like ordinary light when they pass through a second such crystal, that is, they

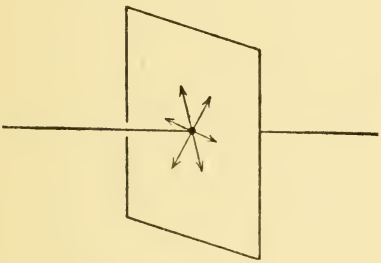


FIG. 62.

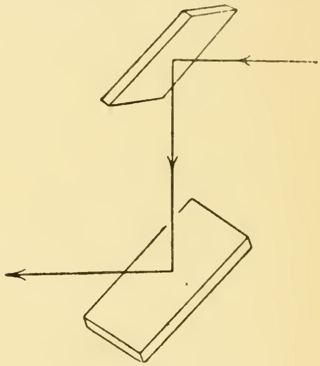


FIG. 63.

do not again resolve into the equally intense rays, but into two of unequal intensity, one of which may under certain circumstances vanish entirely.

In ordinary "white" light the various directions within the plane of a wave are of equal value or equivalent (Fig. 62). In polarized light this is obviously no longer the case. Malus discovered (1808) that polarization is not a peculiarity of the light that has passed through a doubly refracting crystal, but may also be produced by simple reflection. He showed that light which has been reflected from a mirror at a definite angle is reflected by a second mirror to a varying degree, if the latter mirror is turned about the incident ray (Fig. 63). The plane perpendicular to the surface of the mirror and containing the incident and the reflected ray is called the incident plane. The reflected ray is then said to be polarized in the incident plane; this implies no more than that it behaves differently towards a second mirror according to the position of the second incident plane to the first. If these mirrors are perpendicular to each other no reflection at all occurs at the second mirror.

The two rays that emerge out of a crystal of calcite are polarized perpendicularly to each other. If we allow them both to fall on to a mirror at an appropriate angle the one is completely extinguished just when the other is reflected to its full amount.

Fresnel and Arago made the decisive experiment (1816) when they attempted to make two such rays, polarized perpendicularly to each other, interfere. They did not succeed. Fresnel and also Young then drew the inference (1817) that light vibrations must be transversal.

As a matter of fact this deduction makes the peculiar behaviour of polarized light intelligible at once. The vibrations of the ether particles do not occur in the direction of

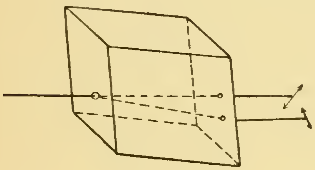


Fig. 64.

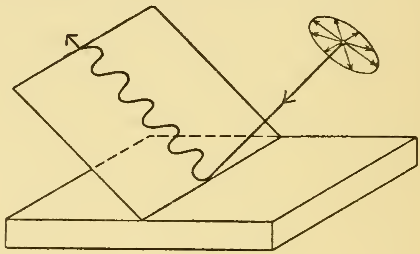


Fig. 65.

propagation but perpendicularly to it, that is, in the plane of the wave (Fig. 62). But every motion of a point in a plane may be regarded as composed of two motions in two directions perpendicular to each other. In dealing with the kinematics of a point we saw that its motion is determined uniquely when its rectangular co-ordinates, which vary with the time, are given. Now a doubly refracting crystal clearly has the property of transmitting the vibrations of light in it at different velocities in two mutually perpendicular directions. Hence, by Huyghens' principle, when these vibrations enter the crystal they will be deflected to different extents, or refracted differently, that is, they will be separated in space. Each of the emergent rays then consists only of vibrations that take place in a certain plane that passes through the direction of the ray, and the planes belonging to these two rays are mutually perpendicular (Fig. 64). Two such vibrations clearly cannot influence each other: they cannot interfere. Now, if a polarized ray enters into a second crystal it is transmitted without being weakened only if its direction of vibration is in just the right position with respect to the crystal, being just that in which this vibration can propagate itself. In all other positions the ray is weakened : in the perpendicular position it is not transmitted at all.

Similar conditions obtain in reflection. If this occurs at the appropriate angle, then only one of the two vibrations, the one parallel and perpendicular to the incident plane, is reflected ; the other penetrates into the mirror and is absorbed (Fig. 65). Whether the reflected vibration is that which takes place in the incident plane or perpendicularly to it cannot, of course, be ascertained. (In Fig. 65 the latter is assumed to be the case.) But this question of the position of vibration with respect to the plane or incidence or the direction of polarization has given rise to elaborate researches, theories and discussions, as we shall presently see.

6. The Ether as an Elastic Solid

After the transversality of light waves had been proved in this way and by numerous experiments, there arose in Fresnel's mind the vision of a future *dynamical theory of light*, which was to derive optical phenomena from the properties of the ether and the forces acting in it, in conformity with the method of mechanics. The ether was necessarily a kind of elastic solid, for it is only in such a substance that mechanical transverse waves can occur. But in Fresnel's time the mathematical *theory of elasticity of solid bodies* had not yet been developed. Possibly he also thought from the outset that the analogy of the ether with material substances was not to be carried too far. At any rate he preferred to investigate the laws of the propagation of light empirically and to interpret them by means of the idea of transversal waves. Above all it was to be expected that the optical phenomena in crystals would shed light on the behaviour of the ether. Fresnel's work in this field is to be ranked among the most beautiful achievements of systematic physics, both in experimental as well as in theoretical respects. Yet we must not digress too far in pursuing details, but must keep in view our problem: how is the ether constituted ?

Fresnel's results appeared to confirm the analogy of light waves with elastic waves. This gave a powerful stimulus to the working out of the theory of elasticity, which had already been begun by Navier (1821) and Cauchy (1822), and to which Poisson (1828) devoted his attention. Cauchy then at once applied the laws derived from elastic waves to optics (1829). We shall try to give an idea of the content of this ether theory, The difficulty involved is that the proper and adequate means of describing changes in continuous deformable bodies is the method of *differential equations.* Since we do not wish to take these as known, all that can be done is to illustrate them by a simple example, and then to add at the end that in the general case the same holds, but in a more complicated way. The non-mathematical reader may perhaps then get a rough idea of what is involved. It will not, however, give him a real estimate of the power of achievement expressed in the physical pictures and in the mathematical methods used. We are fully conscious of the impossibility of entirely satisfying the non-mathematician, but we cannot refrain from attempting to illustrate the mechanics of continua, because all subsequent theories, not only of the elastic ether, but also electrodynamics in all its ramifications, and, above all, Einstein's theory of gravitation, are built up on these conceptions.

A very thin stretched string is in a certain sense a one-dimensional elastic configuration. We shall use it to develop the theory of elasticity. To link up with ordinary mechanics, which deals only with individual rigid solids, we suppose the string to be not continuous but of an atomistic structure, as it were. Let it consist of a series of equal small bodies that are arranged in a line at equal distances from each other (Fig. 66). The particles are to possess inertial mass and each is to exert forces on its two neighbours : these forces are to be such that they resist both an increase and a decrease of the distance between these particles. If we wish to have a concrete picture of such forces, we need only think of small spiral springs that are fixed between the particles. These resist compression as well as extension. But such a representation must not be taken literally. Forces of this kind, indeed, constitute just the essential phenomena of elasticity.

Now if the first particle is displaced a little in the longitudinal or in the transverse direction, it immediately acts on the second particle ; the latter in its turn passes the action on to the next, and so forth. The disturbance of the equilibrium of the first particle thus passes along the whole series like a short wave and finally also reaches the last particle. This process does not, however, occur infinitely quickly. At every particle a small fraction of time is lost because the particle, owing to its inertia, does not instantaneously respond to the impulse. For the force does not produce an instantaneous displacement but an acceleration, that is, a change of velocity during a small interval of time, and the change of velocity again requires time to produce its displacement. Only when this displacement has reached its full value does the force act to its full extent on the next particle, and from then onwards the process repeats itself with a loss of time that is dependent on the mass of the particles. If the force that arises through the displacement of the first particle were to influence the last particle of the series directly, the action would occur instantaneously, According to Newton's theory of gravitation this is actually supposed to be the case in the mutual attraction of the heavenly bodies. The force with whi h one acts on the other is always directed at the point momentarily occupied by the other and is determined by the distance separating these points at that moment. Newtonian gravitation is said to be an *action at a distance*, for it acts between points at a distance although there is no intervening medium to convey this action.

In contrast with this our series of equidistant points is the simplest model of *contiguous action* or *action by contact.* For the action exerted by the first point on the last is transferred by the intervening masses, and hence does not occur instantaneously but with a loss of time. The force exerted by a

Fig. 66.

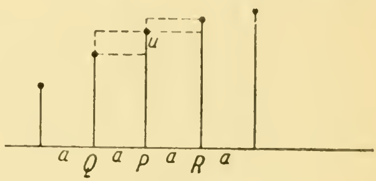


Fig. 67.

particle on its neighbours is certainly still imagined as an action at a distance, although only at a small distance. We may, however, suppose these distances between the particles to grow smaller and smaller, their number becoming correspondingly greater and greater, but in such a way that their total mass remains the same. The chain or particles then passes over into the limiting conception of a *material continuum.* The forces act between infinitely near particles, and the laws of motion assume the form of differential equations. They express mathematically the physical conception of contiguous action.

We shall pursue this limiting process of the laws of motion a little further for the case of our chain of mass-particles. Let us consider purely transverse displacements (Fig. 67). In the theory of elasticity it is assumed that a particle P is pulled back by its neighbour Q more strongly in proportion to the amount that Q is displaced transversely beyond P. If u is the excess of the transverse displacement of P beyond that of Q, and if a is the original distance between the particles along the straight line, then the restoring force is to be proportional to the ratio $\frac{u}{a} = d$, which is called the *deformation.* We set

$$K = p \cdot \frac{u}{a} = pd$$

where p is a constant number which is clearly equal to the force if the deformation d be chosen equal to 1. p is called the *elastic constant.*

Now the same particle likewise experiences a force $K' = p\frac{u'}{a} = pd'$ from its other neighbour R. But except in the singular case, when the deflection of P is exactly a maximum, the particle R will be more strongly displaced than P, and hence will not pull back the latter, but will tend to increase its displacement. Thus K′ will work against K.

The resultant force on the particle P is the difference of these forces

$$K - K' = p(d - d').$$

And this determines the motion of P according to the fundamental dynamical formula: mass *times* acceleration *equals* force

$$= mb = K - K' = p(d - d').$$

Now let us suppose the number of particles to be increased more and more, but their mass to be decreased in the same ratio so that the mass per unit of length always retains the same value. Let there be n particles in each unit of length, so that $n \cdot a = 1$, that is, $n = \frac{1}{a}$. The mass per unit of length is $mn = \frac{m}{a}$. This linear quantity is called the *density of mass,* and is designated by ρ. By dividing the above equation by a, we get

$$\frac{m}{a}b = \rho b = \frac{K - K'}{a} = p\frac{d - d'}{a}$$

and here we have configurations quite similar to those which occurred in the definitions of the conceptions, velocity and acceleration. For just as the velocity was the ratio of the path x to the time t, $v = \frac{x}{t}$, wherein the time t is to be considered quite short for an accelerated motion, so we have here the deformation $d = \frac{u}{a}$, the ratio of the relative displacement to the original distance, wherein the latter is to be regarded as extremely small. Just as the acceleration was before defined as the ratio of the change of velocity to the time, $b = \frac{w}{t} = \frac{v - v'}{t}$, so we have here the quantity $f = \frac{d - d'}{a}$, which measures in a fully analogous manner the change of the deformation from point to point.

Exactly as the velocity v and the acceleration b retain their sense and their finite values for time intervals that are arbitrarily small, so the quantities d and f retain their meaning and finite values no matter how small the distance a becomes. All these are so-called *differential co-efficients*, $v = \frac{x}{t}$ and $d = \frac{u}{a}$ being such of the first order, $b = \frac{v - v'}{t}$ and $f = \frac{d - d'}{a}$ such of the second order.

Thus the equation of motion becomes a differential equation of the second order,

$$\rho b = pf \quad . \quad . \quad . \quad . \quad (36)$$

both with respect to the time change as well as to the space change of the event. *All* laws of contiguous action in theoretical physics are of this type. If, for example, we are dealing with elastic bodies that are extended in all directions, we get two analogously formed members for the other two space dimensions. Moreover, precisely similar laws hold in the theory of electric and magnetic events. Finally, the gravitational theory of Einstein has also been brought into such a form.

We have here yet to remark that laws of action at a distance may be written in a form similar to that of formulæ for contiguous action. For instance, if we strike out the member ρb in our equation (36), that is, if we assume that the density of mass is extremely small, then a displacement of the first particle will at the same moment call up a force acting on the last particle, because the inertia of the intervening members has dropped out. Thus we really have the transmission of a force with infinite velocity, a true action at a distance. Nevertheless the law $pf = o$ appears in the form of a differential equation, as a contiguous action. Such *laws of pseudo-contiguous action* will be met with in the theory of electricity and magnetism, where they have really prepared the way for the true laws of contiguous action. The essential factor in the latter is the inertial member that is responsible for the finite velocity of transmission of disturbances of equilibrium, that is, the generation of waves.

Two quantities occur in the law (36) that determine the physical character of the substance: the mass per unit of volume or the density ρ, and the elastic constant p. If we write $b = \frac{p}{\rho} f$, we see that for a given deformation, that is for a given f, the acceleration becomes greater in proportion as p becomes greater and ρ becomes smaller. Thus p is just a measure of the elastic rigidity of the substance, and ρ is a measure of the inertial mass, and it is clear that an increase of the rigidity accelerates the motion, an increase of the inertia retards it. Accordingly the velocity c of a wave will depend only on the ratio $\frac{p}{\rho}$. For the more quickly the wave travels the greater are the accelerations of the individual particles of the substance. The exact law for this relationship is found by the following considerations.

Each individual point-mass executes a simple periodic motion of the kind which we investigated earlier (II, 11, p. 34). We showed there that in it the acceleration is connected with the deflection x according to formula (11)

$$b = (2\pi\nu)^2 x$$

where ν is the number of vibrations per second. If we insert in place of ν the time of vibration according to formula (34), p. 84, $T = \frac{1}{\nu}$, we get

$$b = \left(\frac{2\pi}{T}\right)^2 x.$$

The same argument that has here been used for succession in time may also be applied for succession in space, and must lead to relations that correspond entirely. We have simply to replace the quantity f (the second space-coefficient) and the time of vibration T (the period in time succession) by the wave-length λ (the " space-period "). We thus get the formula

$$f = \left(\frac{2\pi}{\lambda}\right)^2 x.$$

If we form the quotient of the two expressions for b and f the factor $(2\pi)^2 x$ cancels out, and there remains

$$\frac{b}{f} = \frac{\lambda^2}{T^2}.$$

Now, on the one hand we have by formula (35), p. 85, that $\frac{\lambda}{T} = c$, on the other hand by (36), p. 97, that $\frac{b}{f} = \frac{p}{\rho}$. Hence it follows that

$$c^2 = \frac{p}{\rho} \text{ or } c = \sqrt{\frac{p}{\rho}} \quad . \quad . \quad . \quad (37)$$

This relation holds for all bodies, no matter whether they be gaseous, liquid, or solid. But there is the following difference.

In *liquids and gases* there is no elastic resistance to the lateral displacement of the particles, but only to the change of volume. Hence *only longitudinal waves* can propagate themselves in such substances, their velocities being determined according to formula (37), by the elastic constant p which is decisive in such changes of volume.

On the other hand, in *solid bodies*, on account of the elastic rigidity which opposes lateral displacements, *three waves, one longitudinal and two transverse*, with different velocities, can transmit themselves in each direction. This is due to the fact that the compressions and rarefactions of the longitudinal waves involve an elastic constant p, which is different from that which comes into action for the lateral distortions due to the transverse vibrations.

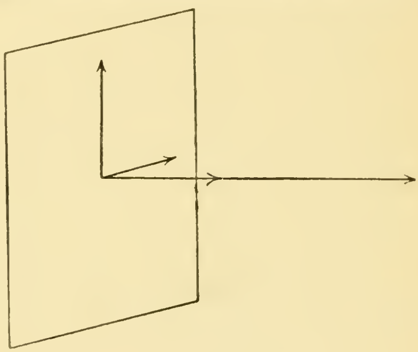


FIG. 68.

Moreover, in non-crystalline bodies the two transverse waves have, indeed, different directions of vibration, perpendicular to each other, but they have the same velocity C_t; the longitudinal wave has a different velocity C_l (Fig. 68).

All these facts allow themselves to be confirmed by experiments on acoustic waves in solid bodies.

We now return to the starting-point of our reflections, namely, to the elastic theory of light.

This consists in identifying the ether as the carrier of light vibrations with a solid elastic body. The light waves are then, as it were, to be sound waves in this hypothetical medium.

Now, what properties are to be ascribed to this elastic ether? In the first place the enormous velocity of propagation c

requires either that the elastic ridigity p be very great or that the density of mass ρ be very small, or that both conditions hold simultaneously. But since the velocity of light is different in different substances, either the ether within a material body must be condensed or its elasticity must be changed or again both may be true simultaneously. We see that different courses are here open to us. The number of possibilities is still further increased by the fact that, as we saw above (IV, 5, p. 93), experiment cannot decide whether the vibrations of polarized light are parallel or perpendicular to the plane of polarization (the incident plane of the polarizing mirror).

Corresponding to the indefinite nature of the problem we also find an innumerable number of different theories of the elastic ether in history. We have already mentioned the most important authors ; to the mathematicians Poisson, Fresnel, Cauchy, and Green there is to be added for the first time a German physicist of note, Franz Neumann, who became the teacher of the generation of physicists of his own country characterized by Helmholtz, Kirchhoff and Clausius.

Nowadays we feel surprise at the amount of ingenuity and labour that was expended on the problem of comprehending optical phenomena in their totality as motions of an elastic ether having the same properties as those possessed by material elastic solids. It seems to us that the principle which defines explanations as the reduction of unknown things to known things was overstrained. For we now know that the nature of elastic solids is by no means simple and is certainly not to be regarded as known. The physics of the ether has shown itself to be simpler and more easily intelligible than the physics of matter, and modern research is directed at tracing back the constitution of matter, as a secondary phenomenon, to the properties of the fields of force that represent the remains of the ether of the older theory. This change in the programme of science is not least due to the failures attending the attempts to build up a logical theory of the elastic ether.

One objection to the latter theory, which seems of importance, is that an all-pervading ether (which fills astronomic space) of great rigidity, which it must have as the carrier of the rapid vibrations of light, would necessarily offer resistance to the motion of heavenly bodies, in particular, to that of the planets. Astronomy has never detected departures from Newton's laws of motion that would point to such a resistance. Stokes (1845) partly disposed of this objection by remarking that the conception of solidity of a body is in itself really something relative and depends on the relation of the deforming forces to time. A piece of pitch—sealing-wax and glass behave similarly—when struck with a hammer splits cleanly. But if it is loaded with a weight, the latter sinks gradually, although perhaps only slowly, into the pitch as if pitch were a very viscous fluid. Now, the forces that occur in light vibrations change stupendously quickly (600 billion times per sec.) compared with the relatively slow processes that occur in planetary motions in the course of time ; the ratio of these forces is much more extreme than that of the hammer blow to the superimposed weight. Therefore the ether may function for light as an elastic solid and yet give way completely to the motion of the planets.

Now, even if we wish to content ourselves with this astronomic space filled with pitch, serious difficulties arise out of the laws of the propagation of light themselves. Above all, we have to take into account that in elastic solids a longitudinal wave always occurs conjointly with the two transversal waves. If we follow out the refraction of a wave at the boundary of two media, and if we assume that the wave vibrates purely transversally in the first medium, then a longitudinal wave must arise in the second medium together with the transverse wave. All attempts to escape this consequence of the theory by making more or less arbitrary changes have been doomed to failure. Extraordinary hypotheses were suggested, such as that the ether opposes to compression an infinitely small or an infinitely great resistance compared with its rigidity towards transversal distortions. In the former case the longitudinal waves would travel forward infinitely slowly, in the latter infinitely quickly, and would at any rate not manifest themselves as light. A physicist, MacCullagh (1839), went so far as to construct an ether that departed altogether from the model of elastic bodies. For whereas in these the particles oppose a resistance to every change of their distance from each other, but follow pure twists without resistance, MacCullagh's ether was to behave in just the contrary way. We cannot here enter into the theory. However strange it may appear, it is nevertheless of importance as the fore-runner of the electromagnetic theory of light. It leads to almost the same formulæ as the latter, and is actually able to give an account of optical phenomena that is to a considerable degree correct. But its weakness is that it disclosed no relationship between optical phenomena and other physical phenomena. It is clear that by means of arbitrary constructions ether models can be found that allow a certain region of phenomena to be represented. Such inventions acquire a value as contributions to our knowledge only when they lead to a fusion of two formerly unconnected physical regions. This is the great advance achieved by Maxwell when he fitted optics into the scheme of electromagnetic phenomena.

7. The Optics of Moving Bodies

Before we pursue this development further we wish to pause and to ask how the doctrine of the elastic ether behaves towards the space-time problem and relativity. Whereas in our optical investigations so far we have taken no account of the position or motion of the bodies that emit, receive, or allow the passage of light, we shall now concentrate our attention on just these conditions.

The space of mechanics is regarded as empty wherever there are no material bodies present. The space of optics is filled with ether. But here the ether is for us actually a kind of matter that has a certain mass, density, and elasticity. Accordingly we can immediately apply Newtonian mechanics with its doctrine of space and time to the universe full of ether. This universe then no longer consists of isolated masses that are separated by empty spaces but is completely filled with the thin mass of the ether, in which the coarse masses of matter are floating. The ether and matter act on each other with mechanical forces and move according to the Newtonian laws. Thus Newton's standpoint is logically applicable to optics. The question is only whether observation is in agreement with it.

But this question cannot be answered simply by unambiguous experiments. For the state of motion of the ether outside and inside matter is not known, and we are free to think out hypotheses about it. Thus we must put the question in the form : is it possible to make assumptions about the mutual actions of the motions of the ether and of matter such that all optical phenomena are thereby explained ?

We now call to mind the doctrine of the principle of relativity of classical mechanics. According to it absolute space exists only in a restricted sense ; for all inertial systems that move rectilinearly and uniformly with respect to each other may be regarded with equal right as being at rest in space. The first hypothesis that suggests itself to us concerning the luminiferous ether is the following

The ether in astronomic space far removed from material bodies is at rest in an inertial system.

For if this were not the case parts of the ether would be accelerated. Centrifugal forces would arise in it and would bring about changes of density and elasticity, and we should expect that the light from stars would have given us indications of this.

In form this hypothesis satisfies the classical principle of relativity. If the ether is counted among material bodies, then motions of translations of bodies with respect to the ether are just as much relative motions as those of two bodies with respect to each other, and a common motion of translation of the ether and all matter should be capable of detection either mechanically or optically.

But the physics of material bodies *alone*, without the ether, need no longer satisfy the principle of relativity. A common translation of all matter in which the ether does not participate, that is, a relative motion with respect to the latter, could very well be ascertained by optical experiments. Then the ether would practically define a system of reference that is absolutely at rest. The question which is important above all else for the sequel is whether the observable optical phenomena depend only on the relative motions of material bodies or whether the motion in the sea of ether makes itself remarked.

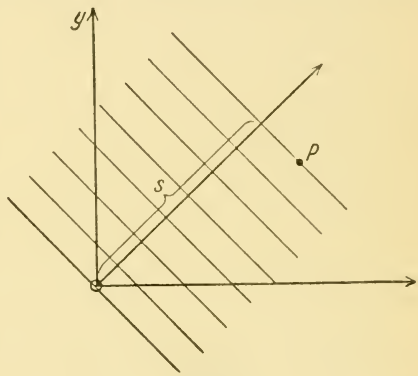


FIG. 69.

A light wave has three characteristics:

1. The vibration number or frequency.
2. The velocity.
3. The direction of propagation.

We shall now investigate systematically what influence relative motions of the bodies emitting and receiving light with respect to each other and to the transmitting medium, be it the ether in free astronomic space or be it a transparent substance, have on these three characteristics.

We shall apply the following method. We consider a train of waves, which leaves the zero-point o in any direction at the time $t = o$, and we count the individual waves that pass over any point P up to the time t. This number is evidently quite independent of the co-ordinate system in which the co-ordinates P are measured, whether this system be at rest or moving. We determine this number thus:

The first wave that leaves the zero-point $t = o$ has to advance along a certain distance s (Fig. 69) until it reaches the point P, and it takes the time $\frac{s}{c}$ to do this. From this moment onwards we count the waves that pass over P up to the moment t, that is, during the time $t - \frac{s}{c}$. Now since the light executes ν vibrations in one second, and every wave that passes by corresponds exactly to one vibration, ν waves pass by in one second, and hence $\nu\left(t - \frac{s}{c}\right)$ waves pass over the point P in the time $t - \frac{s}{c}$ secs.

Thus the wave number $\nu\left(t - \frac{s}{c}\right)$ is dependent only on how the two points O and P are situated with respect to each other and to the train of waves, and on how great the interval of time t is between the departure of the first wave at O and the arrival of the last at P. This number has nothing to do with the system of reference. Thus it is an *invariant* in the sense that we have attached to this word above.

This comes out most clearly if we use Minkowski's mode of expression. According to this the departure of the first wave from the zero point at the time $t = o$ is an event, a world-point; the arrival of the last wave at the time t at the point P is another event, a second world-point. But world-points exist without relation to definite co-ordinate systems. And since the wave-number $\nu\left(t - \frac{s}{c}\right)$ is determined by the two world-points, only it is independent of the system of reference, or is invariant.

From this there easily follow, either by intuition or by applying Galilei-transformations, all theorems about the behaviour of the three characteristics of the wave, the frequency, direction, and velocity, when the system of reference is changed. We shall deduce these theorems in order and shall compare them with experience.

8. The Doppler Effect

The fact that the observed *frequency* of a wave depends on the motion both of the source of light and of the observer, each with respect to the intervening medium, was discovered by Christian Doppler (1842). The phenomenon may easily be observed in the case of sound waves. The whistle of a locomotive seems higher when it is approaching the observer and becomes deeper at the moment of passing. The rapidly approaching source of sound carries the impulses forwards so that they succeed each other more rapidly. The motion of an observer moving towards the source has a similar effect; he then receives the waves in more rapid succession. So the same must hold in the case of light. Now the frequency of the light determines its colour ; the rapid vibrations correspond to the violet end of the spectrum, the slower vibrations to the red. Hence when a light source is approaching an observer or *vice versa* the colour of the light inclines a little towards violet ; in the case when either is receding, a little towards red.

This phenomenon has actually been observed.

Now the light which comes from luminescent gases does not consist of all possible vibrations but of a number of separate frequencies. The spectrum that a prism or a spectral apparatus depending on interference exhibits is no continuous band of colour like the rainbow but separate sharp-coloured lines. The frequency of these spectral lines is characteristic of the chemical elements that are emitting light in the flame (spectral analysis by Bunsen and Kirchhoff 1859). The stars, too, have such line spectra, whose lines for the most part coincide with those of the earth's elements. From this it is to be inferred that the matter in the furthermost depths of astronomic space is composed of the same primary constituents. The lines of the stars do not, however, exactly coincide with the corresponding lines on the earth but show small displacements towards the one side for one half of the year, and towards the other during the other half. These changes of frequency are the results of the Doppler effect of the earth's motion about the sun. During the one half of the year the earth moves towards a definite star, and hence the frequency of all the light waves coming from this star are magnified and the spectral lines of the star appear shifted towards the side of rapid frequencies (the violet end), whereas during the second half of the year the earth moves away from the star, and hence the spectral lines are then displaced towards the other side (the red end).

This wonderful picture of the earth's motion in the spectrum of the stars does not, indeed, present itself in an unadulterated form. For it is clear that there will be superposed on it the Doppler effect due to the emission of the light by a moving source. Now, if the stars are not all at rest in the ether, their motion must again manifest itself in a displacement of the spectral lines. This becomes added to that due to the earth's motion, but does not show the annual change, and hence may easily be distinguished and separated from the former. Astronomically this phenomenon is much more important still, for it gives us information about the velocities of even the most distant stars so far as the motion entails an approach towards or a recession from the earth. It is not our object, however, to enter more closely into these investigations.

We are interested above all in the question as to what happens when the observer and the source of light move in the same direction with the same velocity. Does the Doppler effect then vanish, does it depend only on the relative motion of the material bodies, or does it not vanish and thereby betray the motion of bodies through the ether? In the former case the principle of relativity would be fulfilled for the optical phenomena that occur between material bodies.

The ether theory gives the following answer to this question.

The Doppler effect does not only depend on the relative motion of the source of light and of the observer, but also to a slight extent on the motions of both with respect to the ether. But this influence is so small that it escapes observation; moreover, in the case of a common translation of the source of light and of the observer it is rigorously equal to zero.

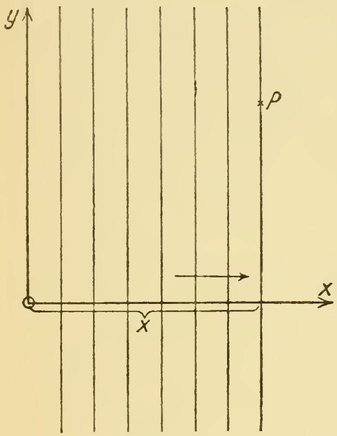


FIG. 70.

The latter point is so clear intuitively that it need hardly be emphasized. It is only necessary to reflect that the waves pass by any two points at rest relatively to each other in the same rhythm, irrespective of whether the two points are at rest in the ether or move with a common motion. Nevertheless the principle of relativity does *not* hold *rigorously*, but only approximately, for the bodies emitting and absorbing the light. We shall prove this.

For this purpose we make use of the theorem above derived concerning the invariance of the wave-number.

Let us allow a train of waves to start out in the x-direction from the zero-point of the system S which is at rest in the ether, and let us count the waves that pass over a point P until the time t has elapsed (Fig. 70). The path which the waves traverse in this time is equal to the x-co-ordinate of the point P. Thus we must set $s = x$, and the wave-number amounts to

$$\nu\left(t - \frac{x}{c}\right).$$

Let us next consider a system S′ moving in the x-direction with the velocity v, and let the observer be at rest in it at a point with the co-ordinate x'. At the time $t = o$ S and S′ are to coincide, and at the time t the observer is just to have reached the point P. Then the same wave-number in the system S′ is equal to

$$\nu'\left(t - \frac{x'}{c'}\right),$$

where ν' and c' denote the frequency and velocity as measured by the moving observer.

We thus have

$$\nu\left(t - \frac{x}{c}\right) = \nu'\left(t - \frac{x'}{c'}\right) \quad . \quad . \quad . \quad (38)$$

where the co-ordinates are connected by the Galilei transformation (29) on p. 65,

$$x' = x - vt \text{ or } x = x' + vt$$

If we insert this we get

$$\nu\left(t - \frac{x' + vt}{c}\right) = \nu'\left(t - \frac{x'}{c'}\right) \quad . \quad . \quad (39)$$

and this must of course hold for all values of x' and t. If we choose, in particular, $t = 1$, $x' = 0$, we get

$$\nu\left(1 - \frac{v}{c}\right) = \nu' \quad . \quad . \quad . \quad (40)$$

That is the desired law. It expresses that an observer moving in the same direction as the light waves measures a frequency ν' that is reduced in the ratio $\left(1 - \frac{v}{c}\right) : 1$.

Conversely we now consider a light wave source that vibrates with the frequency ν_0, and moves in the direction of the x-axis with the velocity v_0. Let an observer at rest in the ether measure the frequency ν. This case is immediately reducible to the preceding one. For it is quite immaterial for our argument whether it is the light source or the observer that is moving, it only depends on the rhythm with which the waves impinge on a moving point. The moving point is now the source of light. We thus get the formula for this case from the preceding case if we replace v in it by v_0 and ν' by ν_0 :

$$\nu\left(1 - \frac{v_0}{c}\right) = \nu_0.$$

But here ν_0 is given as the frequency of the source of light, and

ν, the observed frequency, is being sought. Thus we must solve for ν and we get

$$\nu = \frac{\nu_0}{1 - \frac{v_0}{c}} \quad . \quad . \quad . \quad . \quad (41)$$

The observed frequency, therefore, appears magnified, since the denominator is less than 1, in the ratio $1 : \left(1 - \frac{v_0}{c}\right)$.

We see at once that it is not immaterial whether the observer moves in the one direction or the source in the opposite direction with the same velocity.

For if, in formula (41), we set $v_0 = - v$, it becomes

$$\nu = \frac{\nu_0}{1 + \frac{v}{c}},$$

and this is different from (40). In all practical cases the difference is certainly very small. We saw earlier (IV, 3, p. 82) that the ratio of the velocity of the earth in its orbit around the sun compared with that of light is $\beta = \frac{v}{c} = 1 : 10{,}000$, and similar small values of β hold for all cosmic motions. But we may then write as a very close approximation

$$\frac{1}{1 + \beta} = 1 - \beta;$$

for if we neglect $\beta^2 = \frac{1}{100{,}000{,}000} = 10^{-8}$ compared with 1, we have $(1 + \beta)(1 - \beta) = 1 - \beta^2 = 1$.

This rejection of the square of $\beta = \frac{v}{c}$ will play an important part in the sequel. It is almost always permissible because such exceedingly small quantities as $\beta^2 = 10^{-8}$ are accessible to observation in only a few cases. The phenomena of the optics (and electrodynamics) of moving bodies are nowadays, indeed, classified according to whether they are of the order β or β^2. The former quantities are said to be of the *first order*, and the latter of the *second order* in β. In this sense we may assert the following:

The Doppler effect depends only on the relative motion of the source of light and of the observer if the quantities of the second order are neglected.

We see this, too, if we assume a simultaneous motion of the source of light (velocity v_0) and the observer (velocity v). We then clearly obtain the observed frequency ν' if we insert ν from (41) in (40):

$$\nu' = \nu\left(1 - \frac{v}{c}\right) = \nu_0 \frac{1 - \frac{v}{c}}{1 - \frac{v_0}{c}}.$$

If the source of light and the observer have the same velocity $v_0 = v$, the fraction becomes equal to 1 and we get $\nu' = \nu_0$. Thus the observer notices nothing of a common motion with the source relative to the ether. But as soon as v differs from v_0, a Doppler effect comes about, the amount of which depends not only on the difference of the velocities $v - v_0$. This would allow the motion relative to the ether to be ascertained if the difference were not of the second order and hence much too small to be observed.

We see that the Doppler effect gives no useful practical method of establishing motions with respect to the ether in astronomic space.

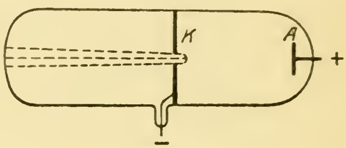


FIG. 71.

We must further add that the Doppler effect has been detected with sources of light on the earth. This required sources of light moving with extremely great speed in order that the ratio $\beta = \frac{v}{c}$ might attain a perceptible value. For this purpose J. Stark (1906) used the so-called *canal rays*. If two electrodes are fixed in an evacuated tube containing hydrogen of very small density, and if one of the electrodes is perforated and made the negative terminal (cathode) of an electric discharge (Fig. 71), we get in the first place the so-called cathode rays, and secondly, as Goldstein discovered in 1886, a reddish luminescence penetrates through the hole or holes of the cathode, due to positively charged hydrogen atoms or molecules moving at a great speed. The velocity of these canal rays is of the order $v = 10^8$ cms. per sec., thus β has the value

$$\beta = \frac{10^8}{3 \cdot 10^{10}} = \frac{1}{300},$$

which is fairly high compared with the astronomic values.

Stark investigated the spectrum of canal rays and found that the bright lines of hydrogen exhibited the displacement that was to be expected on the grounds of the Doppler effect. This discovery became of great importance for atomic physics. But it does not belong to our proper theme.

Finally we have to mention that Beloposki (1895) and Galitzin (1907) proved the existence of a sort of Doppler effect with the help of sources of light on the earth and moving mirrors.

9. The Convection of Light by Matter

We have next to investigate the second characteristic of a source of light, namely, its *velocity*. According to the ether theory the velocity of light is a quantity that is determined by the density of mass and the elasticity of the ether. Thus it has a fixed value in the ether of astronomic space, but a different value in every material body, which will depend on how the matter influences the ether in its interior and carries it along with itself.

If we first treat the velocity of light in astronomic space we must conclude that an observer moving relatively to the ether will measure a velocity different from that measured by our observer at rest. For here the elementary laws of relative motion clearly hold. If the observer moves in the same direction as the light, its velocity relative to the ether will seem diminished by the amount of the velocity v of the observer. Indeed, beings can be imagined that could overtake light. The same result arises from the formulæ above derived that express the general relations between the properties of light as established by two observers moving with translation relatively to each other. If we set $t = 0$, $x' = 1$ in formula (39) we get

$$\frac{\nu}{c} = \frac{\nu'}{c'},$$

and if we insert the value for ν' from (40) in this we get

$$\frac{\nu}{c} = \frac{\nu}{c'}\left(1 - \frac{v}{c}\right),$$

or, since ν cancels out,

$$c' = c\left(1 - \frac{v}{c}\right) = c - v \quad . \quad . \quad . \quad (42)$$

This signifies that the velocity of light in the moving system is determined according to the rules of relative motion.

This may also be interpreted by regarding an observer who is moving through the ether as being in an *ether wind* that blows away from or against the light waves just like the air brushes past a quickly moving motor car and carries the sound with it.

Now, this furnishes us with a means of establishing the motion of, say, the earth or the solar system relative to the ether. We have two essentially different methods of measuring the velocity of light, an astronomical and a terrestrial method. The former, the old process of Rõmer, makes use of the eclipses of Jupiter's satellites ; it measures the velocity of the light that traverses the space between Jupiter and the earth. In the latter method the source of light and the observer participate in the motion of the earth. Do these two methods give exactly the same result or are there deviations that betray motion relative to the ether ?

Maxwell (1879) called attention to the fact that by observing the eclipses of Jupiter's moons it should be possible to ascertain

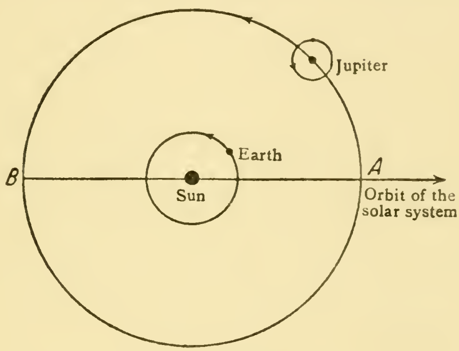


FIG. 72.

a motion of the whole solar system with respect to the ether. Let us suppose the planet Jupiter at the point A of its orbit (Fig. 72), which is the point nearest to the orbit of the sun in the motion of the solar system in the direction shown. (It has been assumed in the diagram that the orbit of Jupiter intersects the orbit of the solar system at A.) In the course of a year Jupiter moves only a short distance away from A, since its time of revolution in its own orbit is about twelve years. In one year the earth traverses its orbit once, and by observing eclipses it is possible to find the time required by the light to travel across the diameter of the earth's orbit. Now, since the whole solar system moves in the direction of the sun towards A the light from Jupiter to the earth runs contrary to this motion, and its velocity appears increased. Let us now wait for six years until Jupiter is situated at the opposite point B of its orbit. The light now runs in the same direction as the solar system, and thus requires a longer time to cross the earth's orbit ; so its velocity appears smaller.

When Jupiter is at A the eclipses of one of his satellites during half a year (of the earth) must be delayed by the amount of time $t_1 = \frac{l}{c + v}$, where l denotes the diameter of the earth's orbit. When Jupiter is at B the delay amounts to $t_2 = \frac{l}{c - v}$. If the solar system were at rest in the ether both delays would be equal to $t_0 = \frac{l}{c}$. Their actual difference, namely,

$$t_2 - t_1 = l\left(\frac{1}{c - v} - \frac{1}{c + v}\right) = \frac{2lv}{c^2 - v^2} = \frac{2lv}{c^2(1 - \beta^2)}$$

for which, by neglecting β^2 in comparison with 1, we may write

$$t_2 - t_1 = \frac{2lv}{c^2} = 2t_0\beta,$$

allows us to determine β and hence also the velocity $v = \beta c$ of the solar system relative to the ether. Now, light takes about eight minutes to travel from the sun to the earth, thus $t_0 = 16$ mins. or $= 1000$ secs. (in round numbers). Thus from a time-difference $t_2 - t_1 = 1$ sec. we should get $\beta = \frac{1}{2000}$ or $v = \beta c = \frac{300000}{2000} = 150$ kms. per sec.

The velocities of the stars relative to the solar system, which may be deduced from the Doppler effect, are mostly of the order 20 kms. per sec., but velocities up to 300 kms. per sec. occur in certain clusters of stars and spiral nebulæ. The accuracy of the astronomic determinations of time has thus far not sufficed to establish a delay in the eclipses of a satellite of Jupiter to the extent of one sec. or less in the course of half a year. Yet it is not out of the question that refinement of the methods of observation will yet disclose such a delay.

An observer situated on the sun, who happened to know the value of the velocity of light in the ether at rest, would also be able to ascertain the motion of the solar system through the ether by means of the eclipses of Jupiter's satellites. To do this he would have to measure the delay in the eclipses during half a revolution of Jupiter in his orbit. The same formula $t_2 - t_1 = 2t_0\beta$ is valid for this, but now t_0 denotes the time that the light requires to traverse the diameter of *Jupiter's* orbit. This value of t_0 is (about $2\frac{1}{2}$ times) greater than the value used above for the earth's orbit, 16 mins., and the delay $t_2 - t_1$ becomes greater in the same proportion. But for the same reason the time of revolution of Jupiter, during which the eclipses must be observed consecutively, is much greater than (about 12 times as great as) an earth year, so that this method, which could also be applied by an observer on the earth, seems to promise no advantage.

At any rate the fact that the accuracy that is nowadays attainable has brought to light not even a delay of several seconds proves that the velocity of the solar system with respect to the ether is not much greater than the greatest known velocities of the stars relative to each other.

We next turn our attention to the terrestrial modes of measuring the velocity of light. Here it is easy to see why they do not allow us to draw conclusions about the motion of the earth through the ether. We have already indicated the ground for this above when mentioning these methods for the first time (IV, 3, p. 82), for the light traverses one and the same path in its journey there and back. It is only a mean velocity during the path to and fro that is actually measured. The deviation of this from the velocity of light c in the ether is, however, a quantity of the second order with respect to β and is not accessible to observation. For if l is the length of path then the time that the light requires for the first journey, in the direction of the earth's motion, is equal to $\frac{l}{c-v}$, and the time for the return journey is $\frac{l}{c+v}$, thus the whole time is

$$l\left(\frac{1}{c+v}+\frac{1}{c-v}\right)=\frac{2lc}{(c+v)(c-v)}=\frac{2lc}{c^2-v^2}.$$

The mean velocity is $2l$ divided by this time, thus it is

$$\frac{c^2-v^2}{c}=c\left(1-\frac{v^2}{c^2}\right)$$

and hence it differs from c by a quantity of the second order.

Besides the direct measurement of the velocity of light there are numberless other experiments in which the velocity of light comes into play. All interference and diffraction phenomena are brought about by making light-waves that travel along different paths meet at the same place and causing them to be superposed on each other. Refraction at the boundary of two bodies arises through light having different velocities in them; thus this velocity enters into the action of all optical apparatus that contains lenses, prisms, and similar things. Is it not possible to think out arrangements in which the motion of the earth and the "ether wind" produced by it make themselves remarked?

Very many experiments have been designed and carried out to discover this motion. The general result of experiments with sources of light on the earth teaches us that not the slightest influence of the ether-wind is ever observable. It is true that up to recent times we have been dealing with experimental arrangements that allow only quantities of the first order in β to be measured. The fact that this must always lead to a negative result easily follows from the circumstance that the true duration of the motion of the light from one place to another is never measured, but only differences of such times for the same light-path or their sum for the motion there and back. For the reason given above we thus see that the quantities of the first order always cancel out.

But we might expect a positive result if we took a source not on the earth but in the heavens. If we direct a telescope at a star, to which the momentary velocity v of the earth is just directed (Fig. 73), the velocity of light in the lenses of the telescope relative to the substance of the glass will be greater by the amount v than if the earth were at rest, and if we look at the same star six months later through the telescope, the velocity of light in the lenses will be smaller by the amount v. Now, since the amount of the refraction in a lens is determined by the velocity of light, we might expect the focus of the lens to have a different position in these two cases. This would be an effect of the first order. For the difference of the velocity of light in the two cases would be $2v$, and its ratio to the velocity in the ether at rest would be $\frac{2v}{c} = 2\beta$.

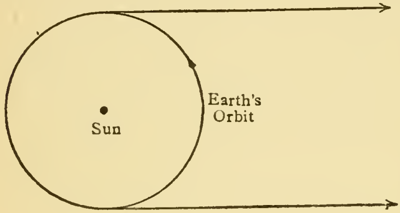


FIG. 73.

Arago actually carried out this experiment, but found no change in the position of the focus. How is this to be explained?

We clearly made the assumption above that the velocity of light in a body that moves in the ether against the ray with the velocity v is greater by just this amount than if the body were at rest in the ether. In other words, we have assumed that material bodies pierce through the ether without carrying it along in the slightest, just like a net that is carried through water by a boat.

The results of experiment teach us that this is manifestly not the case. Rather, the ether must participate in the motion of matter. It is only a question of how much.

Fresnel established that to explain Arago's observation and all other effects of the first order it was sufficient to assume that the ether is only *partly* carried along by matter. We shall forthwith discuss in detail this theory, which has been brilliantly confirmed by experiment.

It was Stokes (1845) above all others who later adopted the more radical standpoint that the ether in the interior of matter shares completely in its motion. He assumed that the earth carries along with itself the ether which is in its interior, and that this ether motion gradually decreases outwards until the state of rest of the ether in the universe is reached. It is clear that then all optical phenomena on the earth occur exactly as if the earth were at rest. But in order that the light that comes from the stars may not experience deflections and changes of velocity in the transitional stratum between the ether of space and the ether convected by the earth, special hypotheses concerning the motions of the ether must be made. Stokes found a hypothesis such as satisfied all optical conditions. But it was shown later not to be in agreement with the laws of mechanics. Numerous attempts at rescuing Stokes's theory have led to no result, and it would have succumbed to internal difficulties even if Fresnel's theory had not been confirmed by Fizeau's experiment (see p. 119 below).

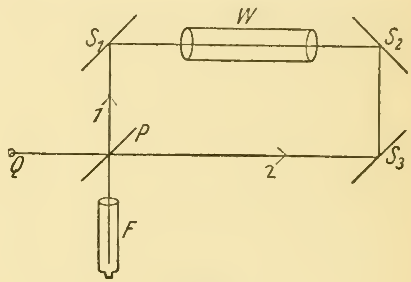


Fig. 74.

Fresnel's idea of partial convection cannot easily be deduced from Arago's experiment because refraction in lenses is a complicated process which concerns not only the velocity but also the direction of the waves. But there is a fully equivalent experiment that was carried out by Hoek (1868) later and which is much easier to follow.

The principle underlying the arrangement of the apparatus is that of the interferometer (Fig. 74). The light falls from the source Q on to a half-silvered glass plate P inclined at 45° to the direction of the ray. This glass plate divides the ray into two parts. The reflected ray (ray 1) strikes consecutively the mirrors S_1, S_2, S_3, that form the corners of a square with P, and on its return to P is partly reflected into the telescope F.

The transmitted ray (ray 2) traverses the same path in the reverse sense and interferes with ray 1 in the field of vision. A transparent body, say a tube W filled with water, is next interposed between S_1 and S_2, and the whole apparatus is mounted so that the straight line connecting S_1 with S_2 can be placed alternately in the same direction as, and opposite to, the earth's motion about the sun. Let the velocity of light in water that is at rest be c_1. This value is a little smaller than the velocity *in vacuo* and the ratio of the one to the other $\frac{c}{c_1} = n$ is called the *refractive index* of water. The velocity of light in air differs only inappreciably from c, and thus the refractive index of air is almost exactly equal to 1. Now the water is carried along by the earth in its orbit. If the ether in the water were not to participate in this motion at all then the velocity of light in the water relative to the absolute ether (in outside space) would be unaltered; that is, it would be equal to c_1, and, for a ray travelling in the direction of the earth's motion, it would be $c_1 + v$, and c_1 relative to the earth. We shall assume neither of these cases to begin with, but shall leave the amount of convection undetermined. Let the velocity of light in the moving water relative to the absolute ether be a little greater than c_1, say $c_1 + \phi$, and hence $c_1 + \phi - v$ relative to the earth. We wish to determine the unknown *convection coefficient* ϕ from experiment. If it is zero, no convection occurs; if it is v, complete convection occurs. Its true value must lie between these limits. We shall, however, make one assumption, namely, that the convection in air may be neglected in comparison with that in water.

Now, let l be the length of the tube of water. Then the ray 1 requires the time $\frac{l}{c_1 + \phi - v}$ to traverse the tube, if the earth is moving in the direction from S_1 to S_2. To traverse the corresponding air distance between S_3 and P the same ray requires the time $\frac{l}{c + v}$. Thus, on the whole, the time that the ray 1 requires to traverse the two equal paths in water and in air is

$$\frac{l}{c_1 + \phi - v} + \frac{l}{c + v}.$$

The ray 2 travels in the reverse direction. It first traverses the air-distance in the time $\frac{l}{c - v}$, then the water-distance in the time $\frac{l}{c_1 - \phi + v}$, and hence altogether it requires for the same distances in the air and water the time

$$\frac{l}{c - v} + \frac{l}{c_1 - \phi + v}$$

Now, experiment shows that the interferences do not shift in the slightest when the apparatus is turned into the direction opposite to that of the earth's velocity or, indeed, into any other position whatsoever. From this it follows that the rays 1 and 2 take equal times, independent of the position of the apparatus with respect to the earth's orbit, that is,

$$\frac{l}{c_1 + \phi - v} + \frac{l}{c + v} = \frac{l}{c - v} + \frac{l}{c_1 - \phi + v}.$$

We can calculate ϕ from this equation. We shall pass over the somewhat circuitous calculation * and shall give only the result which, if we neglect quantities of the second and higher orders, is :

$$\phi = \left(1 - \frac{1}{n^2}\right)v \quad . \quad . \quad . \quad (43)$$

This is the famous *convection formula* of Fresnel, who, indeed, found it by a different, more speculative, process. Before we mention his assumption let us see what the formula actually asserts. According to it the convection is the greater the more the refractive index exceeds the value 1 which it has *in vacuo*. For air c_1 is almost equal to c, and n almost equal to 1, thus ϕ is almost zero, as we predicted above. The greater the refractive power, the more complete is the convection of the light. Now, the velocity of light in a moving body, measured

* The steps of the argument are :

$$\frac{(c + v) + (c_1 + \phi - v)}{(c_1 + \phi - v)(c + v)} = \frac{(c_1 - \phi + v) + (c - v)}{(c - v)(c_1 - \phi + v)},$$

$$(c + c_1 + \phi)(c - v)(c_1 - \phi + v) = (c + c_1 - \phi)(c + v)(c_1 + \phi - v),$$

$$\phi^2 - \phi\frac{v^2 + c^2}{v} = c_1^2 - c^2,$$

$$\phi = \frac{v}{2\beta^2}\left(1 + \beta^2 - \sqrt{(1 - \beta^2)^2 + \frac{4\beta^2}{n^2}}\right),$$

and, approximately, we have

$$\phi = \frac{v}{2\beta^2}\left[1 + \beta^2 - \left(1 - \beta^2 + \frac{2\beta^2}{n^2}\right)\right],$$

$$\phi = \left(1 - \frac{1}{n^2}\right)v.$$

relative to the absolute ether is

$$c_1 + \phi = c_1 + \left(1 - \frac{1}{n^2}\right)v,$$

and relative to the moving body it is

$$c_1 + \phi - v = c_1 + \left(1 - \frac{1}{n^2}\right)v - v = c_1 - \frac{v}{n^2}.$$

This last formula will serve us as a link to Fresnel's interpretation. He assumed that the density of the ether in a material body is different from the density in free ether; let the former be ρ_1 and the latter ρ.

We next imagine the moving body, say, in the form of a beam, whose length is parallel to the direction of motion; let its basic face be of unit area. In the motion of the beam through the ether the front face advances by the distance v in a unit of time (Fig. 75), and this sweeps out a volume v (area of the face multiplied by the height). This volume contains an amount of ether ρv. Thus this enters into the beam through the front face. Here it assumes a new density and will thus move on with a different velocity v with respect to the body, since, for the same reasons as above, its mass must also equal $\rho_1 v_1$, and we get

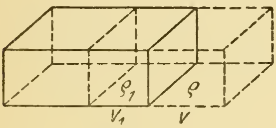


FIG. 75.

$$\rho_1 v_1 = \rho v$$

or

$$v_1 = \frac{\rho}{\rho_1} v.$$

This is in a certain sense the strength of the ether wind in the beam moving with the velocity v. Light which moves with the velocity c_1 relatively to the condensed ether moves relatively to the body with the velocity

$$c_1 - v_1 = c_1 - \frac{\rho}{\rho_1} v.$$

Now we have seen that according to the result of Hoek's experiment the velocity of light relative to the moving body is

$$c_1 - \frac{1}{n^2} v.$$

Consequently we must have

$$\frac{\rho}{\rho_1} = \frac{1}{n^2} = \frac{c_1^2}{c^2}.$$

Thus the condensation $\frac{\rho_1}{\rho}$ is equal to the square of the coefficient of refraction.

Furthermore, we can conclude from this that the elasticity of the ether must be the same in all bodies. For formula (37) on p. 99 tells us that in every elastic medium $c^2 = \frac{p}{\rho}$. Thus in ether $p = c^2\rho$, in matter $p_1 = c_1^2\rho_1$. But according to the above result concerning the condensation of ether in matter these two expressions are the same.

This mechanical interpretation of the convection coefficient by Fresnel has exerted a great influence on the elaboration of the elastic theory of light. But we must not disguise from ourselves that it is open to strong objections. As is well known, rays of light of different colour (frequency) have different refractive indices n, that is, different velocities. Hence it follows that the convection coefficient has a different value for each colour. But this is incompatible with Fresnel's interpretation, for then the ether would have to flow with a different velocity in the body according to the colour. Thus there would be just as many ethers as there are colours, and that is surely impossible.

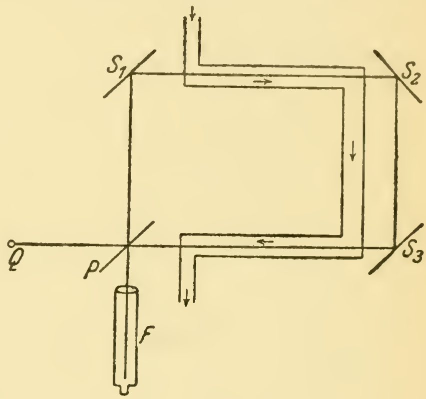


FIG. 76.

The convection formula (43), however, is founded on the results of experiment without regard to the mechanical interpretations. We shall see that it is derived in the electro-magnetic theory of light from ideas concerning the atomic structure of matter and electricity.

It is very difficult to test Fresnel's formula by means of experiments on the earth because it requires that transparent substances be moved with extreme rapidity. Fizeau succeeded in carrying out the experiment (1851) by means of a sensitive interferometer arrangement.

The apparatus used by him is quite similar to that of Hoek, except that *both* light-paths S_1S_2 and S_3 are furnished with tubes in which the water can circulate; they are arranged so that the ray 1 flows directly parallel to the water, and the ray 2 directly against it. Fizeau tested whether the water carries the light along with it by observing whether the interference fringes were displaced when the water was set into rapid motion. This displacement actually occurred, but very much less than to the extent that would correspond to complete convection. Exact measurement disclosed perfect agreement with Fresnel's convection formula (43).

10. Aberration

We shall now discuss the influence of the motion of bodies on the *direction* of light-rays, in particular the question whether the motion of the earth through the ether can be ascertained by observing any phenomena accompanying changes of direction. Here again we have to distinguish whether we are dealing with an astronomic or an earth source of light.

The apparent deflection of the light that reaches the earth from the stars is the *aberration*, which we have already discussed from the point of view of the corpuscular theory (IV, 3, p. 81). Although the explanation there given is very simple, it is correspondingly complicated from the point of view of the wave-theory, for it is easy to see that a deflection of the wave-planes does not occur at all. This is seen most readily in the case where the rays fall perpendicularly to the motion of the observer. For then the wave-planes of this motion are parallel and are so perceived by the moving observer (Fig. 77). But calculation tells us the same. Let us place a stationary co-ordinate system S and a moving system S′ so that the x- and the x'-axis each fall in the direction of motion, and let us count the waves that have passed over any point P from the moment $t = o$ to the moment t. This number is, as we know, $\nu\left(t - \frac{s}{c}\right)$, where s is the path traversed by the waves. In the case of perpendicularly incident waves we clearly have $s = y$.

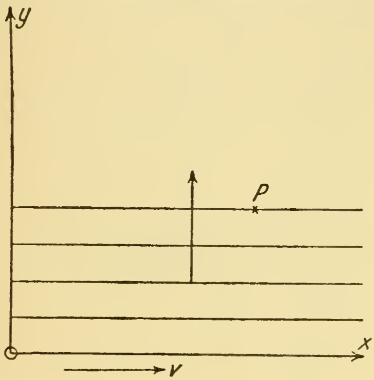


Fig. 77.

The invariance of the wave-number requires that

$$\nu\left(t - \frac{y}{c}\right) = \nu'\left(t - \frac{y'}{c'}\right)$$

if the co-ordinates are transformed into each other by the

Galilei transformation. In this case the y-co-ordinate remains unaltered, and hence we must have

$$\nu = \nu' \text{ and } \frac{\nu}{c} = \frac{\nu'}{c'}, \qquad \text{thus } c = c'.$$

Hence the moving observer sees a wave of exactly the same frequency, velocity, and direction. For if this were altered then the wave-number in S′, besides depending on y', would also have to depend on x'.

Thus it seems as if the wave theory is unable to account for the simple phenomenon of aberration, which has been known for almost 200 years.

But the position is not quite so bad as this would indicate. The reason for the failure of the argument given just above is that the optical instruments with which the observations are made, and which include the naked eye, do not establish the position of the wave that arrives, but accomplish something totally different.

The function of the eye or of the telescope is called *optical representation,* and it consists in combining the rays emitted by a luminescent object into one picture. In this process the vibrational energy of the particles of the object are transported by the light-waves to the corresponding particles of the picture. The paths along which this transference of energy takes place are actually the physical rays. But energy is a quantity which, according to the law of conservation, can move about and be transformed just like a substance, but cannot be created or destroyed. Hence it seems reasonable to apply the laws of the corpuscular theory to the motion of energy. As a matter of fact the simple derivation of the aberration formula given earlier is quite correct if we define the light-rays as the energy-paths of the light-waves and apply the laws of relative motion to them, as if they were streams of projected particles.

But we may also obtain this aberration formula, without applying this conception of rays as energy paths, by following the refraction of waves individually in the lenses or prisms of the optical instrument. For this we require a definite convection theory. Stokes's theory of complete convection can account for aberration only by making assumptions about the motion of the ether which are not admissible. We have already called attention to these difficulties above. Fresnel's theory gives a law of refraction of waves of light at the surface of moving bodies from which the aberration formula follows exactly. The substance of the body through which the light passes does not affect the result, although the value of the convection coefficient is different in every substance. To test this directly Airy (1871) filled a telescope with water and ascertained that the aberration retained its normal value. The aberration is, of course, no longer an effect of the first order if the light-wave and the observer have no motion relative to each other. From this it also follows that in all optical experiments with sources of light on the earth no deflection of the rays through the ether wind occurs. Fresnel's theory succeeds in representing these facts so that they are in agreement with experiment. It is unnecessary to enter into the details.

11. Retrospect and Further Development

We have treated the luminiferous ether as a substance that obeys the laws of mechanics. Thus it satisfies the law of inertia, and hence where there is no matter, in astronomic space, it will be at rest in an appropriate inertial system. Now if we refer all phenomena to a different inertial system, exactly the same laws hold for the motions of bodies and of the ether, hence also for the propagation of light, but, of course, only in as far as they concern accelerations and mutual force effects. We know that the velocity and the direction of a motion are quite different with respect to different inertial systems; for we may regard every body moving in a straight line as at rest merely by choosing a suitable system of reference, namely, one that moves with it. Thus in this almost trivial sense the classical principle of relativity must hold for the ether regarded as a mechanical substance.

From this it follows, however, that the velocity and direction of light-rays must appear different in every inertial system. Thus it was to be expected that it would be possible to ascertain the velocity of the earth or of the solar system by observing optical phenomena at the surface of the earth, which are in the main conditioned by the velocity and direction of the light. But all experiments performed with this end in view led to a negative result. Hence it appears that the velocity and direction of the light-rays are quite independent of the motion of the astronomic body on which the observations are carried out. Or, expressed in other words, optical phenomena depend only on the relative motions of *material* bodies.

This is a principle of relativity which seems quite similar to the classical principle of mechanics, and yet it has a different meaning. For it refers to *velocities* and *directions* of motional events, and in mechanics these are *not* independent of the motion of the system of reference.

Now there are two possible points of view. One of these starts from the assumption that optical observations actually introduce something that is fundamentally new, namely, that light behaves differently from material bodies as regards direction and velocity. So soon as optical observations are taken as convincing evidence this point of view will be adopted if all speculations about the *nature* of light are left out of consideration. We shall see that Einstein finally pursued this path. It, however, requires utter freedom from the conventions of the traditional theory, which is attained only when the Gordian knot of constructions and hypotheses has become so intricate that the only solution left is to cut it.

But in our above discussion we were still living in the most flourishing period of the theory of the mechanical ether. This theory was compelled to regard the optical principle of relativity as a secondary, in a certain sense half accidental phenomenon, brought about by the compensating effect of causes that were acting in opposition to each other. The fact that such is possible to a certain extent is due to the circumstance that it is still open to make hypotheses about how the ether moves and how it is influenced in its motion by moving bodies. Now it is a great achievement of Fresnel's convection hypothesis that it actually accounts for the optical principle of relativity, so far as quantities of the first order are concerned. So long as the accuracy of optical measurements did not attain the great improvement necessary to measure quantities of the second order, this theory sufficed all demands of experiment with one possible exception, to which, curiously enough, very little attention was paid. For if improved accuracy in astronomical measurement should arrive at the result that by observing the eclipses of Jupiter's satellites according to the old method of Römer (see p. 80) an influence of the motion of the solar system on the velocity of light were to be revealed, then certainly the ether theory would be confronted with a problem that would appear insoluble. For it is clear that this effect of the first order could be argued away by no hypothesis about the convection of the ether.

So we recognize the importance of the experimental task of measuring the dependence of optical events on the earth's motion as far as quantities of the second order. Only the solution of this problem can give us a decision as to whether the optical principle of relativity holds rigorously or only approximately. In the former case Fresnel's ether theory would fail; we should then be confronted with a new state of affairs.

Historically, this occurred only about 100 years after Fresnel's time. In the meanwhile the ether theory was developed in other directions. For at the outset there was not *one* ether but a whole series, an optical, a thermal, an electrical, a magnetic ether, and perhaps a few more. A special ether was invented, as a carrier, for every phenomenon that occurs in space. At first all these ethers had nothing to do with each other, but existed in the same space independently of each other, side by side or, rather, interwoven. This state could not, of course, last in physics. Relationships were soon found between the phenomena of different branches that were at first separate, and so there emerged finally *one* ether as the carrier of *all* physical phenomena that bridge over space free of matter. In particular, light showed itself to be an electromagnetic process of vibration, of which the carrier is identical with the medium that transmits electric and magnetic forces. These discoveries first gave the ether theory strong support. At last, indeed, the ether came to be identified with Newtonian space. It was to persist in absolute rest and was to transmit not only electromagnetic effects but also indirectly to generate the Newtonian inertial and centrifugal forces.

We shall next describe the development of the theory. The process has features resembling the trial of a case in court. The ether is alleged to be the universal culprit, the pieces of evidence accumulate overwhelmingly, until at the end the undeniable proof of an alibi, namely, Michelson and Morley's experiment about the quantities of the second order, and its interpretation by Einstein puts an end to the whole business.

CHAPTER V

THE FUNDAMENTAL LAWS OF ELECTRODYNAMICS

1. Electro- and Magneto-statics

THE fact that a certain kind of ore, magnetite, attracts iron, and that rubbed amber (*elektron* in Greek) attracts and holds light bodies was known even to the ancients. But the sciences of magnetism and electricity are products of more recent times, which had been trained by Galilei and Newton to ply Nature with rational questions and to read the answer out of experiment.

The fundamental facts of electric phenomena were established from the year 1600 onwards. We shall recapitulate them briefly. At that time friction was the exclusive means of producing electrical effects. Gray discovered (1729) that metals, when brought into contact with bodies that have been electrified by friction, themselves acquire similar properties. He showed that the electrical effects can be passed along in the metals. This led to the classification of substances as conductors and non-conductors (insulators). It was discovered by du Fay (1730) that electrical action is not always *attraction* but may also be *repulsion*. He interpreted this fact by assuming two fluids (nowadays we call them positive and negative electricity), and he established that similarly charged bodies repel each other, oppositely charged bodies attract each other.

We shall here define the conception of the *electric charge* quantitatively at once. In doing so we shall not follow rigorously the very often circuitous steps of argument that led historically to the enunciation of the conceptions and laws, but we shall rather select a series of definitions and experiments in which the logical sequence comes out most clearly.

Let us imagine a body M that has somehow been electrified by friction. This now acts attractively or repulsively on other electrified bodies. To study this action we shall take small test bodies, say spheres, whose diameters are very small compared with their closest approach to the body M, at which we still wish to investigate the force. If we bring such a test body P near the body M, whose action we wish to study, P experiences a statical force of definite magnitude and direction, which may be measured by the methods of mechanics, say, by balancing it against a weight with the help of levers or threads.

We next take two such test bodies P_1 and P_2, bring them in turn to the same point in the vicinity of M, and measure in each case the forces K_1 and K_2 as regards size and direction. We shall henceforth adopt the convention that opposite forces are to be regarded as being in the same direction, but their values are to have opposite signs attached in calculations.

Experiment shows that the two forces have the same direction, but their values may be different and they may have different signs.

Now let us bring the two test bodies to a different point near M and let us again measure the forces K_1' and K_2' as regards value and direction. They have again the same direction, but in general they have different values and a different sign.

If we next form the ratio $K_1 : K_2$ of the forces at the first point, and then the ratio $K_1' : K_2'$ at the second, it is found that both have the same value, which may be positive or negative :

$$\frac{K_1}{K_2} = \frac{K_1'}{K_2'}.$$

From this result we may conclude :

1. The direction of the force exerted by an electrified body M on a small test-body P does not depend at all on the nature and the electrification of the test-body, but only on the properties of the body M.

2. The ratio of the forces exerted on two test-bodies brought to the same point in turn is quite independent of the choice of the point, that is from the position, nature, and electrification of the body M. It depends only on the properties of the test-bodies.

We now choose a definite test-body, electrified in a definite way, as a unit body, and we ascribe to it the charge or amount of electrification $+1$. With the aid of this we everywhere measure the force that the body M exerts. Let it be denoted by E. Then this also determines the direction of the force K exerted on any other test-body P. The ratio K : E, however, depends only on the test-body P and is called its *electric charge e*. This may be positive or negative according as K and E are in the same or in opposite directions, in the narrower sense. Thus we have

$$\frac{K}{E} = e \text{ or } K = eE \quad . \quad . \quad . \quad (44)$$

The force E on the charge 1 is also called the *electric intensity of field* of the body M. When once the unit charge has been fixed it depends only on the electrical nature of the body M and determines its electrical action in the surrounding space, or as we usually say, its " electric field."

As for the choice of the unit charge, it would be almost impossible to fix this practically by a decree concerning the electrification of a definite test-body ; rather, one will seek to find a mechanical definition for it. This is successfully done as follows :

We can first charge two test-bodies equally strongly. The criterion of equal charges is that they experience the same force from the third body M when placed successively at the same point near it. The two bodies will then repel each other with the same force. We now say that their charge is 1 if this repulsion is equal to the unit of force when the distance between the two test-bodies is equal to unit length. Nothing at all is herein assumed about the dependence of the force on the distance.

Through these definitions the amount of electricity or the electric charge becomes just as much a measurable quantity as length, mass, or force.

The most important law about amounts of electricity, which was enunciated independently in 1747 by Watson and Franklin, is the law that in every electrical process *equal amounts* of positive and negative electricity are always formed. For example, if we rub a glass rod with a piece of silk, the glass rod becomes charged with positive electrification ; an exactly equal negative charge is then found on the silk.

This empirical fact may be interpreted by saying that the two kinds of electrification are *not generated* by friction but are *only separated.* They are represented as two *fluids* that are present in all bodies in equal quantities. In non-electrified " neutral " bodies they are everywhere present to the same amount so that their effects outwards are counterbalanced. In electrified bodies they are separated. One part of the positive electrification, say, has flowed from one body to another, and just as much negative has flowed in the reverse direction.

But it is clearly sufficient to assume *one* fluid that can flow independently of matter. Then we must ascribe to the matter that is free of this fluid a definite charge, say positive, and to the fluid the opposite charge, that is, negative. The electrification consists in the negative fluid flowing from one body to the other. The first will then become positive because the positive charge of the matter is no longer wholly compensated; the other becomes negative because it has an excess of negative fluid.

The struggle between the supporters of these two hypotheses, the *one-fluid theory* and the *two-fluid* theory, lasted a long time, and of course remained fruitless and purposeless until it was decided by the discovery of new facts. We shall not enter further into these discussions, but shall only state briefly that characteristic differences were finally found in the behaviour of the two electrifications; these differences indicated that the positive electrification is actually firmly attached to matter, but the negative can move freely. This doctrine still holds to-day. We shall revert to this point later in dealing with the theory of electrons.

Another controversy gathered round the question as to how the electrical forces of attraction and repulsion are transmitted through space. The first decades of electrical research were not yet carried out under the influence of the Newtonian theory of attraction. Action at a distance seemed unthinkable. Metaphysical theorems were held to be valid, such as that matter can act only at points where it is itself present, and thus diverse hypotheses were evolved to explain electrical forces, that emanations flowed from the charged bodies and exerted a pressure when they impinged on bodies, and similar assumptions. But after Newton's theory of gravitation had begun to reap its victories the idea of a force acting directly at a distance gradually became a habit of thought. For it is, indeed, nothing more than a habit of thought when an idea impresses itself so strongly on minds that it is used as the last principle of explanation. It does not then take long for metaphysical speculation, often in the garb of philosophic criticism, to evolve the proof that the correct or accepted principle of explanation is a logical necessity and that its opposite cannot be imagined. But, fortunately, progressive empirical science does not, as a rule, trouble about this, and, when new facts demand it, it often has recourse to ideas that have been condemned. The development of the doctrine of electric and magnetic forces is an example of such a cycle of theories. At the beginning we see a theory of contiguous action based on *metaphysical* grounds; it is replaced by a theory of action at a distance on Newton's model. At the end this becomes transformed, owing to the discovery of new *facts*, into a general theory of contiguous action again. But this fluctuation is no sign of weakness. For it is not the pictures that are connected with the theories which are the essential features, but the empirical facts and their conceptual relationships. Yet if we follow these we see no fluctuation but only a continuous development full of inner logical consistency. We may justifiably omit the first theoretical attempts of pre-Newtonian times from the series because the facts were known too incompletely to furnish really convincing starting-points for theory. But the fact that the theory of action at a distance then arose in conformity with the model of Newtonian mechanics is founded quite naturally in the nature of electrical facts. A branch of research which had at its disposal only the experimental means of the 18th century could not do otherwise, on the ground of the observations possible at that time, than come to the decision that the electrical and the magnetic forces act at a distance in the same way as gravitation. Nowadays, too, it is absolutely permissible, from the point of view of the highly developed theories of contiguous action of Faraday and Maxwell, to represent electro- and magneto-statics by means of actions at a distance, and when properly used, they always lead to correct results.

The idea that electric forces act like gravitation at a distance was first conceived by Aepinus (1759). He even went so far as to regard gravitation and electricity as effects of the *same* fluid. He supposed, in the sense of the one-fluid theory, that matter devoid of electric fluid would repel other matter, but that there is always a little excess of fluid present that effects the gravitational attraction. Curiously enough he did not succeed in setting up the correct law for the dependence of electrical actions on the distance, but he was able to explain the phenomenon of influence qualitatively. This consists in a charged body acting attractively not only on other charged bodies but also on uncharged bodies, particularly on conducting bodies, for a charge of the opposite sign is induced on the side of the influenced body nearest the acting body, whereas a charge of the same sign is driven to the further side (Fig. 78); hence the attraction outweighs the repulsion.

The true law was presumably first found by Priestley, the discoverer of oxygen (1767). He discovered it by an ingenious indirect way which essentially carries more conviction with it than that of direct measurement. Independently of him Cavendish (1771) derived this law by the same method. But it receives its name from the physicist who first proved it by measuring the forces directly, namely, Coulomb (1785).

The argument of Priestley and Cavendish ran somewhat as follows :

If an electric charge is given to a conductor then it cannot remain in equilibrium in the interior of the conducting substance since particles of the same charge repel each other. Rather, they must tend to the outer surface at which they distribute themselves in a certain way so as to be in equilibrium.

Now experiment teaches very definitely that no electric field exists within a space that is enclosed on all sides by metallic walls, no matter how strongly the envelope is charged. The charges on the outer surface of the empty space must thus distribute themselves so that the force exerted at each point in the interior must vanish. Now, if the empty space has the particular form of a sphere, reasons of symmetry convince us that the charge can only be distributed uniformly over the surface. If ρ is the charge per unit area of surface (density of charge, then amounts of electricity $\rho f_1 + \rho f_2$ are on the two por-

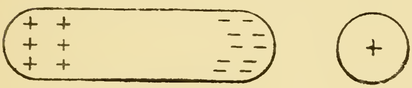


FIG. 78.

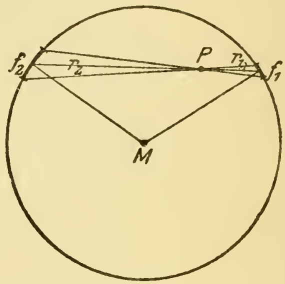


FIG. 79.

tions f_1, f_2 of the surface. The force that a small portion of surface f_1 of this kind exerts on a test-body P situated in the interior of the sphere and carrying the charge e is then $K_1 = e\rho f_1 R_1$, where R_1 denotes the force which is exerted between two units of charge placed at P and f_1, and which somehow depends on the distance r_1 between P and f_1. Now corresponding to each portion of surface f_1 there is an opposite portion f_2, which is obtained by connecting the points of the boundary of f_1 with P and producing these lines through P until they intersect the sphere. The two portions of area f_1 and f_2 are thus cut out of the surface of the sphere by the same double cone with its apex at P (Fig. 79), and the angles between them and the axis of the double cone are equal. The values of f_1 and f_2 are thus in the ratio of the squares of the distances from P :

$$\frac{f_2}{f_1} = \frac{r_2^2}{r_1^2}.$$

The charge ρf_2 on f_2 exerts the force $K_2 = e\rho f_2 R_2$ on P, where K_2 depends on r_2 in some way ; K_2 is of course oppositely directed to K_1.

It readily suggests itself to assume that all the forces acting on P can only then neutralize each other if the forces due to two opposite portions of area exactly counterbalance, that is, when $K_1 = K_2$. It is possible to prove this assumption, but that would take us too far here. If we take it for granted, then it follows that $f_1 R_1 = f_2 R_2$, or

$$\frac{R_1}{R_2} = \frac{f_2}{f_1} = \frac{r_2^2}{r_1^2}.$$

Accordingly $$R_1 r_1^2 = R_2 r_2^2 = c$$

where c is a quantity independent of the distance r. This determines R_1 and R_2, namely,

$$R_1 = \frac{c}{r_1^2}, \quad R_2 = \frac{c}{r_2^2}.$$

Hence, in general, the force R between two unit charges at a distance r apart must have the value

$$R = \frac{c}{r^2}.$$

In conformity with our convention about the unit of electric charge we must set $c = 1$. The force between two unit charges unit distance apart is to be equal to 1. With this convention the force that two bodies carrying the charges e_1 and e_2 and at a distance r apart exert on each other is

$$K = \frac{e_1 e_2}{r^2} \quad . \quad . \quad . \quad . \quad (45)$$

This is *Coulomb's law*. In its formulation we assume that the greatest diameter of the charged bodies is, of course, small compared with their distances apart. This restriction expresses that this law, just like the law of gravitation, is an idealized elementary law. To deduce from it the action of bodies of finite extent we must consider the electricity distributed over them to be divided into small parts, then calculate the effects of all the particles of the one body on all those of the others in pairs and sum them.

Formula (45) fixes the dimensions of quantity of electricity since we have for the repulsion of two equal charges $\frac{e^2}{r^2} = K$, that is, $e = r\sqrt{K}$, hence

$$[e] = [L\sqrt{K}] = \left[L\sqrt{\frac{ML}{T^2}}\right] = \left[\frac{L}{T}\sqrt{ML}\right].$$

This, at the same time, fixes the unit of charge in the C.G.S. system; it must be written $\frac{\text{cm.}\sqrt{\text{grm. cm.}}}{\text{sec.}}$.

The electric intensity of field E, defined by $K = eE$, has the dimensions

$$[E] = \left[\frac{K}{e}\right] = \left[\frac{K}{L\sqrt{K}}\right] = \left[\frac{\sqrt{K}}{L}\right] = \left[\frac{\sqrt{ML}}{LT}\right] = \left[\frac{1}{T}\sqrt{\frac{M}{L}}\right],$$

and its unit is $\frac{1}{\text{sec.}}\sqrt{\frac{\text{grm.}}{\text{cm.}}}$.

After Coulomb's law had been set up electrostatics became a mathematical science. Its most important problem is this: given the total quantity of electricity on conducting bodies, to calculate the distribution of charges on them under the action of their mutual influence and also the forces due to these charges. The development of this mathematical problem is interesting in that it very soon became changed from the original formulation based on the theory of action at a distance to a theory of pseudo-contiguous action, that is, in place of the summations of Coulomb forces there were obtained differential equations in which the intensity of field E or a related quantity called *potential* occurred as the unknown. But we cannot here further discuss these purely mathematical questions in which Laplace (1782), Poisson (1813), Green (1828), and Gauss (1840) have achieved meritorious results. We shall emphasize only one point: in this treatment of electrostatics, which is usually called the *theory of potential,* we are not dealing with a true theory of contiguous action in the sense which we attached to this expression above (IV, 6, p. 95). For the differential equations refer to the change in the intensity of field from place to place, but they contain no member that expresses a change in time. Hence they entail no transmission of electric force with finite velocity but, in spite of their differential form, they represent an instantaneous action at a distance.

The doctrine of *magnetism* developed in the same way as electrostatics. We may, therefore, express ourselves briefly. The most essential difference between these two regions of phenomena is that there are bodies that conduct electricity, whereas magnetism is always bound to matter and can only move with it.

A lozenge-shaped magnetized body, a *magnet needle,* has two *poles,* that is, points from which the magnetic force seems to start out, and again the law holds that like poles repel, unlike poles attract. If we break a magnet in halves, the two parts do not carry opposite magnetic charges, but each part receives a new pole at the broken surface and again represents a complete magnet with two equal but opposite poles. This holds, no matter into how many parts the magnet be broken.

From this it has been concluded that there are indeed two kinds of magnetism as in the case of electricity, that they cannot move freely, and that they are present in the smallest particles of matter, molecules, in equal quantities but separate. Thus each molecule is itself a small magnet with a North and a South pole (Fig. 80). The magnetization of a finite body consists in all the elementary magnets that were originally in complete disorder being brought into the same direction. Then the effects of the alternate North (+) and South (—) poles counterbalance except for those at the two end faces, from which therefore all the action seems to start.

By using a very long thin magnet needle it is possible to arrange so that in the vicinity of the one pole the force of the other becomes inappreciable. Hence in magnetism, too, we may operate with test-bodies, namely, with the poles of very long thin

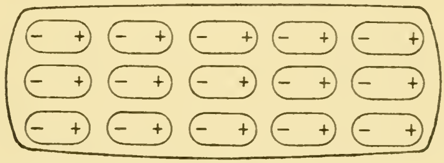


FIG. 80.

magnetic rods. These allow us to carry out all the measurements that we have already discussed in the case of electricity. We thus succeed in defining the *amount of magnetism* or the *pole strength* p and the magnetic intensity of field H. The magnetic force that a pole p experiences in the field H is

$$K = pH \quad . \quad . \quad . \quad . \quad (46)$$

The unit of pole is chosen so that two unit poles at unit distances apart exert the repulsive force 1 on each other. The law according to which the force between two poles p_1 and p_2 changes with the distance was also found by Coulomb from direct measurement. Just like Newton's law of attraction, it has the form

$$K = \frac{p_1 p_2}{r^2} \quad . \quad . \quad . \quad . \quad (47)$$

Clearly the dimensions of magnetic quantities are the same as those of the corresponding electric quantities, and their units have the same notation in the C.G.S. system.

The mathematical theory of magnetism runs fairly parallel with that of electricity. The most essential difference is that the true quantities of magnetism remain attached to the molecules, and that the measurable accumulations that condition the occurrence of poles in the case of finite magnets arise only owing to the summation of molecules that point in the same direction.

2. Voltaic Electricity and Electrolysis

The discovery of so-called contact electricity by Galvani (1780) and Volta (1792) is so well known that we may pass it by here. For however interesting Galvani's experiments with frogs' legs and the resulting discussion about the origin of electric charges may be, we are here more concerned with formulating conceptions and laws clearly. Hence we shall recount only the facts.

If two different metals be dipped into a solution (Fig. 81), say, copper and zinc into dilute sulphuric acid, the metals manifest electric charges that exert exactly the same action as frictional electricity. According to the fundamental law of electricity charges of both sign occur on the metals (poles) to the same amount. The system composed of the solution and the metals, which is also called *Voltaic element* or *cell,* thus has the power of separating the two kinds of electricity. Now, it is remarkable that this power is apparently inexhaustible, for if the poles are connected by a wire, so that their charges flow around and neutralize each other, then, as soon as the wire is again removed, the poles are still always charged. Thus the element continues to keep up the supply of electricity so long as the wire connexion is maintained. Hence a continuous flow of electricity must be taking place. How this is to be imagined in detail depends on whether the one-fluid or the two-fluid theory is supported. In the former case only one current is present, in the latter two opposite currents, one of each fluid, flow.

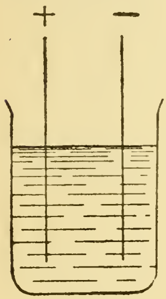


Fig. 81.

Now, the *electric current* manifests its existence by showing very definite effects. Above all it heats the connecting wire. Everyone knows this fact from the metallic threads in our electric glow-lamps. Thus the current continually produces heat-energy. Whence does the Voltaic element derive the power of producing electricity continually and hence thereby indirectly generating heat ? According to the law of conservation of energy, wherever one kind of energy appears during a process another kind of energy must disappear to the same extent.

The source of energy is the chemical process in the cell. The one metal dissolves as long as the current flows, whereas a constituent of the solution separates out on the other. Complicated chemical processes may take place in the solution itself. We have nothing to do with these but satisfy ourselves with the fact that the Voltaic element is a means of generating electricity in unlimited quantities and of producing considerable electric currents.

But we shall now have have to consider the reverse process, in which the electric current produces a chemical decomposition. For example, if we allow the current between two undecomposable wire leads (electrodes), say of platinum, to flow through slightly acidified water, the latter resolves into its components, hydrogen and oxygen, the hydrogen coming off at the negative electrode (cathode), the oxygen at the positive electrode (anode). The quantitative laws of this process of "*electrolysis*," discovered by Nicholson and Carlisle (1800), were found by Faraday (1832). The far-reaching consequences of Faraday's researches for the knowledge of the structure of matter are well known; it is not the consequences themselves that lead us to discuss these researches but the fact that Faraday's laws furnished the means of measuring electric currents accurately, and hence allowed the system of electromagnetic conceptions to be still further elaborated.

This experiment in electrolytic dissociation can be carried out not only with a Voltaic current, but just as well with a current discharge, which occurs when oppositely charged metallic bodies are connected by a wire. Care must indeed be taken that the quantities of electricity that take place in the discharge are sufficiently great. We have apparatus for storing electricity, so-called *condensers*, whose action depends on the induction principle, and which give such powerful discharges that measurable amounts are decomposed in the electrolytic cell. The amount of the charge that flows through the cell may be measured by the above discussed methods of electrostatics. Now, Faraday discovered the law that twice the charge produces twice the dissociation, three times the charge three times the dissociation, in short, that the amount m of dissociated substance (or of one of the products of dissociation) is proportional to the quantity e of electricity that has passed through the cell:

$$Cm = e.$$

The constant C also depends on the nature of the substances and of the chemical process.

A second law of Faraday regulates this dependence. It is known that chemical elements combine together in perfectly definite proportions to form compounds. The quantity of an element that combines with 1 grm. of the lightest element, hydrogen, is called its *equivalent weight.* For example, in water (H_2O) 8 grms. of oxygen (O) are combined with 1 grm. of hydrogen (H), hence oxygen has the equivalent weight 8. Now Faraday's law states that the same quantity of electricity that separates out 1 grm. of hydrogen is able to separate out an equivalent weight of every other element, thus, for example, 8 grms. of oxygen.

Hence the constant C need only be known for hydrogen, and then we get it for every other substance by dividing this value by the equivalent weight of the substance. For we have for 1 grm. of hydrogen

$$C_0 . 1 = e$$

and for any other substance with the equivalent weight μ

$$C\mu = e.$$

By dividing these equations we get

$$\frac{C}{C_0} = \frac{1}{\mu}, \text{ i.e. } C = \frac{C_0}{\mu}.$$

Thus $C_0 = e$ is the exact quantity of electricity that separates out 1 grm. of hydrogen. Its numerical value has been determined by exact measurements and amounts in the C.G.S. system to

$$C_0 = 2{\cdot}90 . 10^{14} \text{ units of charge per gramme} \quad . \quad . \quad (48)$$

Now we may combine Faraday's two laws into the one formula:

$$e = \frac{C_0}{\mu} m \quad . \quad . \quad . \quad . \quad (49)$$

Thus electrolytic dissociation furnishes us with a very convenient measurement of the quantity of electricity e that has passed through the cell during a discharge. We need only determine the mass m of a product of decomposition that has the equivalent weight μ and then we get the desired quantity of electricity out of equation (49). In this it is of course a matter of indifference whether this electricity is obtained from the discharge of charged conductors (condensers) or whether it comes from a Voltaic cell. In the latter case the electricity flows continuously with constant strength; the quantity that passes per unit of time through any cross-section of the conducting circuit, and hence also through the decomposing cell, is called the *intensity of current* or *current strength.* This may be measured simply by allowing the Voltaic current to flow for a unit of time (1 sec.) through the electrolytic cell and then determining the mass m of a product of dissociation. Then equation (49) again gives us the charge e, which is equal to the current strength. If the current flows not for 1 sec. but for t secs., then the quantity of electricity e and the mass m of each product of decomposition separated out is t times as great. Hence the intensity of current J is

$$J = \frac{e}{t} = \frac{C_0}{\mu} \cdot \frac{m}{t} \quad . \quad . \quad . \quad . \quad (50)$$

Its dimensions are

$$[J] = \left[\frac{e}{T}\right] = \left[\frac{L}{T}\sqrt{K}\right] = \left[\frac{L}{T^2}\sqrt{ML}\right]$$

and its unit is

$$\frac{\text{cm.}\sqrt{\text{grm. cm.}}}{\text{sec.}^2}$$

3. Resistance and Heat of Current

We must next occupy ourselves a little with the process of conduction or flowing itself. It has been customary to compare the electric current with the flowing of water in a conducting tube and to apply the conceptions there valid to the electrical process. If water is to flow in a tube there must be some driving force. If the water be allowed to flow down from a higher vessel through an inclined tube to a lower vessel, gravitation is the driving force (Fig. 82). This is so much the greater, the higher the upper surface of the water is above the lower. But the velocity of the current of water or its current strength depends not only on the amount of the impulse given by gravitation but also on the resistance that the water experiences in the conducting tube. If this is long and narrow the amount of water sent through per unit of time is less than in the case of a short wide tube. The current strength J is thus proportional to the difference of level V that drives the water and inversely proportional to the resistance W. We set

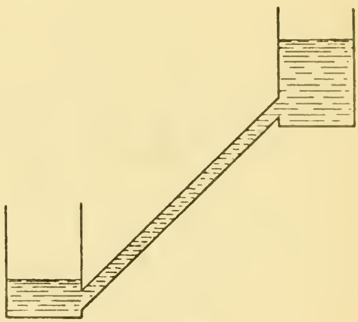


Fig. 82.

$$J = \frac{V}{W} \text{ or } JW = V \quad . \quad . \quad . \quad (51)$$

in which the unit of resistance chosen is that which allows the current of strength 1 to flow when the difference of level is 1.

G. S. Ohm (1826) applied precisely the same ideas to the electric current. The difference of level that effects the flow corresponds to the electric force. For a definite piece of wire of length l we must set $V = El$, where E is the field strength, which is regarded constant along the wire. For if the same electric field acts over a greater length of wire, it furnishes a stronger impulse to the flowing electricity. The force V is also called the *electromotive force* (difference of potential or level). It is moreover identical with the conception of electric potential which we mentioned above (p. 132).

Since the current-strength J and the electric intensity of field E, hence also the potential difference or electromotive force $V = El$, are measurable quantities, the proportionality between J and V expressed in Ohm's law may be tested experimentally.

The resistance W depends on the material and the form of the conducting wire ; the longer and thinner it is, the greater is W. If l is the length of the wire and q the size of the cross-section, then W is directly proportional to l, and inversely proportional to q. We set

$$\sigma W = \frac{l}{q} \text{ or } W = \frac{l}{\sigma q} \quad . \quad . \quad . \quad (52)$$

where the factor of proportionality σ depends further only on the material of the wire V and is called the *conductivity*.

If we substitute W from (52) and $V = el$ in (51), we get

$$JW = J\frac{l}{q\sigma} = V = El.$$

By cancelling l we get

$$\frac{J}{q\sigma} = E \text{ or } \frac{J}{q} = \sigma E.$$

But $\frac{J}{q}$ denotes the current strength per unit cross-section. This is called the *current density* and is denoted by i. We thus have

$$i = \sigma E \quad . \quad . \quad . \quad . \quad (53)$$

In this form Ohm's law is left with only one constant peculiar to the conducting material, namely, the conductivity σ, but in no other way depending on the form of the conducting body (wire).

In the case of insulators $\sigma = 0$. But ideal insulators do not exist. Very small traces of conductivity are always present except in a complete vacuum. There is an unbroken sequence known leading from bad conductors (such as porcelain, amber) over the so-called semi-conductors (such as water and other electrolytes), to the metals, which have enormously high conductivity.

We have already pointed out above that the current heats the conducting wire. The quantitative law of this phenomenon was found by Joule (1841). It is clearly a special case of the law of conservation of energy, in which electric energy becomes transformed into heat. Joule's law states that the heat developed per unit of time by the current J in traversing the potential difference V is

$$Q = JV \quad . \quad . \quad . \quad . \quad (54)$$

where Q is to be measured not in calories but in mechanical units of work. We shall make no further use of this formula, and state it here merely for the sake of completeness.

4. Electromagnetism

Up to that time electricity and magnetism had been regarded as two regions of phenomena that were similar in some respects, but quite separate and self-dependent. A bridge was eagerly sought between the two regions, but for a long time without success. At last Oersted (1820) discovered that magnet needles are deflected by Voltaic currents. In the very same year Biot and Savart discovered the quantitative law of this phenomenon, which Laplace formulated as an action at a distance. It is very important for us for the reason that in it there occurred a constant, peculiar to electromagnetism and of the nature of a velocity, which showed itself later to be identical with the velocity of light.

Biot and Savart established that the current flowing in a straight wire neither attracts nor repels a magnetic pole, but strives to drive it round in a circle around the wire (Fig. 83), so that the positive pole moves in the sense of a right-handed screw turned from below (contrary to the hands of a watch) about the (positive) direction of the current. The gravitative law can be brought into the simplest form by supposing the conducting wire to be divided into a number of short pieces of length l and writing down the effect of these current-elements, from which the effect of the whole current is obtained by summation. We shall do no more than state the law of a current element for the special case in which the magnetic pole lies in the plane that passes through the middle part of the element and is perpendicular to its direction (Fig. 84). Then the force that acts on the magnet pole of unit strength, i.e. the magnetic intensity of field H in this plane, is perpendicular to the line connecting the pole with the mid-point of the current-element, and is directly proportional to the current intensity J and to its length l, and inversely proportional to the square of the distance r:

$$cH = \frac{Jl}{r^2} \quad . \quad . \quad . \quad . \quad (55)$$

Outwardly this formula has again similarity with Newton's law of attraction or with Coulomb's law of electrostatics and magnetostatics, but the electromagnetic force has nevertheless a totally different character. For it does not act in the direction of the connecting line but perpendicular to it. The three

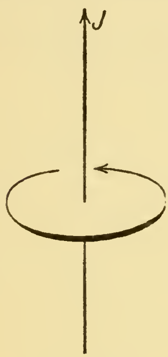


FIG. 83.

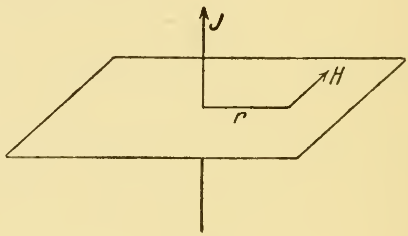


FIG. 84.

directions J, r, H are perpendicular to each other in pairs. From this we see that electrodynamic effects are intimately connected with the structure of Euclidean space; in a certain sense they furnish us with a natural rectilinear co-ordinate system.

The factor of proportionality c introduced in formula (55) is completely determined since the distance r, the current-strength J, and the magnetic field H are measurable quantities. It clearly denotes the strength of that current which when it flows through a piece of conductor of length 1 produces the magnetic field 1 at unit distance. It is customary and often convenient to choose in place of the unit of current that we have introduced (namely, the quantity of static electricity that flows through the cross-section per unit of time) and that is called the electrostatic unit, this current of strength c in electrostatic measure as the unit of current; it is then called the *electromagnetic unit of current.* Its use brings with it the advantage that the equation (55) assumes the simple form $H = \frac{JI}{r^2}$ or $J = \frac{Hr^2}{l}$, whereby the measurement of the strength of a current is reduced to that of two lengths and of a magnetic field. Most practical instruments for measuring currents depend on the deflection of magnets by currents or the converse, and hence give the current strength in electromagnetic measure. To express this in terms of the electrostatic measure of current first introduced the constant c must be known; for this, however, only one measurement is necessary.

Before we speak of the experimental determination of the quantity c, we shall get an insight into its nature by means of a simple dimensional consideration. According to (55) it is defined by $c = \frac{Jl}{Hr^2}$.

Now the following dimensional formulæ hold:

$$J = \left[\frac{e}{T}\right], \qquad H = \left[\frac{p}{R^2}\right]$$

hence the dimensions of c become

$$[c] = \left[\frac{eL}{pT}\right].$$

But we know that the electric charge e and the magnetic strength of pole p have the same dimensions because Coulomb's law for electric and magnetic force is exactly the same. Hence we get

$$[c] = \left[\frac{L}{T}\right],$$

that is, c has the dimensions of a velocity.

The first exact measurement of c was carried out by Weber and Kohlrausch (1856). These experiments belong to the most memorable achievements of physical precision measurements, not only on account of their difficulty but also on account of the far-reaching consequences of the result. *For the value obtained for c was* 3.10^{10} *cm./sec., which is exactly identical with the velocity of light.*

This equality could not be accidental. Numerous thinkers, above all Weber himself, the mathematicians Gauss and Riemann, and the physicists Neumann, Kirchhoff, Clausius felt the close relationship that the number $c = 3.10^{10}$ cm./sec. established between two great realms of science, and they sought to discover the bridge that necessarily led from electromagnetism to optics. Riemann came very near to solving the problem, but this was actually accomplished by Maxwell, after Faraday's wonderful and ingenious method of experimenting had brought to light new facts and new views. We shall next pursue this development.

5. Faraday's Lines of Force

Faraday came from no learned academy ; his mind was not burdened with traditional ideas and theories. His sensational rise from a book-binder's apprentice to the world-famous physicist of the Royal Society is well known. The world of his ideas, which arose directly and exclusively from the abundance of his experimental experiences, was just as free from the conventional scheme as his life. We discussed above his researches on electrolytic dissociation. His method of trying all conceivable changes in the conditions of experiment led him (1837) to insert a non-conductor like petroleum and turpentine between the two metal plates (electrodes) of the electrolytic cell in place of a conducting fluid (acid or solution of a salt). These non-conductors did not dissociate, but they were not without influence on the electrical process. For it is found that when the two metal plates are charged by a definite Voltaic battery with a definite potential difference, they take up totally different charges according to the substance that happens to be between them (Fig. 85). The non-conducting substance thus influences the power of taking up electricity or the *capacity* of the system of conductors composed of the two plates, which is called a *condenser.*

The discovery impressed Faraday so much that from that time onwards he gave up the usual idea that electrostatics was based on the direct action of electric charges at a distance, and developed a peculiar new interpretation of electric and magnetic phenomena, which is to be called a theory of contiguous action. What he learned from the experiment above described was the fact that the charges on the two metal plates do not simply act on each other through the intervening space, but that this intervening space plays an essential part in the action. From this he concluded that the action of this medium is propagated from point to point, and is thus an action by contact or a contiguous action.

We are familiar with the contiguous action of elastic forces in deformed rigid bodies. Faraday, who always kept to empirical facts, did indeed compare the electric contiguous action in non-conductors with elastic tensions, but he took good care not to apply the laws of the latter to electrical phenomena. He used the graphical picture of "*lines of force*" that run in the direction of the electric intensity of field from the positive charges through the insulator to the negative charges. In the case of a plate-condenser the lines of force are straight lines perpendicular to the planes of the plate (Fig. 86). Faraday regards the lines of force as the true substratum of electric events; for him they are actually material configurations that move about, deform themselves, and hereby bring about electrical effects. For Faraday the charges play a quite subordinate part, as the places at which the lines of force start out or end. He was strengthened in this view by those experiments which prove that in conductors the total electric charge resides on the surface whilst the interior remains quite free. To give a drastic proof of this he built a large cage fitted out all round with metal, into which he entered with sensitive

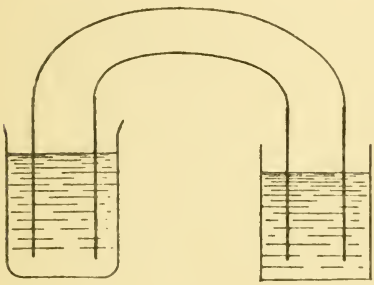


FIG. 85.

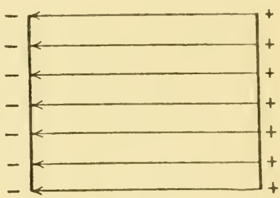


FIG. 86.

electrical measuring instruments. He then had the cage very strongly charged, and found that in the interior not the slightest influence of the charges was to be detected. Above (V, 1, p. 130) we used just this fact to derive Coulomb's law of action at a distance. But Faraday concluded from it that the charge was not the primary element of electrical events, and that it could not be imagined as a fluid which had the power of exerting forces at a distance. Rather, the primary element is the state of tension of the electric field in the non-conductors, which is represented by the picture of the lines of force. The conductors are in a certain sense holes in the electric field, and the charges in them are only fictitious conceptions, invented to explain the pressures and tensions arising through the strains in the field as actions at a distance. Among the non-conductors or dielectric substances there is also the *vacuum*, the *ether*, which we here again encounter in a new form.

This strange view of Faraday's at first found no favour among the physicists and mathematicians of his own time. The view of action at a distance was maintained, and this was possible even when the "dielectric" action of non-conductors discovered by Faraday was taken into account. Coulomb's law only needed to be altered a little; to every non-conductor there is assigned a peculiar constant ϵ, its "dielectric constant," which is defined by the fact that the force acting between two charges e_1, e_2 embedded in the non-conductor is smaller in the ratio $1 : \epsilon$ than that acting *in vacuo*:

$$K = \frac{1}{\epsilon} \frac{e_1 e_2}{r^2} \quad . \quad . \quad . \quad . \quad (56)$$

For a vacuum $\epsilon = 1$, for every other body $\epsilon > 1$

With this addition the phenomena of electrostatics could actually all be explained even when the dielectric properties of non-conductors were taken into account. We have already mentioned above that electrostatics had formally long ago passed over into a theory of pseudo-contiguous action, the so-called theory of potential. This likewise easily succeeded in assimilating the dielectric constant ϵ. Nowadays we know that this really already signified that the mathematical formulation of Faraday's conception of lines of force had been obtained. But as this method of potential was regarded only as a mathematical artifice, the antithesis between the classical theory of action at a distance and Faraday's idea of contiguous action still remained.

Faraday developed precisely similar views about magnetism. He discovered that the forces between two magnet poles likewise depend on the medium that happens to lie between them, and this again led him to the view that the magnetic forces, just like the electric forces, are produced by a peculiar state of tension in the intervening media. The lines of force served him to represent these tensions. They can, as it were, be made actually visible by scattering iron filings over a sheet of paper and holding the latter closely over a magnet (Fig. 87).

The theory of action at a distance leads to the formal introduction of a constant characteristic of the substance, the magnetic penetrability or *permeability* μ, and gives Coulomb's law in the altered form:

$$K = \frac{1}{\mu} \frac{p_1 p_2}{r^2} \quad . \quad . \quad . \quad . \quad (57)$$

Physicists have not, however, remained satisfied with this formal addition, but have devised a molecular mechanism that makes the magnetic and dielectric power of polarization intelligible. We have already seen above that the properties of magnets lead us to regard their molecules themselves as small elementary magnets that are made to point in parallel directions by the process of magnetization. It is assumed that they retain this parallelism of themselves, say, through frictional resistances. Now it may be assumed that in the case of most bodies that do not occur as permanent magnets this friction is wanting. The parallel position is then indeed produced by an external magnetic field, but will at once disappear if the field is removed. Such a substance will then be a magnet only so long as an external field is present. But it need not even be assumed that the molecules are permanent magnets that assume a parallel position. If each molecule contains the two magnetic fluids, then they will separate under the action of the field and the molecule will become a magnet of

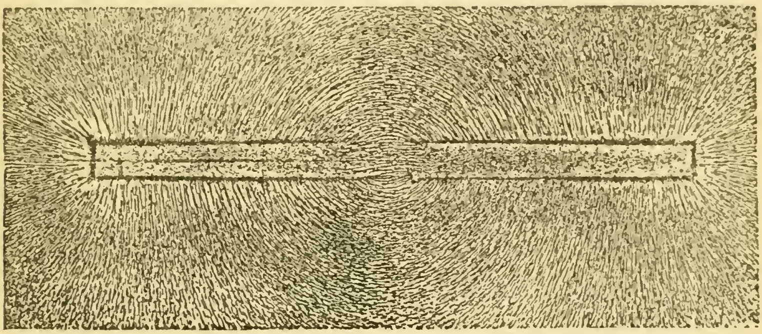


FIG. 87.

itself. But this induced magnetism must have exactly the action that the formal theory describes by introducing the permeability. Between the two magnet-poles (N, S) in such a medium there are formed chains of molecular magnets whose opposite poles everywhere compensate each other in the interior, but end with opposite poles at N and S, and hence weaken the actions of N and S (Fig. 88). (The converse effect, namely, strengthening, also occurs, but we shall not enter into its interpretation.)

Exactly the same as has just been illustrated for magnetism may also be imagined for electricity. A dielectric, in this view, is composed of molecules that are either electric dipoles of themselves and assume a parallel position in an external field or that become dipoles through the separation of the positive and negative electricity under the action of the field. Between two plates of a condenser (Fig. 89) chains of molecules again form whose charges compensate each other in the interior but not on the plates. Through this a part of the charge on the plates is itself neutralized, and a new charge has to be imparted to the plates to charge them up to a definite tension or potential. This explains how the polarizable dielectric increases the receptivity or capacity of the condenser.

According to this idea of the theory of action at a distance the effect of the dielectric is an indirect one. The field in the vacuum is only an abstraction. It signifies the geometrical distribution of the force that is exerted on an electric test-body carrying a unit charge. But the field in the dielectric is in a state of real physical change, the molecular displacement of the two kinds of electricity.

Faraday's theory of contiguous action knows no such difference between the field in the ether and in insulating matter.

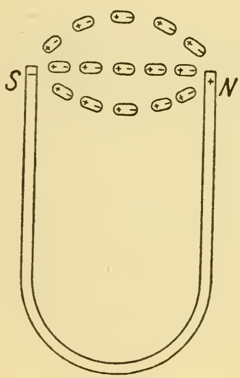


FIG. 88.

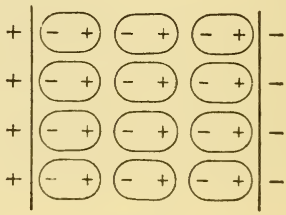


FIG. 89.

Both are dielectrics. For the ether the dielectric constant $\epsilon = 1$, for other insulators ϵ differs from 1. If the graphical picture of electric displacement is correct for matter, it must also hold for the ether. This idea plays a great part in the theory of Maxwell, which is essentially nothing else than the translation of Faraday's idea of lines of force into the exact language of mathematics. Maxwell assumes that in the ether, too, the genesis of an electric or a magnetic field is accompanied by " displacements " of the fluids. It is not necessary for this purpose to imagine the ether to have an atomic structure, yet Maxwell's idea comes out most clearly if we imagine ether molecules which become dipoles just like the material molecules in the field. The field is not, however, the *cause* of the polarization, but the displacement is the *essence* of the state of tension which we call electric field. The chains of ether molecules *are* the lines of force and the charges at the surface of the conductors *are nothing more* than the end-charges of these chains. If there are material molecules present besides the ether particles, the polarization becomes strengthened and the charges at the ends become greater.

Are Faraday's and Maxwell's ideas or those of the theory of action at a distance right ?

So long as we confine ourselves to electrostatic and magnetostatic phenomena, both are fully equivalent. For the mathematical expression of Faraday's idea is what we have called a theory of pseudo-contiguous action, because it does, indeed, operate with differential equations but recognizes no finite velocity of propagation of tensions. Faraday and Maxwell, however, themselves disclosed those events which, analogously to the inertial effects of mechanics, effect the delay in the transference of an electromagnetic state from point to point and hence bring about the finite velocity of propagation. These events are the magnetic induction and the displacement current.

6. Magnetic Induction

After Oersted had discovered that an absolute current produces a magnetic field and Biot and Savart had formulated this fact as an action at a distance, Ampère discovered (1820) that two Voltaic currents exert forces on each other, and he in turn succeeded in expressing the law underlying this phenomenon in the language of the theory of action at a distance. This discovery had far-reaching consequences, for it made it possible to reduce magnetism to a form of electricity. According to Ampere small closed currents are supposed to flow in the molecules of magnetized bodies. He showed that such currents behaved exactly like elementary magnets. This idea has stood the test of examination ; from his time onwards magnetic fluids became superfluous. Only electricity was left, which when at rest produced the electrostatic field, and when flowing the magnetic field besides. Ampère's discovery may also be expressed thus : According to Oersted a wire in which the current J_1 is flowing produces a magnetic field in its neighbourhood. A second wire in which the current J_2 is flowing then experiences force effects in this magnetic field. Thus this field clearly tends to deflect or accelerate flowing electricity.

Hence the following question suggests itself. Can the magnetic field also set electricity that is at rest into motion ? Can it produce or " induce " a current in the second wire which is initially without a current ?

Faraday found the answer to this question (1831). He discovered that a static magnetic field has not the power of producing a current, but that one arises as soon as the magnetic field is changed. For example, when he suddenly approached a magnet to a closed conducting wire, a current flowed in the wire so long as the motion lasted; or when he produced the magnetic field by means of a primary current a short impulse of current occurred in the secondary wire whenever the first current was started or stopped.

From this it is clear that the induced electric force depends on the velocity of alteration of the magnetic field in time. Faraday succeeded in formulating the quantitative law of this phenomenon with the help of his lines of force. We shall give it such a form that its analogy with Biot and Savart's law comes out clearly. We imagine a bundle of parallel lines of magnetic force that constitute a magnetic field H. We suppose a circular conducting wire placed around this sheath (Fig. 90). If the intensity of field H changes in the small interval of time t by the amount $\mathbf{h}$ we call $\frac{\mathbf{h}}{t}$ its velocity of change or the change in the number of lines of force. If we represent the lines of force as chains of magnetic dipoles (which, however, according to Ampère, is not allowed), then in the change of H a displacement of the magnetic quantities will occur in every ether molecule, or a "magnetic displacement current" will occur whose current strength per unit of area or current density is given by $j = \frac{\mathbf{h}}{t}$. If the field H is not in the ether but in a substance of permeability μ, the density of the magnetic displacement current is $j = \mu\frac{\mathbf{h}}{t}$. Thus the magnetic current $J = qj = q\mu\frac{\mathbf{h}}{t}$ passes through the cross-section q, that is, through the surface of the circle formed by the conducting wire.

Now, according to Faraday, this magnetic current produces all around it an electric field E, which encircles the magnetic current exactly as the magnetic field H encircles the electric current in Oersted's experiment, only in the reverse direction. It is this electric field E that drives the induced current around in the conducting wire; it is also present even if there is no conducting wire in which the current can form.

We see that the magnetic induction of Faraday is a perfect parallel to the electromagnetic discovery of Oersted. The quantitative law, too, is the same. According to Biot and Savart the magnetic field H produced by a current-element of length l and of strength J (Cf. Fig. 84, p. 140) in the middle plane perpendicular to the element is perpendicular to the connecting line r and to the current direction, and has the value $H = \frac{Jl}{cr^2}$ [Formula (55), p. 140].

Here exactly the same holds when electric and magnetic quantities are exchanged and when the sense of rotation is reversed (Fig. 91). The induced electric intensity of field in the central plane is given by $E = \frac{Jl}{cr^2}$.

In it the same constant c, the ratio of the electromagnetic to the electrostatic unit of current occurs, which was found by Weber and Kohlrausch to be equal to the velocity of light. It can easily be seen from considerations of energy that this must be so.

A great number of the physical and technical applications

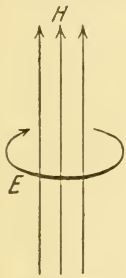


FIG. 90.

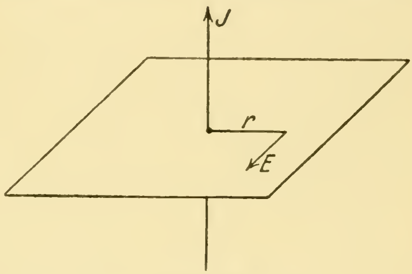


FIG. 91.

of electricity and magnetism depend on the law of induction. The transformer, the induction coil, the dynamo, and innumerable other apparatus and machines are appliances for inducing electric currents by means of changing magnetic fields. But however interesting these things may be, they do not lie on our road of investigation, the final goal of which is to examine the relationship of the ether with the space problem. Hence we turn our attention at once to the representation of Maxwell's theory, whose object was to combine all known electromagnetic phenomena to one uniform theory of contiguous action.

7. Maxwell's Theory of Action by Contact

We have already stated above that soon after Coulomb's law had been set up electrostatics and magnetostatics were brought into the form of a theory of pseudo-contiguous action. Maxwell now undertook to fuse this theory with Faraday's ideas, and to elaborate it so that it also included the newly discovered phenomena of dielectric and magnetic polarization, of electromagnetism, and magnetic induction.

Maxwell took as the starting-point of his theory the idea already mentioned above that an electric field E is always accompanied by an electric displacement ϵE not only in matter, for which ϵ is greater than 1, but also in the ether, where $\epsilon = 1$. We explained above how the displacement can be visualised as the separation and flowing of electric fluids in the molecules.

The first fact that Maxwell established was that in the light of this idea of displacement Coulomb's law was essentially nothing more than an inference from the law of indestructibility of electricity.

Let us imagine a metallic sphere embedded in a medium whose dielectric constant is ϵ (Fig. 92). In this sphere we construct a concentric sphere of radius 1 and another of radius r. Now, let the metal sphere be charged with an amount of electricity $+ e$. Then, according to Maxwell, a displacement of the positive electricity outwards must occur in every molecule in order that the amount of electricity contained in any arbitrary volume remain constant. And the amount of electricity transported across the surface of a sphere of radius 1 is to be measured by ϵE according to Maxwell. The same amount of electricity will pass through every concentric sphere since otherwise an accumulation of charges would occur in the dielectric. And since the surfaces of two spheres are in the ratio of the squares of the radii, an amount of electricity $r^2 \epsilon E$ passes through the sphere of radius r.

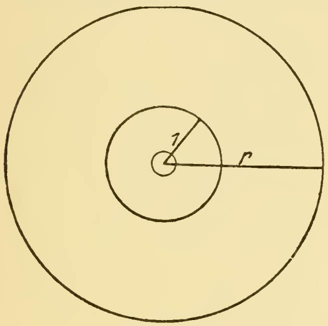


FIG. 92.

Now this must also be exactly equal to the charge e of the metal sphere at which the displacement comes to an end; thus we have $r^2 \epsilon E = e$, or

$$E = \frac{e}{\epsilon r^2}.$$

But this is Coulomb's law in the generalized form (56), (p. 144); E is the force exerted by the charge e on unit charge at the distance r.

If we are dealing, not with spheres, but with arbitrary charged bodies, Maxwell's fundamental idea still remains the same: the field is determined by the condition that the displacement ϵE of the electricity outwards in the delectric or the "divergence" of ϵE (div. ϵE) across any arbitrarily small closed surface just compensates the charges that occur in the interior of the surface. By denoting the charge per unit of volume or the *density of charge* of the electricity by p, we write symbolically

$$\text{div.}\,\epsilon E = p \quad . \quad . \quad . \quad . \quad (58)$$

This is to serve us only as a mnemonic for the law formulated above. But Maxwell showed that it is possible to derive a definite differential expression for the conception of divergence. Hence to mathematicians formula (58) signifies a differential equation, a law of contiguous action.

Exactly the same considerations apply to magnetism, but with one important difference: according to Ampère no real magnets exist, no magnetic quantities, but only electromagnets. The magnetic field is always to be produced by electric currents, whether they be conduction currents in wires or molecular currents in the molecules. From this it follows that the magnetic lines of force never end, that is, they either merge into themselves again or stretch to infinity. This is so in the case of an electromagnet, a coil through which a current is flowing (Fig. 93); the magnetic lines of force run rectilinearly through the interior of the coil, partly joining outside and partly going off to infinity. If we consider the coil enclosed between two planes A and B, then just as much "magnetic displacement" μH will enter through A as goes out through B. As, by the way, the displacement picture is unsuitable in this case, we usually say *magnetic induction* instead of displacement. Hence just as many lines of force will go out through any closed surface as enter into it, or the total divergence of magnetism through an arbitrary closed surface is nil:

$$\text{div.}\,\mu H = 0 \quad . \quad . \quad . \quad . \quad (59)$$

This is Maxwell's formula of contiguous action for magnetism.

We now come to Biot and Savart's law of electromagnetism. To convert this into a law of contiguous action we suppose the electric current not to be flowing in a thin wire, but to be distributed uniformly with the density $i = \frac{J}{q}$ over a circular cross section q, and we then ask what is the magnetic intensity of field H at the edge of the cross section (Fig. 94). But by Biot and Savart's law this is everywhere in the direction of the tangent to the circle and, according to formula (55), (p. 140), it has the value $H = \frac{Jl}{cr^2}$, where r is the radius of the circle, and l the length of the current element. Now the cross section, being circular, is πr^2, hence we may write formula (55) thus: $\frac{cH}{\pi l} = \frac{J}{\pi r^2} = \frac{J}{q} = i$, and this holds for every cross section, however small, and for every length, however short. On the left, then, there is a certain differential quantity of the magnetic field, and the law states that this quantity is proportional to the current-density. We cannot here carry out the mathematical investigation as to how this differential quantity is formed. It has to take into account not only the intensity but also the direction of the magnetic field, and since this encircles or curls round the direction of the current, the dif-

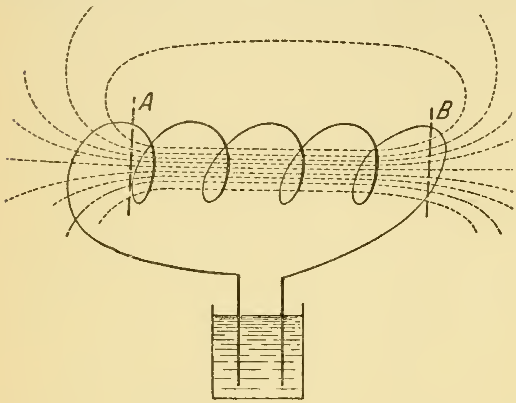


FIG. 93.

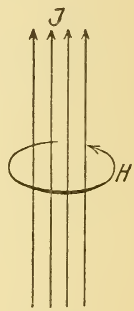


FIG. 94.

ferential operation is called "curl" of the field H (written, curl H). Accordingly we write symbolically

$$c \text{ curl } H = i \quad . \quad . \quad . \quad . \quad (60)$$

and again regard this formula only as a mnemonic for the relationships between the intensity and direction of the magnetic field H and the intensity of current i. To the mathematician, however, it is a differential equation of the same kind as the law (58).

Now exactly the same holds for magnetic induction, but we shall write the opposite sign to indicate the opposite sense of rotation:

$$c \text{ curl } E = -j \quad . \quad . \quad . \quad (61)$$

The four symbolic formulæ (58) to (61) show wonderful symmetry. Formal agreement of this kind is by no means a matter of indifference. It exhibits the underlying simplicity of phenomena in nature, which remains hidden from direct perception owing to the limitations of our senses, and reveals itself only to our analytical faculty.

8. The Displacement Current

But this symmetry is not perfect; for i denotes the density of the electric current of conduction, that is, a transportation or convection of electric currents along finite distances, whereas j is the time change of the magnetic field, and can be interpreted as a displacement current only on the basis of the very artificial hypothesis of ether dipoles.

Now Maxwell remarked (1864) that what sufficed for the magnetic field should hold no less for the electric field. The idea of dipoles compels us also to assume a *dielectric displacement current,* which flows in non-conductors when the electric field E varies. If **e** is the change of E in the time t, then the density of the dielectric displacement current must be set equal to $\epsilon\frac{\mathbf{e}}{t}$.

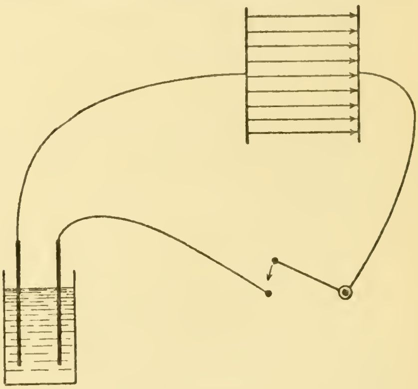


Fig. 95.

This Maxwellian theory, which seems almost trivial in our description, is of the greatest importance, for it became the key to the electromagnetic theory of light. We shall make its meaning clear by considering a concrete example. Suppose the poles of a galvanic cell to be connected with the plates of a condenser by means of two wires, and let there be a key in one of the two connexions (Fig. 95). If the latter is depressed, a short current flows which charges up the two plates of the condenser; between them an electric field E is thereby produced. Before Maxwell's time, this phenomenon was regarded as an "open circuit." Maxwell, however, asserts that during the growth of the field E a displacement current flows between the condenser plates, and the current becomes added to the conduction current so that the circuit becomes complete. So soon as the condenser plates are charged, both currents, the conduction and the displacement current, cease.

Now, the essential point is that Maxwell affirms that, just like the conduction current, the displacement current also produces a magnetic field according to Biot and Savart's law. That this is actually so has not only been proved by the success of Maxwell's theory in predicting numerous phenomena, but was also confirmed directly later by experiment.

In a semi-conductor a conduction and a displacement current will be present simultaneously. For the former, Ohm's law, $i = \sigma E$, holds; (53) (p. 138), for the latter, Maxwell's law, $i = \frac{\epsilon \mathbf{e}}{t}$. If both are present simultaneously we thus have $i = \epsilon\frac{\mathbf{e}}{t} + \sigma E$. There is no conduction current for magnetism, we always have $j = \mu\frac{\mathbf{h}}{t}$ in that case. If we insert this in our symbolic equations (58) to (61), we get:

$$\left\{\begin{array}{ll} (a)\ \text{div.}\ \epsilon E = \rho & (c)\ c\ \text{curl}\ H - \epsilon\frac{\mathbf{e}}{t} = \sigma E \\ (b)\ \text{div.}\ \mu H = 0 & (d)\ c\ \text{curl}\ E + \mu\frac{\mathbf{h}}{t} = 0 \end{array}\right. \qquad (62)$$

These are *Maxwell's laws,* which have remained the foundation of all electromagnetic and optical theories up to our own time. To the mathematician they are perfectly definite differential equations. To us they are mnemonics which state:

(*a*) Wherever an electric charge occurs, an electric field arises of such a kind that in every volume the charge is exactly compensated by the displacement.

(*b*) Through every closed surface just as much magnetic displacement passes outwards as comes inwards.

(*c*) Every electric current, be it a conduction or a displacement current is surrounded by a magnetic field.

(*d*) A magnetic displacement current is surrounded by an electric field in the reverse sense.

Maxwell's "field equations," as they are called, constitute a true theory of contiguous action or action by contact, for, as we shall presently see, they give a finite velocity of propagation for electromagnetic forces.

At the time when they were set up, however, faith in direct action at a distance, according to the model of Newtonian attraction, was still so deeply rooted that a considerable interval elapsed before they were accepted. For the theory of action at a distance had also succeeded in mastering the phenomena of induction by means of formulæ. This was done by assuming that moving charges exert in addition to the Coulomb attraction also the special actions at a distance that depend on the amount and direction of the velocity. The first hypotheses of this kind were due to Neumann (1845). A particularly famous law is that which was set up by Wilhelm Weber (1846); similar formulæ were given by Riemann (1858), and Clausius (1877). All these theories have in common the idea that all electrical and magnetic actions are to be explained by means of forces between elementary electrical charges or, as we say nowadays, "electrons." They were thus precursors of the present-day theory of electrons, with an essential factor omitted, however, namely, the finite velocity of propagation of the forces. These theories of electrodynamics, based on action at a distance, gave a complete explanation of the motive forces and induction currents that occur in the case of closed conduction currents. But in the case of "open" circuits, that is, condenser charges and discharges, they were doomed to failure, for here the displacement currents come into play, of which the theories of action at a distance know nothing. It is to Helmholtz that we are indebted for appropriate experimental devices, allowing us to decide between the theories of action at a distance and action by contact. He succeeded in carrying the experiment out with a certain measure of success, and he himself became one of the most zealous pioneers of Maxwell's theory. But it was his pupil, Hertz, who secured the victory for Maxwell's theory by discovering electromagnetic waves.

9. The Electromagnetic Theory of Light

We have already mentioned above (V, 4, p. 141) the impression which the coincidence, established by Weber and Kohlrausch, of the electromagnetic constant c with the velocity of light made upon the physicists of the day. And there were still further indications that there is an intimate relation between light and electromagnetic phenomena. This was shown most strikingly by Faraday's discovery (1834) that a polarized ray of light which passes by a magnetized body is influenced by it. When the beam is parallel to the magnetic lines of force its plane of polarization becomes turned. Faraday himself concluded from this that the luminiferous ether and the carrier of electromagnetic lines of force must be identical. Although his mathematical powers were not sufficient to allow him to convert his ideas into quantitative laws and formulæ, his conceptions were very abstract and were in no wise confined within the narrow limits of the trivial view which accepted as known what was familiar. Faraday's ether was no elastic medium. It derived its properties, not by analogy, from the *apparently* known material world, but from exact experiments and from the consequent relationships that were *really* known. Maxwell continued Faraday's work. His talents were akin to those of Faraday, but they were supplemented by a complete mastery of the mathematical means available at the time.

We shall now make clear to ourselves that the propagation of electromagnetic forces with finite velocity arises out of Maxwell's field laws (62). In doing so we shall confine ourselves to events that occur *in vacuo* or in the ether. The latter has no conductivity, that is, $\sigma = 0$, and no true charges, that is, $\rho = 0$; and its dielectric constant and permeability are equal to 1, that is, $\epsilon = 1$, $\mu = 1$. The first two field equations (62) then assert that

$$\text{div. E} = 0 \qquad \text{div. H} = 0 \quad . \quad . \quad . \quad (63)$$

or that all lines of force are either closed or run off to infinity. Even if only to obtain a rough picture of the processes we shall imagine to ourselves individual closed lines of force.

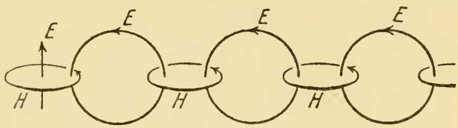


FIG. 96.

The other two field equations are then

$$\frac{\mathbf{e}}{t} = c \text{ curl H} \qquad \frac{\mathbf{h}}{t} = -c \text{ curl E} \quad . \quad . \quad (64)$$

We now assume that, somewhere in a limited space, there is an electric field E which alters by the amount **e** in the small interval of time t; then $\frac{\mathbf{e}}{t}$ is its rate of change. According to the first equation, a magnetic field immediately coils itself around this electric field, and its rate of change is also proportional to $\frac{\mathbf{e}}{t}$. The magnetic field, too, will alter in time, say, by **h**, during a successive small interval, t. Again, in accordance with the second equation, its rate of change $\frac{\mathbf{h}}{t}$ immediately induces an interwoven electric field. In the following interval of time the latter again induces an encircling magnetic field, according to the first equation, and so this chain-like process continues with finite velocity (Fig. 96).

This is, of course, only a very rough description of the process, which actually propagates itself in all directions continuously. Later we shall sketch a better picture.

What interests us particularly here is the following. We know from mechanics that the finite velocity of propagation of elastic waves is due to the delays that occur as a result of the mass inertia which acts when the forces are transmitted in the body from point to point. The mass inertia, however, is determined by the acceleration, and this is the rate of change of the velocity; $b = \frac{w}{t}$, where w is the change of the velocity $v = \frac{x}{t}$ in the small time t. Thus the retardation is clearly due to the *double* differentiation.

Now, the case is exactly the same here. The rate of change of the electric field $\frac{\mathbf{e}}{t}$ first determines the magnetic field H, and then the rate of change $\frac{\mathbf{h}}{t}$ of the latter determines the electric field E at a neighbouring point. The advance of the electric field alone from point to point is thus conditioned by two differentiations with respect to time, that is, by an expression which is formed quite analogously to acceleration. It is due to this alone that electromagnetic waves exist. If one of the two partial effects were to occur without loss of time, no propagation of the electric force in the form of waves would occur. This helps us to realize the importance of Maxwell's displacement current, for it is just this rate of change $\frac{\mathbf{e}}{t}$ of the electric field.

We shall now give a picture of the propagation of an electromagnetic wave which will be truer to the actual process. Let two metal spheres have strong but opposite equal charges $+ e$ and $- e$, so that a strong electric field exists between them. Next let a spark occur between the spheres. The charges then neutralize each other, the field collapses at a great rate of change $\frac{\mathbf{e}}{t}$. The figure shows how the magnetic and electric lines of force then encircle each other alternately (Fig. 97). In our diagram the magnetic lines of force are drawn only in the medium plane between the spheres, the electric lines of force only in the plane of the paper, perpendicular to the medium plane. The whole figure is, of course, spherically symmetrical about the line connecting the centres of spheres. Each successive loop of the lines of force is weaker than its immediate predecessor, because it lies further outwards and has a bigger circumference. Accordingly, the inner part of a loop of electric force does not quite counterbalance the outer part of its predecessor, particularly as it enters into action a little late.

If we pursue the process along a straight line which is perpendicular to the line connecting the centres of the spheres, say along the x-axis, then we see that the electric and magnetic forces are always perpendicular to this axis; moreover, they are perpendicular to each other. This is true of any direction of propagation. Thus, the electromagnetic wave is rigorously transversal. Furthermore, it is polarized, but we still have the choice of regarding either the electric or the magnetic intensity of field as the determining factor of the vibration.

It is beyond our scope here to prove that the velocity of propagation is exactly equal to the constant c that occurs in the formulæ. Yet it is in itself probable, for we know that c has the dimensions of a velocity. Further, since according to Weber and Kohlrausch the value of c is equal to that of the velocity of light, c, Maxwell was able to conclude that *light waves are nothing other than electromagnetic waves.*

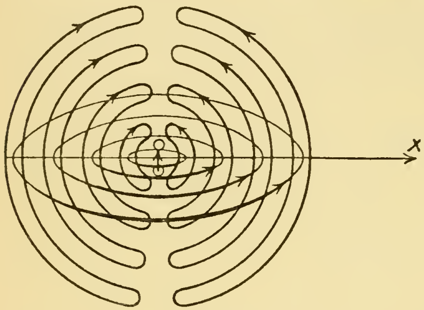


FIG. 97.

One of the inferences which Maxwell drew was soon confirmed experimentally to a certain extent, for he calculated the velocity of light c_1 in an insulator which was not appreciably magnetizable ($\mu = 1$, $\sigma = 0$). This velocity can then depend, besides on c, only on the dielectric constant ϵ; for when $\mu = 1$ and $\sigma = 0$ this dielectric constant is the only constant that occurs in the formulæ (62). Maxwell found that $c_1 = \frac{c}{\sqrt{\epsilon}}$. This leads to the value $n = \frac{c}{c_1} = \sqrt{\epsilon}$ for the refractive index.

Thus it should be possible to determine the refrangibility of light from the dielectric constant as given by purely electrical measurements. For some gases, for example, hydrogen, carbon dioxide, air, this is actually the case, as was shown by L. Boltzmann. For other substances Maxwell's relation $n = \sqrt{\epsilon}$ is not correct, but in all these cases the refractive index is not constant but depends on the colour (frequency) of the light. This shows that dispersion of the light introduces a disturbing effect. We shall return to this later and shall deal with it from the point of view of the theory of electrons. At any rate, it is clear that the dielectric constant as determined statically agrees more closely with the square of the refractive index the slower the vibrations or the longer the waves of the light that is used; infinitely slow waves are, of course, identical with a stationary state. Comparatively recent researches into the region of longest heat and light rays (in the infra-red) by Rubens have completely confirmed Maxwell's formula.

Concerning the more geometrical laws of optics, reflection, refraction, double refraction, and polarization in crystals, and so forth, the electromagnetic theory of light causes to vanish all the difficulties that were quite insuperable for the theories of the elastic ether. In the latter, the greatest obstacle was the existence of longitudinal waves, which appeared when light crossed the boundary between two media, and which could be removed only by making quite improbable hypotheses about the constitution of the ether. The electromagnetic waves are always strictly transversal. Thus this difficulty vanishes. Maxwell's theory is almost identical formally with the ether theory of MacCullagh, as we mentioned above (IV, 6, p. 101); without repeating the calculations we can take over most of his deductions.

We cannot here enter further into the later development of electrodynamics. The bond between light and electromagnetism became ever closer. New phenomena were continually being discovered, which showed that electric and magnetic fields exerted an influence on light. Everything subjected itself to Maxwell's laws, the certainty of which continued to grow.

But the striking proof of the oneness of optics and electrodynamics was given by Heinrich Hertz (1888) when he showed that the velocity of propagation of electromagnetic force was finite, and when he actually produced electromagnetic waves. He made sparks jump across the gaps between two charged spheres, and by this means generated waves such as are represented by our diagram (Fig. 97). When they encountered a circular wire with a small gap in its circuit they called up in it currents which were manifested by means of small sparks at the gap. Hertz succeeded in reflecting these waves and in making them interfere. This enabled him to measure their wave-length and to calculate their velocity, which came out exactly equal to c, that of light. This directly confirmed Maxwell's hypothesis. Nowadays the Hertzian waves of the great wireless stations travel over the earth without cessation, and bear their tribute to the two great investigators, Maxwell and Hertz, of whom the one predicted their discovery and the other actually produced them.

10. The Electromagnetic Ether

From this time onwards there was only *one* ether, which was the carrier of all electric, magnetic, and optical phenomena. We know its laws, Maxwell's field equations, but we know little of its constitution. Of what do the electromagnetic fields actually consist, and what is it that executes vibrations in the waves of light?

We recall to mind that Maxwell took the conception of displacement as the foundation of his argument, and we interpreted this visually as meaning that in the smallest parts or molecules of the ether, just as in the molecules of matter, an actual displacement and a separation of the electric (or magnetic) fluid occurs. So far as this idea concerns the process of electric polarization of *matter*, it is well founded; and it is also adopted in the newer modification of Maxwell's theory, the theory of electrons, for numerous experiments have rendered certain that matter is constituted molecularly, and that every molecule carries displaceable charges. But this is by no means the case for the free ether; here Maxwell's idea of displacement is purely hypothetical, and its only value is that it visualizes the abstract laws of the field.

These laws state that with every change of displacement in time there is associated an electromagnetic field of force which arises. Can we form a mechanical picture of this relationship?

Maxwell himself indicated mechanical models for the constitution of the ether, and applied them with success at that time. Lord Kelvin was particularly inventive in this direction, and strove unceasingly to comprehend electromagnetic phenomena as actions of concealed mechanical motions and forces.

The rotational character of the relationship between electric currents and magnetic fields, and its reciprocal character suggests to us to regard the electric state of the ether as a linear displacement, the magnetic state as a rotation about an axis, or conversely.

In this way we arrive at ideas that are related to MacCullagh's ether theory. According to this the ether was not to generate elastic resistances against distortions in the ordinary sense, but resistances against the absolute rotation of its elements of volume. It would take us much too far to count up the numerous and sometimes very fantastic hypotheses that have been put forward about the constitution of the ether.

If we were to accept them literally, the ether would be a monstrous mechanism of invisible toothed wheels, gyroscopes and gears intergripping in the most complicated fashion, and of all this confused mass nothing would be observable but a few relatively simple forces which would present themselves as an electromagnetic field.

There are also less cumbersome, and, in some cases, ingenious theories in which the ether is a fluid whose rate of flow represents, say, the electric field, and whose vortices represent the magnetic field. Bjerknes has sketched a theory in which the electric charges are imagined as pulsating spheres in the ether fluid, and he has shown that such spheres exert forces on each other which exhibit considerable similarity with the electromagnetic forces.

If we next inquire into the meaning and value of such theories, we must grant them the credit of having suggested, even if rather seldom, new experiments, and of having led to the discovery of new phenomena. More often, it is true, elaborate and laborious experimental researches have been instituted to decide between two ether theories, both of which were equally improbable and fantastic. In this way much effort has been wasted without reason. Even nowadays there are some people who regard the mechanical explanation of the electromagnetic ether as a postulate of reason. Such theories still continue to crop up, and, naturally, they become more and more abstruse since the abundance of the facts to be explained grows, and hence the difficulty of the task increases without cessation.

Heinrich Hertz consciously turned his mind away from all mechanistic speculations. We give the substance of his own words: "The interior of all bodies, including the free ether, can, from an initial state of rest, experience some disturbances which we call electrical and others which we call magnetic. We do not know the nature of these changes of state, but only the phenomena which their presence calls up." This definite renunciation of a mechanical explanation is of great importance from the point of view of method. It opens up the avenue for the great advances which have been made by Einstein's researches. The mechanical properties of solid and fluid bodies are known to us from experience, but this experience concerns only their behaviour in a crude sense. It may be, and this has been supported by more recent molecular researches, that these visible, crude properties are a sort of appearance, an illusion, due to our clumsy methods of observation, whereas the actual occurrences between the smallest elements of structure, the atoms, molecules, and electrons, take place according to quite different laws. It is, therefore, a naïve prejudice that every continuous medium, like the ether, must behave like the apparently continuous fluids and solids of this world, which is accessible to us through our coarse senses. The properties of the ether must rather be ascertained by studying the events that occur in it independently of all other experiences. The result of these researches may be expressed as follows. The state of the ether may be described by two directed magnitudes, which bear the names electric and magnetic strength of field, E and H, and the changes of which in space and time are connected by Maxwell's equations. Under certain circumstances, mechanical, thermal, and chemical actions on matter, which are capable of being observed, are conditioned by the state of the ether.

Everything that goes beyond these assertions is superfluous hypothesis and mere fancy. It may be objected that such an abstract view undermines the inventive power of the investigator, which is stimulated by visual pictures and analogies, but Hertz' own example contradicts this opinion, for rarely has a physicist been possessed of such wonderful ingenuity in experiment, although as a theorist he recognized only pure abstraction as valid.

11. Hertz' Theory of Moving Bodies

A more important question than the pseudo-problem of the mechanical interpretation of ether events is that concerning the influence of the motions of bodies, among which must be counted, besides matter, the ether, on electromagnetic phenomena. This brings us back from a more general standpoint to the investigations which we made earlier (IV, 7, p. 102) into the optics of moving bodies. Optics is now a part of electrodynamics, and the luminiferous ether is identical with the electromagnetic ether. All the inferences that we made earlier from the optical observations with regard to the behaviour of the luminiferous ether must retain their validity, since they are obviously quite independent of the mechanism of light vibrations; for our investigation concerned only the geometrical characteristics of a light wave, namely, frequency (Doppler effect), velocity (convection), and direction of propagation (aberration).

We have seen that up to the time when the electromagnetic theory of light was developed only quantities of the first order in $\beta = \frac{v}{c}$ were accessible to measurement. And the result of these observations could briefly be expressed thus as the "optical principle of relativity": optical events depend only on the relative motions of the involved material bodies that emit, transmit, or receive the light. In a system of reference moving with translation all inner optical events occur just as if the ether were at rest.

Two theories were proposed to account for this fact. That of Stokes assumed that the ether inside matter was completely carried along by the latter; the second, that of Fresnel, on the other hand, was satisfied by supposing only a partial convection, the amount of which could be derived from experiments. We have seen that Stokes' theory, when carried to its logical conclusion, became involved in difficulties, but that Fresnel's represented all the phenomena satisfactorily.

In the electromagnetic theory exactly the same two standpoints are possible, either complete convection, as advocated by Stokes, or the partial convection of Fresnel. The question is whether purely electromagnetic observations will allow us to come to a decision about these two hypotheses.

Hertz was the first to apply the hypothesis of complete convection to Maxwell's field equations. In doing so, he was fully conscious that such a procedure could be only provisional, because the application to optical events would lead to the same difficulties as those which brought Stokes' theory to grief. But the simplicity of a theory which required no distinction to be made between the motion of ether and of matter led him to develop it extensively and to discuss it. This brought to light that the induction phenomena in moving *conductors*, which are by far the most important for experimental physics and technical science, are correctly represented by Hertz' theory. Disagreements with experimental results occur only later in finer experiments in which the displacements in *non-conductors* play a part. We shall investigate all possibilities in succession.

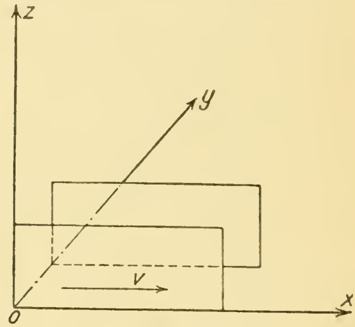


FIG. 98.

1. Moving conductors—(*a*) in the electrical field ;
(*b*) in the magnetic field.
2. Moving insulators— (*a*) in the electrical field ;
(*b*) in the magnetic field.

(1*a*) A conductor acquires surface charges in an electric field. If it is moved, it carries them along with itself. But moving charges must be equivalent to a current, and hence must produce an enveloping magnetic field according to Biot and Savart's law. To picture this to ourselves we imagine a plate condenser whose plates are parallel to the *xz*-plane (Fig. 98). Let them be oppositely charged and let them have an amount of electricity e on each unit area of the plate. Now, let one plate be moved with respect to the ether in the direction of the x-axis, with the velocity v. Then a *convection current* arises. The moving plate is displaced in the unit of time by a length v. If its cross section perpendicular to the x-axis is q, then an amount of electricity eqv passes through a plane that is parallel to the yz-plane, hence a current of the density ev flows. This must exert exactly the same magnetic action as a conduction current of density $i = ev$, flowing through the plate when it is at rest.

This was confirmed experimentally in Helmholtz' laboratory by H. A. Rowland (1875), and later, more accurately, by A. Eichenwald. Instead of the plate moving rectilinearly, a rotating metal disc was used.

(1b) When conductors are moved about in a magnetic field, electric fields arise in them, and hence currents are produced. This is the phenomenon of *induction by motion*, already discovered by Faraday and investigated quantitatively

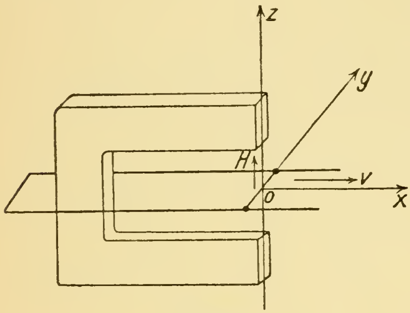


FIG. 99.

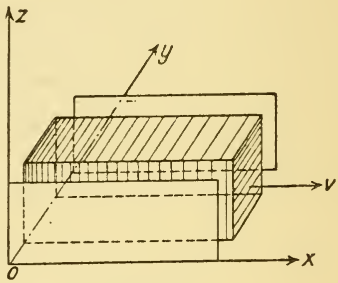


FIG. 100.

by him. The simplest case is this. Let the magnetic field H produced, say, by a horse-shoe magnet, be parallel to the z-axis (Fig. 99). Let there be a straight piece of wire of unit length parallel to the y-axis, and let this be moved with the velocity v in the direction of the x-axis. Hertz' theory then tells us that an electric field parallel to the negative direction of the z-axis is induced in this wire. If the wire is now made part of a closed circuit by sliding it on the two opposite arms of a U-shaped piece of wire, but so that the U takes no part in the motion, as is indicated in the Fig., then an induction current flows in the wire. This is given most simply by stating Faraday's law of induction thus: the current induced in a wire which forms part of a closed circuit is proportional to the change of the number of lines of force per second or of the magnetic displacement μH which is enclosed by the wire. Through the motion of the above wire this number clearly increases by μHv per second. Hence the induced electric intensity of field is equal to $\mu H \frac{v}{c}$.

This law is the basis of all machines and apparatus of physics and electro-technical science in which energy of motion is transformed by induction into electromagnetic energy; these include, for example, the telephone and dynamo machines of every kind. Hence the law may be regarded as having been confirmed by countless experiments.

(2*a*) We suppose the motion of a non-conductor in an electric field to be realized thus: let a movable disc composed of the substance of the non-conductor be placed between the two plates of the condenser of Fig. 98 (see Fig. 100). If the condenser is now charged, an electric field E arises in the disc, and a displacement ϵE is induced which is perpendicular to the plane of the plates, that is, parallel to the y-direction. This causes the two boundary faces of the insulating disc to be charged equally and oppositely to the metal plates facing them, respectively. Concerning the amount of this charge we know the following. On page 150 we saw that, according to Maxwell's view, Coulomb's law gives a relation between the amount of the displacement around a charged sphere, and its charge e—namely, for a sphere of radius r we have

$$E = \frac{e}{\epsilon r^2} \text{ or } \epsilon E = \frac{e}{r^2}.$$

But this sphere has the surface $4\pi r^2$, hence the charge per unit of surface is

$$\frac{e}{4\pi r^2} = \frac{\epsilon E}{4\pi}.$$

If we apply this to the case of the condenser, the surface density on the bounding planes of the isolating plate will be just as great as that on the metal plates, and it will be connected with the electrical field by

$$e = \frac{\epsilon E}{4\pi}.$$

If the insulating layer is now moved in the direction of the x-axis with the velocity v, then, according to Hertz, the ether in the layer will be carried along completely. Hence, also, the field E and the charges $e = \frac{\epsilon E}{4\pi}$ produced by it on the bounding planes will be carried along.

Therefore the moving charge of a bounding surface again represents a current of density $\frac{\epsilon E}{4\pi}v$, and must generate, according to Biot and Savart's law, a magnetic field.

W. C. Röntgen proved experimentally (1885) that this was the case, but the deflection of the magnet-needle that he observed was much smaller than it should have been from Hertz' theory. According to his measurements, it is as if not the whole ether is carried along by the disc, but only a part. The other part remains at rest. If the disc were to consist of pure ether, then we should have $\epsilon = 1$, and the charge produced would equal $\frac{E}{4\pi}$. Röntgen's experiments, however, show that only the excess of the charge over this amount, that is $\frac{\epsilon E}{4\pi} - \frac{E}{4\pi} = \frac{E}{4\pi}(\epsilon - 1)$ participates in the motion of the matter. We shall interpret this result simply later. Here we merely establish that, as was to be expected according to the well-known facts of optics, Hertz' theory of complete convection also fails to explain purely electromagnetic phenomena.

Eichenwald (in 1903) confirmed Röntgen's result very strikingly by allowing the charged metal plates to take part in the motion. These give a convection current of the amount ev; according to Hertz this insulating layer ought, on account of the opposite and equal charges, exactly to compensate this current. But Eichenwald found that this was not the case. Rather, he obtained a current which was entirely independent of the material of the insulator. This is exactly what is to be expected according to Röntgen's results of partial convection. For the current due to the insulator is $\left(\frac{\epsilon E}{4\pi} - \frac{E}{4\pi}\right)v$, of which the first member is compensated by the convection current ev, and so we are left with the current $\frac{E}{4\pi}v$, which is independent of the dielectric constant ϵ.

(2*b*) We imagine a magnetic field parallel to the z-axis, produced, say, by a horse-shoe magnet, and a disc of non-conducting material moving through the field in the direction of the x-axis (Fig. 101). As there are no non-conductors that are appreciably magnetizable, we shall assume $\mu = 1$. Let the two bounding faces of the disc, which are perpendicular to the y-axis, be covered with metal, and let these surface layers be connected to an electrometer by means of sliding contacts, so that the charges that arise on them can be measured.

This experiment corresponds exactly with the induction experiment discussed under (1*b*), except that a moving dielectric now takes the place of the moving conductor. The law of induction is applicable in the same way. It demands the existence of an electrical field $E = vH$, acting in the magnetic direction of the y-axis, if the thickness of the disc is unity. Hence, according to Hertz' theory, the two superficial layers must exhibit opposite charges of the surface density $\frac{\epsilon E}{4\pi} = \frac{\epsilon v H}{4\pi}$, which cause a deflection of the electrometer. The experiment was carried out by H. F. Wilson, in 1905, with a rotating dielectric, and it did, indeed, confirm the existence of the charge produced, but again to a lesser extent, namely, corresponding to a surface density $(\epsilon - 1)\frac{vH}{4\pi}$. It again seemed as if not the whole ether took part in the motion of matter, but only a part which is greater in proportion as the latter is more

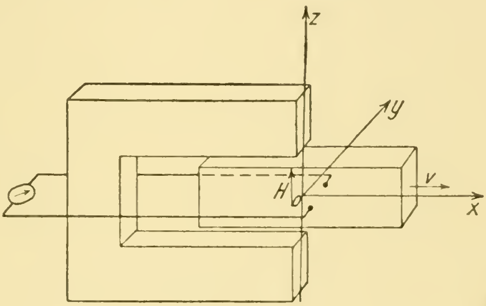


FIG. 101.

strongly dielectric than a vacuum. Here too, then, Hertz' theory fails.

In all these four typical phenomena what counts is clearly only the relative motion of the field-producing bodies with respect to the conductor or insulator investigated. Instead of moving this in the x-direction, as we have done, we could have kept it at rest and moved the remaining parts of the apparatus in the negative direction of the x-axis. The result would have to be the same. For Hertz' theory recognizes only relative motions of bodies, the ether being also reckoned as a body. In a system moving with translation all events happen, according to Hertz, as if it were at rest; that is, the classical principle of relativity holds.

But Hertz' theory is incompatible with the facts, and soon had to make way for another, which took exactly the opposite point of view with regard to relativity.

12. The Electron Theory of Lorentz

It is the theory of H. A. Lorentz (proposed in 1892) that signified the climax and the final step of the physics of the material ether.

It is a one-fluid theory of electricity that has been still further developed atomistically, and this, too, is, as we shall see presently, what determines the part which it allocates to the ether.

The fact that electric charges have an atomic structure, that is, occur in very small indivisible quantities, was first stated by Helmholtz (in 1881) in order to make intelligible Faraday's laws of electrolysis (p. 135). Actually, it was only necessary to assume that every atom in an electrolytic solution enters into a sort of chemical bond with an atom of electricity or an *electron*, to understand that a definite amount of electricity always separates out equivalent amounts of substances.

The atomic structure of electricity proved of particular value for explaining the phenomena which are observed in the passage of the electric current through a rarified gas.

Here it was first discovered that positive and negative electricity behave quite differently. If two metal electrodes are introduced into a glass tube and if a current is made to pass between them (Fig. 102), very complicated phenomena are produced so long as gas is still present at an appreciable pressure in the tube. But if the gas is pumped out more and more, the phenomena become simpler and simpler. When the vacuum is very high the negative electrode, the cathode K, emits a ray of bluish light rectilinearly, taking no account of where the positive pole, the anode A, is situated. These *cathode rays*, which were discovered by Plücker (1858) were regarded by some physicists as light rays, for, as Hittorf (1869) showed, they threw shadows of solid bodies that were interposed in their path. Others regarded them as a material emanation that was shot out by the cathode. Crookes, who upheld this view (1879), called these rays the "fourth state" of matter. A fact that spoke in favour of the material nature of the rays was, above all, that they were deflected by a magnet, and, indeed, just like a stream of negative electricity. The greatest share in investigating the nature of cathode rays was taken by Sir J. J. Thomson and P. L. Lenard. It was successfully shown that the negative charge of the rays could be caught up directly. Furthermore, they are deflected by an electric field applied perpendicularly to their path, and this deflection is opposite to the direction of the field, which again proves the charge to be negative.

The conviction that the nature of cathode rays is corpuscular grew to certainty when physicists succeeded in deducing quantitative conclusions concerning their velocity and their charge.

If we picture the cathode ray as a stream of small particles of mass m, then clearly it will be the less deflected by a definite electrical or magnetic field the greater its velocity; just as a rifle-bullet travels with more "rush" the greater its velocity. Now it is possible to produce cathode rays that can be strongly deflected, that is, slow cathode rays. These may be accelerated artificially so strongly that their initial velocity may be neglected in comparison with their final velocity. To achieve this a wire net or grid N is placed before the cathode K (Fig. 103) and is strongly charged positively. The negative cathode ray particles are then strongly accelerated in the field between the cathode and the grid, and they fly through the meshes of the grid with a velocity that is essentially due only to this accelera-

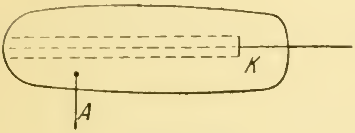


FIG. 102.

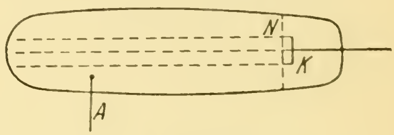


FIG. 103.

tion. But this may be calculated from the fundamental equation of mechanics

$$mb = \mathrm{K} = e\mathrm{E}$$

where e is the charge, and E is the field strength. We are here clearly dealing with a case analogous to that of "falling" bodies, in which the acceleration is not equal to that of gravity g but to $\frac{e}{m}\mathrm{E}$. If the ratio $\frac{e}{m}$ were known, the velocity v could be found from the laws for falling bodies. But there are two unknowns, $\frac{e}{m}$ and v, and hence another measurement is necessary if they are to be determined. This is obtained by applying a lateral magnetic force. In discussing Hertz' theory (V, 11, 1*b*, p. 164) we saw that a magnetic field H calls up in a body moving perpendicularly to H an electric intensity of field $\mathrm{E} = \frac{v}{c}\mathrm{H}$, which is perpendicular both to H and to v. Hence a deflecting force $e\mathrm{E} = e\frac{v}{c}\mathrm{H}$ will act on every cathode ray particle, so that there will be an acceleration $b = \frac{e}{m}\frac{v}{c}\text{H}$ perpendicular to the original motion. This may be found by measuring the lateral deflection of the ray. Hence we have a second equation for determining the two unknowns, $\frac{e}{m}$ and v.

The determinations carried out by this or a similar method have led to the result that, for velocities that are not too great, $\frac{e}{m}$ has a definite constant value :

$$\frac{e}{m} = 5\cdot31 \, . \, 10^{17} \text{ electrostatic units per grm.} \qquad (65)$$

On the other hand, in dealing with electrolysis (V, 3, formula (48), p. 136), we stated that 1 grm. of hydrogen carries an amount of electricity $C_0 = 2\cdot90 \, . \, 10^{14}$. If we now make the readily suggested assumption that the charge of a particle is in each case the same, namely, an atom of electricity or an *electron*, we must conclude that the mass of the cathode ray particle m must bear the following ratio to that of the hydrogen atom m_{H} :

$$\frac{m}{m_{\text{H}}} = \frac{e}{m_{\text{H}}} : \frac{e}{m} = \frac{2\cdot90 \cdot 10^{14}}{5\cdot31 \cdot 10^{17}} = \frac{1}{1830} \qquad (66)$$

Thus, the cathode ray particles are nearly 2000 times lighter than hydrogen atoms, which are the lightest of all chemical atoms. This result leads us to conclude that the cathode rays are a current of pure atoms of electricity.

This view has excellently stood the test of innumerable researches. Negative electricity consists of freely-moving electrons, but positive electricity is bound to matter, and never occurs without it. Thus, recent experimental researches have confirmed and given a precise form to the old hypothesis of the one-fluid theory. The amount of the charge e of the individual electron has also been successfully determined. The first experiments of this type were carried out by Sir J. J. Thomson (1898). The underlying idea is : little drops of oil or water, or little spheres of metal of microscopic or sub-microscopic dimensions, which are produced by condensation of vapour or electrical "spotting" in air, fall with constant velocity, since the friction of the air prevents acceleration. By measuring the rate of fall, the size of the particles can be determined, and then their mass M is obtained by multiplying their size by the density. The weight of such a particle is then Mg, where $g = 981$ cm/sec^2 is the acceleration due to gravity. Now, such particles may be charged electrically by subjecting the air to the action of Röntgen rays or the rays of radioactive substances. If an electric field E, which is directed vertically upwards, is then applied, a sphere carrying the charge e is pulled upwards by it, and if the electric force eE is equal to the weight Mg, the sphere will remain poised in the air. The charge e may then be calculated from the equation $eE = Mg$. Millikan (1910), who carried out the most accurate experiments of this sort, found that the charge of the small drops is always an exact multiple of a definite minimum charge. Thus we shall call this the *elementary electrical quantum*. Its value is

$$e = 4{\cdot}77 \,.\, 10^{-10} \text{ electrostatic units} \quad . \quad . \quad (67)$$

It must be mentioned that the results of Millikan's experiments are disputed by Ehrenhaft, but it is probable that the values below e obtained by the latter for the elementary charge are due to his having used spheres which were too small, and which gave rise to secondary phenomena.

The absolute value of the elementary charge plays no essential part in Lorentz' theory of electrons. We shall now depict the physical world sketched by Lorentz.

The material atoms are the carriers of positive electricity, which is indissolubly connected with them. In addition they also contain a number of negative electrons, so that they appear to be electrically neutral with respect to their surroundings. In non-conductors the electrons are tightly bound to the atoms; they may only be displaced slightly out of the positions of equilibrium, by which the atom becomes a dipole. In electrolytes and conducting gases it may occur that an atom has one or more electrons too many or too few; it is then called an *ion* or a *carrier*, and it wanders in the electric field carrying electricity and matter simultaneously. In metals the electrons move about freely, and experience resistance only when they collide with the atoms of the substance. Magnetism comes about through the electrons in certain atoms moving in closed orbits and hence representing Ampère molecular currents.

The electrons and the positive atomic charges swim about in the sea of ether, in which an electromagnetic field exists in accordance with Maxwell's equations. But we must set $\epsilon = 1$, $\mu = 1$ in them, and, in place of the density of the conduction current, we have the convection current ρv of the electrons. The equations thus become

$$\left.\begin{array}{ll} \operatorname{div} E = \rho & \operatorname{curl} H - \dfrac{1}{c}\dfrac{\mathbf{e}}{t} = \dfrac{\rho v}{c} \\ \operatorname{div} H = 0 & \operatorname{curl} E + \dfrac{1}{c}\dfrac{\mathbf{h}}{t} = 0 \end{array}\right\} \quad . \quad . \quad (68)$$

and include the laws of Coulomb, Biot, and Savart, and Faraday in the usual way.

Thus *all* electromagnetic events consist fundamentally of the motions of electrons and of the fields accompanying them. All matter is an electrical phenomenon. The various properties of matter depend on the various possibilities of motion of the electrons with respect to atoms, as in the manner just now described. The problem of the theory of electrons is to derive the ordinary equations of Maxwell from the fundamental laws (68) for the individual invisible electrons and atoms, that is, to show that material bodies appear to have, according to their nature, respectively, a conductivity σ, a dielectric constant, ϵ and a permeability μ.

Lorentz has solved this problem and has shown that the theory of electrons not only gives Maxwell's laws in the simplest case, but, more than this, also makes it possible to explain numerous finer facts which were insoluble for the descriptive theory or could be accounted for only with the aid of artificial hypotheses. These facts comprise, above all, the finer phenomena of optics, colour dispersion, the magnetic rotation of the plane of polarization (p. 155) discovered by Faraday, and similar interactions between light waves and electric or magnetic fields. We cannot enter further into this extensive and mathematically complicated theory, and shall restrict ourselves to the question which is of primary interest to us: what part does the ether take in the motions of matter?

Lorentz proclaimed the very radical thesis, which had never before been asserted with such definiteness:

The ether is at rest in absolute space.

In principle this identifies the ether with absolute space. Absolute space is no vacuum, but a something with definite properties, whose state is described when two directed quantities are given, the electrical field E and the magnetic field H, and, as such, it is called the ether.

This assumption goes still a little further than the theory of Fresnel. In the latter the ether of astronomic space was at rest in an inertial system, which we might also call absolute rest. But the ether inside material bodies is partly carried along by them.

Lorentz can dispense even with this partial convection and yet arrives at practically the same result. To see this, we consider the phenomenon that occurs in a dielectric between the plates of a condenser. When the latter is charged, a field perpendicular to the plate arises (Fig. 104), and this displaces the electrons in the atoms of the dielectric substance and transforms them into dipoles, as we explained earlier (pp. 146 and 171).

The dielectric displacement in Maxwell's sense is ϵE, but only a part of it is due to the actual displacement of the electrons. For a vacuum has the dielectric constant $\epsilon = 1$, and hence the displacement E; consequently the true value of the electronic displacement is $\epsilon E - E\ (\epsilon - 1) E$. Now we have seen that the experiments of Röntgen and Wilson on the phenomena in moving insulators affirm that actually only this part of the displacement takes part in the motion. Thus Lorentz' theory gives a correct account of electromagnetic facts without needing to have recourse in any way to the motion of matter.

The fact that the convection of light comes out in exact agreement with Fresnel's formula (43), (p. 117) is made plausible by the following argument.

As in Wilson's experiment, we consider a dielectric body which moves in the x-direction with the velocity v and in which a light ray travels in the same direction (Fig. 105). Let this

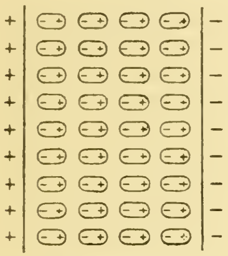


FIG. 104.

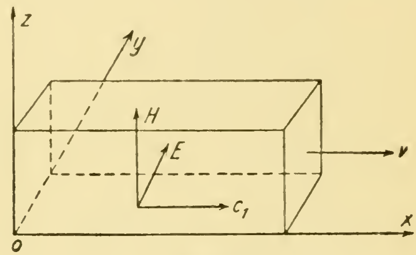


FIG. 105.

ray consist of an electrical vibration E parallel to the y-axis and a magnetic vibration parallel to the z-axis. Now, we know from Wilson's experiment that such a magnetic field in the moving body produces a corresponding displacement of the value $(\epsilon - 1)v$H in the y-direction. From this we get a superposed electrical displacement if we divide by ϵ. Thus the total electrical field is

$$E + \frac{\epsilon - 1}{\epsilon} v H.$$

If the convection were not partial but complete, as is assumed in Hertz' theory, we should have only ϵ in place of $\epsilon - 1$, thus the total field would have the value $E + vH$. We see that on account of the partial convection v is to be replaced by

$$\frac{\epsilon - 1}{\epsilon} v.$$

Thus this value should correspond to the absolute velocity of the ether within matter according to Fresnel's theory, that is, to the convection coefficient called ϕ in optics, cf. formula (43). And this is precisely the case actually. For, according to Maxwell's electromagnetic theory (p. 158) the dielectric constant ϵ is equal to the square of the index of refraction n, i.e. $\epsilon = n^2$. If we insert this value, we get

$$\frac{\epsilon - 1}{\epsilon} v = \frac{n^2 - 1}{n^2} v = \left(1 - \frac{1}{n^2}\right) v = \phi,$$

in agreement with formula (43) (p. 117).

We recall to ourselves that Fresnel's theory encountered difficulties through colour dispersion. For, if the refractive index n depends on the frequency (colour) of the light, so also will the convection coefficient ϕ. But the ether can be carried along in only *one* definite way, and not differently for each colour. This difficulty vanishes entirely for the theory of electrons, for the ether remains at rest, and it is the electrons situated in the matter that are carried along; and colour dispersion is due to their being forced into vibration by the light and reacting, in turn, on the velocity of the light.

We cannot enter further into the details of this theory and its many ramifications, but we shall recapitulate the result as follows.

Lorentz' theory presupposes the existence of an ether that is absolutely at rest. It then proves that, in spite of this, all electromagnetic and optical phenomena depend only on the relative motions of translation of material bodies, so far as terms of the first order in β come into account. Hence it accounts for all known phenomena, above all for the fact that the absolute motion of the earth through the ether cannot be shown by experiments on the earth involving only quantities of the first order (this is the optical or, rather, the electromagnetic principle of relativity).

But an experiment of the first order may be imagined which would be just as little capable of being explained by Lorentz' theory as by all the theories previously discussed: this would be a failure of Römer's method for determining an absolute motion of the whole solar system (see pp. 80 and 111).

The deciding point for Lorentz' theory is whether it stands the test of experiments that allow quantities of the second order in β to be measured. For they should make it possible to establish the absolute motion of the earth through the ether. Before we enter into this question, we have yet to discuss an achievement of Lorentz' theory of electrons through which its range became greatly extended, namely, the electro-dynamic interpretation of inertia.

13. Electromagnetic Mass

The reader will have remarked that from the moment when we left the elastic ether and turned our attention to the electrodynamic ether, we had little to say about mechanics. Mechanical and electrodynamic phenomena each form a realm for themselves. The former take place in absolute Newtonian space, which is defined by the law of inertia and which betrays its existence through centrifugal forces; the latter are states of the ether which is at rest in absolute space. A comprehensive theory, such as that of Lorentz aims at being, cannot allow these two realms to exist unassociated side by side.

Now we have seen that physicists have not succeeded satisfactorily in reducing electrodynamics to terms of mechanics in spite of incredible effort and ingenuity.

The bold idea of the converse then suggests itself: cannot mechanics be reduced to terms of electrodynamics?

If this could be carried out successfully the absolute abstract space of Newton would be transformed into the concrete ether. The inertial resistances and centrifugal forces would have to appear as physical actions of the ether, say, as electromagnetic fields of particular form, but the principle of relativity of mechanics would have to lose its strict validity and would be true, like that of electrodynamics, only approximately, for quantities of the first order in $\beta = \frac{v}{c}$.

Science has not hesitated to take this step, which entirely inverts the order of rank of conceptions. And although the doctrine of the ether which is absolutely at rest had later to be dropped, this revolution, which ejected mechanics from its throne and raised electrodynamics to sovereign power in physics, was not in vain. Its result has retained validity in a somewhat altered form.

We saw above (p. 157) that the propagation of electromagnetic waves comes about through the mutual action of electrical and magnetic intensity of field, producing an effect which is analogous to that of mechanical inertia. An electromagnetic field has an inertial power of persistence quite similar to that of matter. To generate it, work must be performed, and when it is destroyed this work again appears. This is observed in all phenomena that are connected with electromagnetic vibrations, for example, in the various forms of apparatus of wireless telegraphy. A wireless sending station contains an electric oscillator, which consists essentially (Fig. 106) of a spark gap F, a coil S, and a condenser K, that is, two metal plates that are separated from each other, and these instruments, connected by wires, form an "open" circuit. The condenser is charged until a spark jumps across the gap at F. This causes the condenser to become discharged, and the quantities of electricity that have been stored flow away. They do not simply neutralize each other, but shoot beyond the state of equilibrium and again become collected on the condenser plates, only with reversed signs, just like a pendulum swings past the position of equilibrium to the opposite side. When the condenser has thus been charged up afresh, the electricity again flows back and swings to and fro until its energy has been used up in warming the conducting wires or in being passed on to other parts of the apparatus, for example, the emitting antenna. Thus the swinging of the electricity to and fro proves

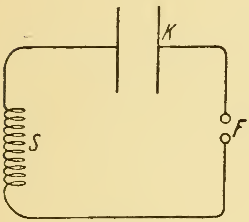


FIG. 106.

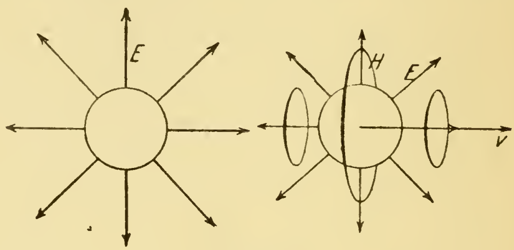


FIG. 107.

the inertial property of the field, which exactly corresponds to the inertia of mass of the pendulum bob. Maxwell's theory represents this fact correctly in all details. The electromagnetic vibrations that occur in a definite set of apparatus can be predicted by calculations from the equations of the field.

This led J. J. Thomson to infer that the inertia of a body must be increased by an electric charge which is imparted to it. Let us consider a charged sphere at first at rest and then moving with the velocity v. The stationary sphere has an electrostatic field with lines of forces which run radially outwards, the moving sphere has, in addition, a magnetic field with circular lines and forces that encircle the path of the sphere (Fig. 107). For a moving charge is a convection current and produces a magnetic field in accordance with Biot and Savart's law. Both states have the inertial property above described. The one can be transformed into the other only by the performance of work. The force that is necessary to set the stationary sphere into motion is thus greater for the charged than for the uncharged sphere. To accelerate the moving charged sphere still further, the magnetic field H must clearly be strengthened. Thus, again, an increase of force is necessary.

We remember that a force K that acts for a short time t represents an impulse $J = Kt$, which produces a change of velocity w in a mass m in accordance with the formula (7) (II, 9, p. 31):

$$mw = J.$$

If the mass carries a charge, a definite impulse J will produce a smaller change of velocity, and the remainder J′ will be used to charge the magnetic field. Thus we have

$$mw = J - J'.$$

Now, calculation gives the rather natural result that the impulse J′ necessary to increase the magnetic field is the greater, the greater the change of velocity w, and, indeed, it is approximately proportional to this change of velocity. Thus we may set $J' = m'w$, where m' is a factor of proportionality, which, moreover, may depend on the state, that is, the velocity v, before the change of velocity occurs. We then have

$$mw = J - m'w$$

or

$$(m + m')w = J.$$

Thus, it is as if the mass m is augmented by an amount m', which is to be calculated from the electromagnetic field equations, and which may be dependent on the velocity v. The exact value of m' for any velocities v may be calculated only if assumptions are made about the distribution of the electric charge over the moving body. But the limiting value for velocities that are small compared with that of light c, that is, for small values of β, is obtained independently of such assumptions as

$$m_0 = \frac{4}{3}\frac{U}{c^2}, \quad . \quad . \quad . \quad . \quad (69)$$

where U is the electrostatic energy of the charges of the body.

We have seen that the mass of the electron is about 2000 times smaller than that of the hydrogen atom. Hence the idea suggests itself that the electron has, perhaps, no "ordinary" mass at all, but is nothing other than an "electric" charge in itself, and that its mass is electromagnetic in origin.

Is such an assumption reconcilable with the knowledge that we have of the size, charge, and mass of the electron?

Since the electrons are to be the structural elements of atoms, they must at any rate be small compared with the size of atoms. Now we know from atomic physics that the radius of atoms is of the order 10^{-8} cms. Thus the radius of the electron must be essentially smaller than 10^{-8} cms. If we imagine the electron as a sphere of radius a with the charge e distributed over its surface, then, as may be derived from Coulomb's law, the electrostatic energy is $U = \frac{1}{2}\frac{e^2}{a}$. Hence, by (69), the electromagnetic mass becomes

$$m_0 = \frac{4}{3}\frac{U}{c^2} = \frac{2}{3}\frac{e^2}{ac^2}.$$

From this we can calculate the radius a:

$$a = \frac{2}{3} \cdot \frac{e}{c^2} \cdot \frac{e}{m_0}.$$

On the right-hand side we know all the quantities; $\frac{e}{m_0}$ from the deflection of the cathode rays [formula (65), p. 170], e from Millikan's measurements [formula (67), p. 171], and c is the velocity of light. If we insert the values given, we get

$$a = \frac{2}{3} \cdot \frac{4{\cdot}77 \cdot 10^{-10}}{9 \cdot 10^{20}} \cdot 5{\cdot}31 \cdot 10^{17} = 1{\cdot}88 \cdot 10^{-13} \text{ cms.},$$

a length which is about 100,000 times smaller than the radius of an atom.

Thus the hypothesis that the mass of the electron is electromagnetic in origin does not conflict with the known facts. But this does not prove the hypothesis.

At this stage theory found strong support in refined observations of cathode rays, and of the β-rays of radio-active substances, which are also ejected electrons. We explained above how electric and magnetic action on these rays allow us to determine the ratio of the charge to the mass, $\frac{e}{m}$, and also their velocity v, and that at first a definite value for $\frac{e}{m}$ was obtained, which was independent of v. But, on proceeding to higher velocities, a decrease of $\frac{e}{m}$ was found. This effect was particularly clear, and could be measured quantitatively, in the case of the β-rays of radium, which are only slightly slower than light. It was incompatible with the ideas of the electron theory that an electric charge should depend on the velocity. But that the mass should depend on the velocity was certainly to be expected if the mass was to be electromagnetic in origin. To arrive at a quantitative theory, it is true, definite assumptions had to be made about the form of the electron and the distribution of the charge on it. M. Abraham (1903) regarded the electron as a rigid sphere, with a charge distributed uniformly over the interior or over the surface, and he showed that both assumptions lead to the same dependence of the electromagnetic mass on the velocity, namely, to an increase of mass with increasing velocity. The more quickly the electron travels, the more the electromagnetic field resists a further increase of velocity. The increase of m explains the observed decrease of $\frac{e}{m}$, and Abraham's theory agrees very well quantitatively with the results of measurement of Kaufmann (1901), if it is assumed that there is no "ordinary" mass present with the electromagnetic mass.

Thus, the object of tracing the inertia of electrons back to electromagnetic fields in the ether was attained. At the same time a further perspective presented itself. Since atoms are the carriers of positive electricity, and also contain numerous electrons, perhaps their mass is also electromagnetic in origin? In that case, mass as the measure of the inertial persistence would no longer be a primary phenomenon, as it is in elementary mechanics, but a secondary consequence of the structure of the ether. Newton's absolute space, which is defined only by the mechanical law of inertia, thus becomes superfluous; its part is taken over by the ether, whose electromagnetic properties are so well known. In this way a very concrete solution of the problem of space would be obtained, and it would be one that would be in conformity with physical thought.

We shall see (V, 15, p. 184) that new facts contradict this view. But the relationship between mass and electromagnetic energy, which was first discovered in this way, denotes a fundamental discovery, the deep significance of which was brought into due prominence only when Einstein proposed his theory of relativity.

We have yet to add that, besides Abraham's theory of the rigid electron, other hypotheses were set up and worked out mathematically. The most important is that of H. A. Lorentz (1904), which is closely connected with the theory of relativity. He assumed that during its motion the electron contracts in the direction of motion; so that from a sphere it becomes a flattened spheroid of revolution, the amount of the flattening depending in a definite way on the velocity. This hypothesis seems at first sight strange. It certainly gives an essentially simpler formula for the way the electromagnetic mass depends on the velocity than Abraham's theory, but this does not justify it. The true criterion is given in the development which Lorentz' theory of electrons had to follow beyond quantities of the second order in consequence of experimental researches, to which we shall presently direct our attention. Lorentz' formula then gained a universal significance in the theory of relativity. We shall return to the experimental decision between it and Abraham's theory later (VI, 7, p. 221).

After the theory of electrons had reached the stage above described, towards the close of the new century, the possibility of forming a uniform physical picture of the world seemed to have drawn near, one which would reduce all forms of energy, including mechanical inertia, to the same root, to the electromagnetic field in the ether. Only one form of energy still remained outside the system, namely, gravitation; yet it seemed reasonable to hope that that, too, would allow itself to be interpreted as an action of the ether.

14. Michelson and Morley's Experiment

Even twenty years before, however, the base of the whole structure had suffered a fissure, and, whilst additions were being made above, supports and substitutes had to be applied below.

We have several times emphasized that the decisive experiments for the theory of the stationary ether had to be such as involved the measurement of quantities of the second order in β. This would necessarily bring to light whether the ether wind sweeps past a quickly moving body and disperses the light waves, as is demanded by theory.

Michelson and Morley (1881) were the first to carry out successfully the most important experiment of this type with Michelson's interferometer (IV, 4, p. 88), which he had refined to a precision instrument of unheard-of efficiency.

In investigating the influence of the earth's motion on the velocity of light (IV, 9, p. 113), it has been found that the time taken by a ray of light to pass to and fro along a distance l parallel to the earth's motion differs only by a quantity of the second order from the value which it has when the earth is at rest. We found earlier that this time was

$$t_1 = l\left(\frac{1}{c+v} + \frac{1}{c-v}\right) = \frac{2lc}{c^2 - v^2}$$

for which we may also write

$$t_1 = \frac{2l}{c} \cdot \frac{1}{1-\beta^2}.$$

If this time could be so accurately measured that the fraction $\frac{1}{1-\beta^2}$ could be distinguished from 1 in spite of the extremely small value of the quantity β^2, we should have means of proving the existence of an ether wind.

But it is by no means possible to measure "light" times (i.e. the short times taken by light rays to traverse certain distances). Interferometric methods give us rather only differences of the times taken by light to traverse various different routes, but with the amazing accuracy that is necessary for this purpose.

For this reason Michelson and Morley caused a second ray of light to traverse a path AB of the same length l to and fro, but perpendicularly to the earth's orbit (Fig. 108). Whilst the light passes from A to B, the earth moves a short distance forward so that the point B arrives at the point B′ of the ether. Thus the true path of the light in the ether is AB′, and if it takes a time t to cover this distance, then $AB' = ct$. During the same time A has moved on to the point A′ with the velocity v, thus $AA' = vt$. If we now apply Pythagoras' theorem to the right-angled triangle AA′B, we get

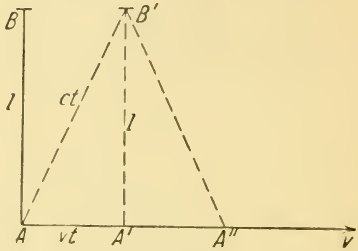


Fig. 108.

$$c^2t^2 = l^2 + v^2t^2$$

or

$$t^2(c^2 - v^2) = l^2, \qquad t^2 = \frac{l^2}{c^2 - v^2} = \frac{l^2}{c^2} \cdot \frac{1}{1-\beta^2},$$

$$t = \frac{l}{c} \cdot \frac{1}{\sqrt{1-\beta^2}}.$$

The light requires exactly the same time to do the return journey, for the earth shifts by the same amount, so that the initial point A moves from A′ to A″.

Thus the light takes the following time for the journey to and fro :

$$t_2 = \frac{2l}{c} \cdot \frac{1}{\sqrt{1-\beta^2}}.$$

The difference between the times taken to cover the same distance parallel and perpendicular to the earth's motion is thus

$$t_1 - t_2 = \frac{2l}{c}\left(\frac{1}{1-\beta^2} - \frac{1}{\sqrt{1-\beta^2}}\right).$$

Now, by neglecting members of higher order than the second in β (as we did on p. 108) we may approximate by replacing $\frac{1}{1-\beta^2}$ by $1+\beta^2$, and $\frac{1}{\sqrt{1-\beta^2}}$ by $1+\beta^2/2$.*

Hence we may write to a sufficient degree of approximation

$$t_1 - t_2 = \frac{2l}{c}\left[(1+\beta^2) - (1+\tfrac{1}{2}\beta^2)\right] = \frac{2l}{c}\cdot\frac{\beta^2}{2} = \frac{l}{c}\beta^2.$$

The retardation of the one light wave compared with the ether is thus a quantity of the second order.

This retardation may be measured with the help of Michelson's interferometer (Fig. 109). In this (cf. p. 113) the light coming from the source Q is divided at the half-silvered plate P into two rays which run in perpendicular directions to the mirrors S_1 and S_2, at which they are reflected and sent back to the plate P. From P onwards they run parallel into the telescope F, where they interfere. If the distances S_1P and S_2P are equal, and if the one arm of the apparatus is placed in the direction of the earth's motion, the case just discussed is exactly realized. Thus the two rays reach the field of vision with a difference of time $\frac{l}{c}\beta^2$. Hence the interference fringes are not exactly situated where they should be if the earth were at rest. But if we now

* For if x is a small number, whose square may be neglected, we have

$$(1+x)(1-x) = 1 - x^2 = 1 \text{ (approx.)},$$

hence

$$1 + x = \frac{1}{1-x}.$$

Furthermore,

$$\begin{aligned}(1-x)(1+\tfrac{1}{2}x)^2 &= (1-x)(1+x+x^2/4)\\ &= (1-x)(1+x) \text{ approx.}\\ &= 1 - x^2 \text{ approx.}\\ &= 1 \text{ approx.}\end{aligned}$$

Thus

$$(1+\tfrac{1}{2}x)^2 = \frac{1}{1-x}$$

$$1 + \tfrac{1}{2}x = \frac{1}{\sqrt{1-x}}.$$

If we replace x by β^2 in the two approximate formulæ obtained, we get the approximations used in the text:

$$\frac{1}{1-\beta^2} = 1 + \beta^2, \qquad \frac{1}{\sqrt{1-\beta^2}} = 1 + \tfrac{1}{2}\beta^2.$$

turn the apparatus through 90°, until the other arm is parallel to the direction of the earth's motion, the interference fringes will be displaced by the same amount, but in the opposite direction. Hence if we observe the position of the interference fringes whilst the apparatus is being rotated, a displacement should come into evidence which would correspond to the double retardation $2\frac{l}{c}\beta^2$.

If T is the period of vibration of the light used, the ratio of the retardation to the period is $\frac{2l}{cT}\beta^2$, and since by formula (35), (p. 85), the wave length $\lambda = cT$, we may write this ratio as $2\frac{l}{\lambda}\beta^2$.

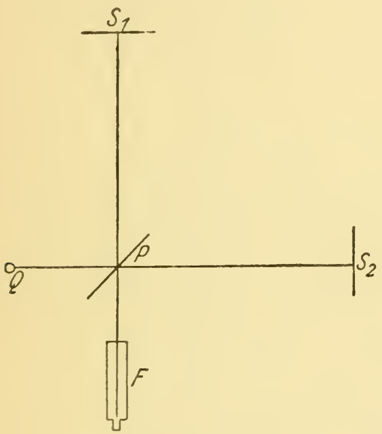


FIG. 109.

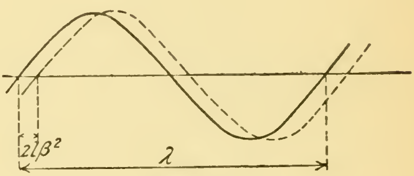


FIG. 110.

Hence, when the apparatus is rotated, the two interfering trains of waves experience a relative displacement whose ratio to the wave length is given by $\frac{2l\beta^2}{\lambda}$ (Fig. 110). The interference fringes themselves arise through the rays which leave the source in different directions, having to traverse somewhat different paths. The distance between two fringes corresponds to a path difference of one wave length, hence the observable displacement of the fringes is the fraction $\frac{2l\beta^2}{\lambda}$ of the width of the fringe.

Now Michelson, in a repetition of his experiment with Morley (1887) carried out on a larger scale, extended the length of the path traversed by the light, by means of several reflections to and fro, to 11 metres $= 1{\cdot}1 . 10^3$ cms. The wave length of the light used was about $\lambda = 5{\cdot}9 . 10^{-5}$ cms. We know that β is approximately equal to 10^{-4}, and hence $\beta^2 = 10^{-8}$. So we get

$$\frac{2l\beta^2}{\lambda} = \frac{2 . 1 . 1 . 10^3 . 10^{-8}}{5 . 9 . 10^{-5}} = 0{\cdot}37,$$

that is, the interference fringes must be displaced by more than $\frac{1}{3}$ of their distance apart when the apparatus is turned through 90°. Michelson was certain that the 100th part of this displacement would still be observable.

But when the experiment was carried out not the slightest sign of the expected displacement manifested itself, and later repetitions, with still more refined means, led to no other result. From this we must conclude that *the ether wind does not exist. The velocity of light is not influenced by the motion of the earth even to the extent involving quantities of the second order.*

15. The Contraction Hypothesis

Michelson and Morley concluded from their experiment that the ether is carried along completely by the moving earth, as is maintained in the elastic theory of Stokes, and in the electromagnetic theory of Hertz. But this contradicts the numerous experiments which prove partial convection. Michelson then investigated whether it was possible to establish a difference in the velocity of light at different heights above the earth's surface, but without a positive result. He concluded from this that the motion of the ether that is carried along by the earth must extend to very great heights above the earth's surface. Thus, then, the ether would be influenced by a moving body at considerable distances. But this is not the case actually, for Oliver Lodge showed (1892) that the velocity of light in the neighbourhood of rapidly-moving bodies is not influenced in the slightest, not even when the light passes through a strong electric or magnetic field, carried along by the body. And all these efforts seem almost superfluous, for even if they had led to an unobjectionable explanation of Michelson's experiment, the whole of the rest of electrodynamics and optics of moving bodies which speaks in favour of partial convection would remain unexplained.

One attempt to explain it, which was, however, developed much later by Ritz (1908), consists in the hypothesis that the velocity of light depends on the velocity of the *source of light.* Yet this assumption is in contradiction to almost all theoretical and experimental results of research. In the first place this would deprive electromagnetic events of their character of contiguous action, for such action consists in the propagation of an effect from one point to the next, being influenced only by the events in the immediate neighbourhood of this point, but not on the velocity of a far distant source of light. For this reason Ritz frankly called his theory a sort of emission theory. But what is emitted was, of course, not supposed to be material particles which obey mechanical laws, but an agent which, when it enters into matter, exerts directed transversal forces, and sets it into vibration. Light vibrations exist then only in matter and not in the ether. The objection that an emission theory is unable to account for interference is clearly unjustified in the case of this view.

But Ritz has not succeeded in bringing his theory into agreement with optical and electromagnetic phenomena. In all cases in which we have to do with relative motions of the source of light and the observer, effects on the vibration number (Doppler effect) do exhibit themselves and also effects on the direction (aberration), but not on the velocity of light (experiments of Arago, p. 114, and Hoek, p. 115). Recently, de Sitter (1913) has shown in an extensive investigation that the velocity of the light which comes from the fixed stars is independent of the motion of these stars.

We have mentioned this theory in spite of its failure, because one idea which it emphasizes is also important for an understanding of the theory of relativity, namely, the fact that all *observable* events are always bound to matter. The "field in the ether" is a fiction, devised to describe the dependence of events in bodies on space and time as simply as possible. We shall revert to this view later.

We now turn to Lorentz' theory of electrons, which was clearly placed in a very difficult position by Michelson and Morley's experiment. The doctrine of the stationary ether seems to demand implacably that an ether wind exists on the earth, and hence stands in flagrant contradiction to the results of Michelson and Morley's experiment. The fact that it did not at once succumb to this proves its inherent strength, which is due to its physical picture of the world being uniform and complete in itself.

Finally, it even overcame this difficulty to a certain extent, albeit by a very strange hypothesis, which was proposed by Fitzgerald (1892) and at once taken up and elaborated by Lorentz.

Let us recall the reflections on which Michelson and Morley's experiment were based. We found that the time taken by a light ray to travel to and fro along a distance l differs according to whether the ray travels parallel or perpendicularly to the earth's motion. In the former case $t_1 = \frac{2l}{c} \cdot \frac{1}{1 - \beta^2}$, in the second, $t_2 = \frac{2l}{c} \frac{1}{\sqrt{1 - \beta^2}}$.

If we now assume that the arm of the interferometer which is directed parallel to the direction of the earth's motion is shortened in the ratio $\sqrt{1 - \beta^2} : 1$, the time t_1 would become reduced in the same ratio, namely,

$$t_1 = \frac{2l\sqrt{1 - \beta^2}}{c(1 - \beta^2)} = \frac{2l}{c} \cdot \frac{1}{\sqrt{1 - \beta^2}}.$$

Thus we should have $t_1 = t_2$.

This hypothesis, the crudeness and boldness of which surprises us, is simply this: *every body which has the velocity v with respect to the ether contracts in the direction of motion by the fraction*

$$\sqrt{1 - \beta^2} = \sqrt{1 - v^2/c^2}.$$

Michelson and Morley's experiment must actually, then, give a negative result, as for both positions of the interferometer t_1 must equal t_2. Furthermore—and this is the important point—such a contraction could be ascertained by no means on the earth, for every earthly measuring rod would be contracted in just the same way. An observer who was at rest in the ether outside the earth would, it is true, observe the contraction. The whole earth would be flattened in the direction of motion and likewise all things on it.

The contraction hypothesis seems so remarkable, indeed almost absurd, because the contraction is not a consequence of any forces, but appears only as an accompanying circumstance of the fact of motion. But Lorentz did not allow this objection to deter him from absorbing this hypothesis in his theory, particularly as *new* experiments confirmed that in the second order of small quantities, too, no effect of the earth's motion through the ether could be detected.

We can neither describe all these experiments here, nor even discuss them in detail. They are partly optical and concern the events involved in reflection and refraction, double refraction, rotation of the plane of polarization, and so forth, and they are partly electromagnetic and concern induction phenomena, the distribution of the current in wires, and so forth. The improved technique of physics allows us nowadays to establish whether an influence of the second order on the earth's motion manifests itself or not in these phenomena. A particularly noteworthy experiment is that of Trouton and Noble (1903), which was intended to detect a torsional force which should occur in a suspended plate-condenser in consequence of the ether wind.

These experiments without exception produced a negative result. There could no longer be a doubt that a motion of translation through the ether cannot be detected by an observer who shares in the motion. Thus, the principle of relativity which holds for mechanics is also valid for all electromagnetic phenomena.

Lorentz next proceeded to bring this fact into harmony with his ether theory. To do this there seemed no other way than to assume the contraction hypothesis and to fuse it into the laws of the electron theory so as to form a uniform whole free from inner contradictions. He first observed that a system of electric charges which keep in equilibrium only through the action of their electrostatic forces, contracts of itself as soon as it is set into motion; or, more accurately, the electromagnetic forces that arise when the system is moving uniformly change the configuration of equilibrium in such a way that every length is contracted in the direction of its motion by the factor $\sqrt{1 - \beta^2}$.

Now, this mathematical theorem leads to an explanation of contraction, if we assume that all physical forces are ultimately electrical in origin or that they at least follow the same laws of equilibrium in uniformly moving systems. The difficulty of regarding all forces as electrical is due to the circumstance that they lead, in accordance with old and well-known theorems, due to Gauss, to charges being in equilibrium, but never in *stable* equilibrium. The forces which bind the atoms to form molecules and the latter to form solid bodies cannot, therefore, simply be electrical. The necessity of assuming non-electric forces comes out most clearly if we enquire into the dynamical constitution of an individual electron. This is supposed to be an accumulation of negative charge, and we must ascribe a finite extent to it, for, as we have seen (p. 178), the energy of a spherically shaped charge of radius a is equal to $\frac{1}{2}\frac{e^2}{a}$, and it becomes infinitely great if a is set equal to nothing. But the component parts of the electron strive to separate, since similar charges repel. Consequently, there must be a new force which keeps them together. In Abraham's theory of the electron it is assumed that an electron is a *rigid* sphere, that is, the non-electric forces are to be so great that they admit of no deformation whatsoever. But it is, of course, possible to make other assumptions.

Now, it suggested itself to Lorentz that the *electron* also experiences the contraction $\sqrt{1 - \beta^2}$. We have already stated above (p. 179) that then a much simpler formula results for the mass of the electron than that arising from Abraham's hypothesis. But, in addition to electromagnetic energy, Lorentz' electrons have also an energy of deformation of foreign origin, which is wanting in the rigid electron of Abraham.

Lorentz next investigated the question whether the contraction hypothesis is sufficient for deriving the principle of relativity. After laborious calculations he established that this was not the case, but he also found (1899) what assumption had to be added in order that all electromagnetic phenomena in moving systems take place just as in the ether. His result is at least just as remarkable as the contraction hypothesis. It is: *a new time measure must be used in a system which is moving uniformly.* He called this time, which differs from system to system, "local time." The contraction hypothesis may clearly be expressed thus: the measure of length in moving systems is different from that in the ether. Both hypotheses together state that space and time must be measured differently in moving systems and in the ether. Lorentz enunciated the laws according to which the measured quantities in various systems may be transformed into each other, and he proved that these transformations leave the field equations of the electron theory unchanged. This is the mathematical content of his discovery. Larmor (1900) and Poincaré (1905) arrived at similar results about the same time.* We shall get to know these relationships presently from Einstein's standpoint in a much clearer form, and so we shall not enter into them here. But we shall make evident to ourselves what consequences the new turn in Lorentz' theory had for the idea of the ether.

In the new theory of Lorentz the principle of relativity holds, in conformity with the results of experiment, for all electrodynamic events. Thus, an observer perceives the same phenomena in his system, no matter whether it is at

* It is interesting historically that the formula of transformation to a moving system, which we nowadays call Lorentz' transformation (see vi, 2, p. 200 formula (72)), were set up by Voigt as early as 1877 in a dissertation which was still founded on the elastic theory of light.

rest in the ether or moving uniformly and rectilinearly. He has no means at all of distinguishing the one from the other. For even the motion of other bodies in the world, which are moving independently of him, always informs him only of relative motion with respect to them and never of absolute motion with respect to the ether. Thus he can assert that he himself is at rest in the ether, and no one can contradict him. It is true that a second observer on another body, moving relatively to the first can assert the same with equal right. There is no empirical and no theoretical means of deciding whether one of them or which is right.

Thus, here we get into exactly the same position with respect to the ether as that into which the classical principle of relativity of mechanics brought us with respect to the absolute space of Newton (III, 6, p. 62). In the latter case we had to admit that it is meaningless to regard a definite place in absolute space as something *real* in the sense of physics. For there is no mathematical means of fixing a place in absolute space or of finding it a second time. In precisely the same way we must now admit that a definite position in the ether is nothing real in the physical sense, and through this the ether itself entirely loses the character of a substance. Indeed, we may even say: if each of two observers who are moving relatively to each other can assert with equal right that he is at rest in the ether, there can be no ether.

Thus, the furthermost development of the ether theory leads to the dissolution of its fundamental conception. But it has required a great effort to admit the emptiness of the ether idea. Even Lorentz, whose ingenious suggestions and laborious efforts have led the ether theory to this crisis, hesitated for a long time to take this step. The reason is this. The ether was conceived for the express purpose of having a carrier of light vibrations, or, more generally, of the electromagnetic forces in empty space. Vibrations without something which vibrates are unthinkable. But we have already pointed out above in discussing Ritz' theory that to assert that in empty space, too, there are *observable* vibrations goes beyond all possible experience. Light or electromagnetic forces are never observable except in connexion with bodies. Empty space free of all matter is no object of observation at all. All that we can ascertain is that an action starts out from one material body and arrives at another material body some time later. What occurs in the interval is purely hypothetical, or, more precisely expressed, arbitrary. This signifies that theorists may use their own judgment in equipping a vacuum with phase quantities (denoting state), fields, or similar things, with the one restriction that these quantities serve to bring changes observed with respect to material things into clear and concise relationship.

This view is a new step in the direction of higher abstraction and in releasing us from common ideas that are apparently necessary components of our world of thought. At the same time, however, it is an approach to the ideal of allowing only that to be valid as constructive elements of the physical world which is directly given by experience, all superfluous pictures and analogies which originate from a state of more primitive and more unrefined experience being eliminated.

From now onwards ether as a substance vanishes from theory. In its place we have the abstract " electromagnetic field " as a mere mathematical device for conveniently describing processes in matter and their regular relationships.*

Whoever is inclined to shrink from such a formal view is advised to think of the following fully analogous abstraction, to which he has long accustomed himself.

To determine a place on the earth's surface trigonometrical signs, denoting the geographical latitude and longitude, are placed on church towers, mountain peaks, and other prominent points. There is, however, nothing of this on the sea. There the meridians of latitude and longitude are only imagined to be drawn, or, as we often say, they are virtual. If a captain on a ship wants to ascertain his position, he converts a point of intersection of one of these imaginary lines into reality by astronomical observations, he converts the virtual point into a real one. The electromagnetic field is to be regarded quite similarly. The solid surface of the earth corresponds to matter, the trigonometrical signs to the physical changes that are ascertainable. But the sea corresponds to a vacuum, the meridians of longitude and latitude to the imagined electromagnetic field, which is virtual until a test body is brought up, which makes the field visible through its own actual changes, just like the boat makes real the geographical position.

Only the reader who has made this view really his own will be able to follow the later development of the doctrine of space and time. Different people find progressive abstraction, objectivation, and relativization easy or difficult, as the case may be. The older peoples of the Continent, Dutch,

* Einstein has recently proposed to call empty space equipped with gravitational and electromagnetic fields the "ether," whereby, however, this word is not to denote a substance with its traditional attributes. Thus, in the "ether," there are to be no determinable points, and it is meaningless to speak of motion relative to the "ether." Such a use of the word "ether" is of course admissible, and, when once it has been sanctioned by usage in this way, probably quite convenient.

French, Germans, Italians, Scandinavians, are most susceptible to these ideas, and are most deeply engaged in elaborating this system. Englishmen, who incline to concrete ideas, are less readily accessible. Americans are fond of attaching themselves to mechanical pictures and models. Even Michelson, whose experimental researches had the greatest share in destroying the ether theory, repudiates a theory of light without the ether as unthinkable. But the younger generation is already being educated in the sense of the new views, and accepts as self-evident what was regarded by the older school as an unheard-of innovation.

If we cast our eyes over its development we see that the ether theory comes to an end with the theory of relativity, which is its closing chapter. Ether as a substance disappears as a superfluous hypothesis, and the principle of relativity comes out the more clearly as the fundamental law of physics. This gives us the task of building up the physical world afresh on this new and surer basis. We now thus arrive at the researches of Einstein.

CHAPTER VI

EINSTEIN'S SPECIAL PRINCIPLE OF RELATIVITY

1. The Conception of Simultaneity

THE difficulties which had to be overcome by applying the principle of relativity to electrodynamical events consisted in bringing into harmony the following two apparently inconsistent theorems:—

1. According to classical mechanics the velocity of any motion has different values for two observers moving relatively to each other.

2. But experiment informs us that the velocity of light is independent of the state of motion of the observer and has always the same value c.

The older ether theory endeavoured to get rid of the contradiction between these two laws by dividing the velocity of light into two components, (*a*) the velocity of the luminiferous ether, and (*b*) the velocity of light with respect to the ether; of these two (*a*) can be appropriately determined by convection coefficients. This theory, however, was successful in eliminating the contradiction only with regard to quantities of the first order. To maintain the law of constancy of the propagation of light Lorentz' theory had to introduce a special length measure and time measure for every moving system. The law then comes about as the expression of a sort of "physical illusion."

In 1905 Einstein recognized that Lorentz contractions and local times were not mathematical devices and physical illusions but involved the very foundations of space and time.

Of the two laws 1 and 2, the first is purely theoretical and conceptual in character whereas the second is founded on fact.

Now, since the second law, that of the constancy of the velocity of light, must be regarded as being experimentally established with certainty, nothing remains but to give up the first law and hence the principles of determination of space and time as hitherto effected. Thus there must be an error in these principles or at least a prejudiced view, due to a confusion of habits of thought with logical consistency, and this we all realize to be an obstacle to progress.

Now, the *conception of simultaneity* is a prejudice of this type.

It is regarded as self-evident that there is sense in the statement that an event at the point A, say the earth, and an event at the place B, say the sun, are simultaneous. It is assumed that conceptions like " moment of time," " simultaneity," " earlier," " later," and so forth have a meaning in themselves *a priori* which is valid for the whole universe. This was Newton's standpoint, too, when he postulated the existence of an absolute time or duration of time (III, 1, p. 50), which was to pass " uniformly and without reference to any external object whatsoever."

But there is certainly no such time for the quantitative physicist. He sees no meaning in the statement that an event at A and an event at B are simultaneous, since he has no means of deciding the truth or the incorrectness of this assertion.

For, to be able to decide whether two events at different points are simultaneous we must have clocks at every point which we know with certainty to go at the same rate or to beat " synchronously." Thus the question resolves into this : can we define a means of testing the equal rate of beating of two clocks situated at different points ?

Let us imagine the two clocks at A and B a distance l apart at rest in a system of reference S. Now there are two methods of regulating the clocks to go at the same rate.

1. We may take them to the same point, regulate them there so that they go in unison and then restore them to A and B respectively.

2. We may use time-signals to compare the clocks.

Both processes are adopted in practice. A ship at sea carries with it a chronometer which beats accurately and which has been regulated in accordance with the normal clock in the home-port, and, moreover, it receives time-signals by wireless telegraphy.

The fact that these signals are regarded as necessary proves our lack of confidence in " transported " time. The practical weakness of the method of transportable clocks consists in the circumstance that the smallest error in the beating increases continually. But even if the assumption is made that there are ideal clocks free from error (such as the physicist is convinced he has in the atomic vibrations that lead to the emission of light), it is logically inadmissible to base on them the definition of time in systems moving relatively to each other. For the equal beating of two clocks, however good they may be, can be tested directly, that is, without the intervention of signals only if they are at rest relatively to each other. It cannot be established without signals that they maintain the same rate when in relative motion. This would be a pure hypothesis which we should have to endeavour to avoid if we wished to adhere to the principles of physical research. This impels us to adopt the method of time-signals to define time in systems moving relatively to each other. If this allows us to arrive at a method of measuring times which is free from contradictions, we shall have to investigate subsequently how an ideal clock has to be constituted in order that it should always show the " right " time in systems moving arbitrarily (see VI, 5, p. 214).

Let us picture a long series of barges B, C, D, drawn by a steam-tug A over the sea. Suppose there is no wind but that the fog is so thick that each ship is invisible to the others. Now, if the clocks on the barges and the tug are to be compared, sound signals will be used. The tug A sends out a shot at 12 o'clock and when the sound is audible on the barges, the bargemen will set their clocks at 12. But it is clear that in doing so they commit a small error, since the sound requires a short time to arrive from A at B, C, . . . If the velocity of sound c is known this error can be eliminated. c is equal to about 340 metres per sec. If the barge B is at a distance $l = 170$ metres behind A, the sound will take $t = \frac{l}{c} = \frac{170}{340} = \frac{1}{2}$ sec. to travel from A to B, and hence the clock at B must be set $\frac{1}{2}$ sec. after 12 at the moment when the sound reaches it. But this correction, again, is right only if the tug and barges are at rest. As soon as they are in motion the sound clearly requires less time to get from A to B, because the barge B is moving towards the sound wave. If we now wish to apply the true correction we must know the absolute velocity of the ships with respect to the air. If this is unknown it is impossible to compare times absolutely with the help of sound. In clear weather we can use light in place of sound. Since light travels enormously more quickly the error will at any rate be small, but in a question of principle the absolute magnitude is of no account. If we imagine in place of the tug and the barges a heavenly body in the sea of ether, and light signals in place of sound signals, all our reflections remain valid and unchanged. There is no more rapid messenger than light in the universe. We see that the theory of the absolutely stationary ether leads to the conclusion that an absolute comparison of times can be carried out in moving systems only if we know the motion with respect to the ether.

But the result of all the experimental researches was that it is impossible to detect motion with respect to the ether by physical means. From this it follows that absolute simultaneity can likewise be ascertained in no way whatsoever.

The paradox contained in this assertion vanishes if we make clear to ourselves that to compare times by means of light signals we must know the exact value of the velocity of light, but that the measurement of the latter again entails the determination of a length of time. Thus we are obviously moving in a vicious circle.

Now, even if we cannot attain absolute simultaneity it is possible, as Einstein has shown, to define a relative simultaneity for all clocks that are relatively at rest with respect to each other, and yet it is not necessary to know the value of the velocity of the signals.

We shall first show this for the case of the tug and the barges. When they are at rest we can make the clocks in the boats A and B (Fig. 111) go at the same rate by the following means. We place a boat C exactly half way between the boats A and B and send off a shot from C. Then it must be heard simultaneously at A and B.

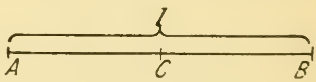


FIG. 111.

Now if the boats S are in motion we may clearly apply the same method. If it does not occur to the bargemen that they are moving relatively to the air they will be convinced that the clocks at A and B go equally quickly.

Suppose a second series of boats S′, whose barges A′, B′, C′ are at exactly the same distances apart as the corresponding boats of the first series S, compare their clocks in exactly the same way. If, now, the one series overtakes the other, whether the latter be at rest or not, the ships will pass each other, and at a certain moment A will coincide with A′ and B with B′, and the bargemen can test whether their clocks agree. Of course they will find that they do not. Even if A and A′ should accidentally be beating synchronously, B and B′ will not do so.

This will bring the error to light. When the boats are in motion the signal from the middle-point C evidently takes more time to arrive at the preceding ship A and less time to arrive at the following ship B than when the boats are at rest, because A is moving away from the sound wave whilst B is moving towards it, and this difference varies with the velocities of the two series.

Now in the case of sound, *one* system has the correct time, namely, that which is at rest relative to the air. In the case of light, however, it is not possible to assert this because absolute motion with respect to the luminiferous ether is a conception which, according to all experience, has no physical reality. The method we illustrated for sound just now to regulate clocks is, of course, also possible with light. The clocks at A and B are set so that every flash of light sent out from the middle-point C of the distance AB reaches the clocks A and B just as their hands are in the same position. In this way every system can have the synchronism of its clocks adjusted. But when two such systems meet each other and if the clocks A, A′ agree in time then B, B′ will exhibit different positions of the hands. Each system may claim with *equal* right that it has the correct time, for each can assert that it is at rest, since all physical laws are the same in each.

But when two claim what, by its very meaning, can belong to only one, it must be concluded that the claim itself is meaningless. *There is no such thing as absolute simultaneity.*

Whoever has once grasped this will find it difficult to understand why it took many years of exact research until this simple fact was recognized. It is a repetition of the old story of the egg of Columbus.

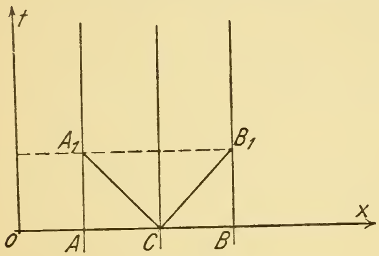


FIG. 112.

The next question is whether the method of comparing clocks which we have introduced leads to a consistent relative conception of time.

This is actually the case. To see this we shall use Minkowski's representation of events or world-points in an *xt*-plane, in which we restrict ourselves to motions in the *x*-direction and thus omit those in the *y*- and *z*-directions Fig. 112).

The points A, B, and C, that are at rest on the *x*-axis, are represented in the *xt*-co-ordinate system as three parallels to the *t*-axis. Let the point C lie midway between A and B. At the moment $t = 0$ a light-signal is to be sent out from it in both directions.

We assume that the system S is " at rest," that is, that the velocity of light is the same in both directions. Then the light-signals moving to the right and left are represented by straight lines which are equally inclined to the *x*-axis, and which we call " *light-lines.*" We shall make their inclination 45°; this evidently amounts to saying that the same distance which represents the unit length 1 cm. on the *x*-axis in the figure signifies the very small time $\frac{1}{c}$ secs. on the *t*-axis, which the light takes to traverse the distance 1 cm.

The *t*-values of the points of intersection A_1, B_1, of the light-lines with the world-lines of the points A and B give the times at which the two light-signals arrive. We see that A_1 and B_1 lie on a parallel to the *x*-axis, that is, they are simultaneous.

The three points, A, B, C, are next to be moving uniformly with the same velocity. Their world-lines are then again parallel, but inclined to the *x*-axis (Fig. 113). The light-signals are represented by the same light-lines, proceeding from C, as above, but their points of intersection, A_1', B_1' with the world-lines A, B do *not* now lie on a parallel to the *x*-axis; thus they are *not* simultaneous in the *xt*-co-ordinate system, and B_1' is later than A_1'. On the other hand an observer moving with the system can with equal right assert that A_1', B_1' are simultaneous events (world-points). He will use an *x't'*-co-ordinate system S′ in which the points A_1', B_1' lie on a parallel to the *x'*-axis. The world-lines of the points A, B, C are, of course, parallel to the *t'*-axis, since A, B, C are at rest in the system S′ and hence their *x'*-co-ordinates have the same values for all *t'*'s.

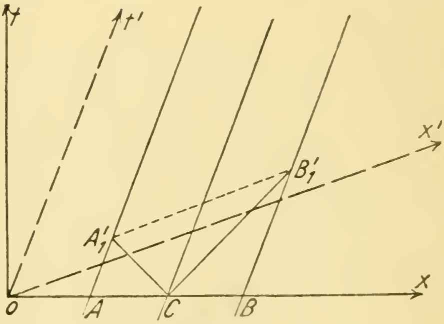


FIG. 113.

From this it follows that the moving system S′ is represented in the *xt*-plane by an oblique co-ordinate system *x't'*, in which *both* axes are inclined to the original axes.

We now recall that in ordinary mechanics the inertial systems in the *xt*-plane are likewise represented by oblique co-ordinates with *t*-axes arbitrarily directed, the *x*-axis, however, always remaining the same (III, 7, p. 66). We have already pointed out that from the mathematical point of view this is a flaw which is eliminated by the theory of relativity. We now see clearly how this happens as a result of the new definition of simultaneity. At the same time a glance at the figure convinces us without calculation that this definition must be consistent in itself. For it signifies nothing other than that we use oblique instead of rectangular co-ordinates.

The units of length and time in the oblique system are not yet determined by the construction, for this makes use only of the fact that in a system S light is propagated with equal velocity in all directions, but not of the law that the velocity of light has the same value c in all inertial systems. If we enlist the help of the latter, too, we arrive at the complete kinematics of Einstein.

2. Einstein's Kinematics and Lorentz Transformations

We once more repeat the hypothesis of Einstein's kinematics.

1. *The Principle of Relativity.*—There are an infinite number of systems of reference (inertial systems) moving uniformly and rectilinearly with respect to each other, in which all physical laws assume their simplest form (originally derived for absolute space or the stationary ether).

2. *The Principle of the Constancy of the Velocity of Light.*—In all inertial systems the velocity of light has the same value when measured with length-measures and clocks of the same kind.

Our problem is to derive the relations between lengths and times in the various inertial systems. In doing so we shall again restrict ourselves to motions that occur parallel to a definite direction in space, the x-direction.

Let us consider two inertial systems S and S′ which have the relative velocity v. The origin of the system S′ thus has the co-ordinate $x = vt$ relative to the system S at the time t. Its world-line is characterized in the system S′ by the condition $x' = 0$. The two equations must denote the same and hence $x - vt$ must be proportional to x'. We set

$$ax' = x - vt.$$

According to the principle of relativity, however, both systems are fully equivalent. Thus we may equally well apply the same argument to the motion of the origin of S relative to S′, except that now the relative velocity v has the reverse sign. Therefore $x' + vt'$ must be proportional to x, and, on account of the equivalence of both systems, the factor of proportionality α will be the same in each case :

$$ax = x' + vt'.$$

From this and the preceding equation t' may be expressed in terms of x and t. We get

$$vt' = ax - x' = ax - \frac{x - vt}{a} = \frac{1}{a}\{(a^2 - 1)\,x + vt\}$$

thus

$$at' = \frac{a^2 - 1}{v}x + t.$$

This equation, combined with the first, allows x' and t' to be calculated when x and t are known. The factor of proportionality α is as yet indeterminate, but it must be chosen so that the principle of the constancy of the velocity of light remains preserved.

The velocity of a uniform motion is represented in the system S by $U = \frac{x}{t}$ and in the system S′ by $U' = \frac{x'}{t'}$. If we divide the two equations which allow x' and t' to be expressed by x and t into each other, the factor α cancels, and we find

$$U' = \frac{x'}{t'} = \frac{x - vt}{\frac{\alpha^2 - 1}{v} x + t}.$$

If we divide the numerator and the denominator on the right by t and introduce $U = \frac{x}{t}$ we get

$$U' = \frac{U - v}{\frac{\alpha^2 - 1}{v} U + 1} \quad . \quad . \quad . \quad (70)$$

If, in particular, we are concerned with the uniform motion of a ray of light along the x-axis then, by the principle of the constancy of the velocity of light, we must have $v = U'$, and the value of each is that of the velocity of light c. If, then, we set $U = c$ and at the same time $U' = c$, we must have

$$c = \frac{c - v}{\frac{\alpha^2 - 1}{v} c + 1} \qquad \text{or} \qquad \frac{\alpha^2 - 1}{v} c^2 + c = c - v.$$

Hence it follows that

$$\alpha^2 - 1 = -\frac{v^2}{c^2} = -\beta^2 \qquad \text{or} \qquad \alpha^2 + \beta^2 = 1.$$

This gives us the factor of proportionality α, namely

$$\alpha = \sqrt{1 - \beta^2} \quad . \quad . \quad . \quad (71)$$

The transformation formulæ now become

$$\alpha x' = x - vt \qquad \alpha t' = -\frac{\beta^2}{v} x + t.$$

We shall write them down once again in full, adding the co-ordinates y and z that are perpendicular to the direction of motion and that do not change:

$$x' = \frac{x - vt}{\sqrt{1 - \frac{v^2}{c^2}}} \quad y' = y \quad z' = z \quad t' = \frac{t - \frac{v}{c^2}x}{\sqrt{1 - \frac{v^2}{c^2}}} \quad . \quad (72)$$

These rules, according to which the place and time of a world-point in the system S′ may be calculated, are said to constitute a *Lorentz transformation.* They are exactly the same formulæ as Lorentz found by difficult arguments involving Maxwell's field equations (see V, 15, p. 186).

If we wish to express x, y, z, t by means of x', y', z', t' we must solve the equations. We can deduce from the equivalence of both systems S and S′ without calculation that the formulæ given by solution must have the same forms except that v becomes changed into $-v$. A calculation shows that, actually

$$x = \frac{x' + vt'}{\sqrt{1 - \frac{v^2}{c^2}}} \quad y = y' \quad z = z' \quad t = \frac{t' + \frac{v}{c^2}x'}{\sqrt{1 - \frac{v^2}{c^2}}}.$$

Particular interest attaches to the limiting case in which the velocity v of the two systems becomes very small in comparison with the velocity of light. We then arrive directly at the Galilei transformation [formula (29), p. 65]. For if $\frac{v}{c}$ can be neglected in comparison with 1, we get from (72)

$$x' = x - vt \quad y' = y \quad z' = z \quad t' = t.$$

This helps us to understand how, on account of the small value that $\frac{v}{c}$ has in most practical cases, Galilean and Newtonian mechanics satisfied all requirements for some centuries.

3. Geometrical Representation of Einstein's Mechanics

Before we seek to interpret the content of these formulæ we shall interpret the relationships which they exhibit between two inertial systems according to the manner introduced by Minkowski, namely, geometrically, in the four-dimensional *world xyzt.* In doing so we may leave the co-ordinates y, z, that remain unchanged, out of the question and restrict ourselves to a consideration of the xt-plane. All kinematical laws, then, appear as geometrical facts in the *xt*-plane. The reader is, however, strongly advised to practise translating the relationships obtained in geometrical form back into the ordinary language of kinematics. Thus, a world-line is to be taken as denoting the motion of a point, the intersection of two world-lines the meeting of the two moving points, and so forth. It is possible to simplify the processes represented in the figures by taking a ruler, passing it along the *t*-axis parallel to the *x*-axis, and keeping in view the intersections of the edge of the ruler with the world-lines. These points, then, move to and fro on the edge and give a picture of the progress of the motion in space.

As we have seen, every inertial system S (VI, I, p. 197) is represented by an oblique set of axes in the *xt*-plane. The fact that one among them is rectangular must be regarded as an accidental circumstance and plays no particular part.

Every point in space may be regarded as the point of origin of a light-wave which spreads out spherically and uniformly in all directions. Of this spherical wave only two light-signals are present which pass along the *x*-direction here alone taken into consideration. One of these moves to the left, the other to the right. Thus they are represented in the *xt*-plane by two intersecting straight lines which are, of course, entirely independent of the choice of the position of reference, since they connect time events, world-points, with each other, namely, the points of space reached successively by the light-signal.

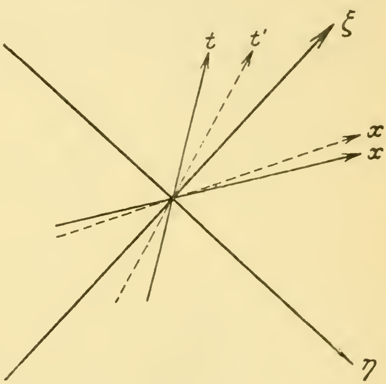


FIG. 114.

We draw these *light-lines* for a world-point which is at the same time to be the origin of all the *xt*-co-ordinate systems considered, and, indeed, we draw them as two mutually perpendicular straight lines. We choose these as the axes of an $\xi\eta$-co-ordinate system (Fig. 114).

This brings to view one of the chief characteristics of Einstein's theory. The $\xi\eta$-system is uniquely determined and fixed in the "world," although its axes are not straight lines in space but are formed by world-points which are reached by a light-signal emitted from the origin. This invariant or "absolute" co-ordinate system is thus highly abstract in its nature. We must accustom ourselves to seeing such abstractions in the modern theory replace the concrete idea of the ether. Their strength lies in the fact that they contain nothing that goes beyond the conceptions necessary to interpret the results of experience.

The *calibration curves* which cut off the units of length and time from the axes of an arbitrary inertial system xt must be rigidly connected with this absolute system of reference $\xi\eta$. These calibration curves must be represented by an invariant law and the question is to find it.

The light-lines themselves are invariant. The ξ-axis ($\eta = 0$) is represented in a system of reference S by the formula $x = ct$, and in another system of reference S′ by the formula $x' = ct'$, for these formulæ express that the velocity of light has the same value in both systems. Now we shall transform the difference $x' - ct'$ which is equal to zero for the points of the η-axis into terms of the co-ordinates x, t by means of the Lorentz transformation (72). It then follows that

$$\begin{aligned} x' - ct' &= \frac{1}{\alpha}\left\{(x - vt) - c\left(t - \frac{v}{c^2}x\right)\right\} \\ &= \frac{1}{\alpha}\left\{x\left(1 + \frac{v}{c}\right) - ct\left(1 + \frac{v}{c}\right)\right\} \\ &= \frac{1 + \beta}{\alpha}(x - ct). \end{aligned}$$

From this we see that when $x - ct = 0$, so is $x' - ct' = 0$.

For the η-axis ($\xi = 0$), $x = -ct$, and $x' = -ct'$. If we carry out the corresponding transformation from x' and ct' in terms of x, t, we have only to change c into $-c$ and β into $-\beta$ in the above (whereas $\alpha = \sqrt{1 - \beta^2}$ remains unchanged) and we get

$$x' + ct' = \frac{1 - \beta}{\alpha}(x + ct).$$

Out of these formulæ we easily read an invariant form. For $(1 + \beta)(1 - \beta) = 1 - \beta^2 = \alpha^2$, hence if we multiply the two equations the factor becomes 1 and we find that

$$(x' - ct')(x' + ct') = (x - ct)(x + ct)$$

or

$$x'^2 - c^2t'^2 = x^2 - c^2t^2$$

that is, the *expression*

$$G = x^2 - c^2t^2 \quad . \quad . \quad . \quad . \quad (73)$$

is an invariant. On account of its fundamental character we call it the *ground invariant.*

It serves in the first place for determining the units of length and time in an arbitrary system of reference S.

To do this we enquire what are the world-points for which G has the value $+ 1$ or $- 1$.

Clearly $G = 1$ for the world-point $x = 1$, $t = 0$. But this is the end-point of a scale of unit length whose other end is applied to the origin at the moment $t = 0$. As this holds in the same way for all systems of reference S, we recognize that the world-points for which $G = 1$ define the stationary unit of length for any arbitrary system of reference, as we shall presently show in greater detail.

In the same way $G = - 1$ for the world-point $x = 0$, $t = \frac{1}{c}$. Thus this world-point is correspondingly connected with the unit of time of the clock which is at rest in the system S.

Now, it is very easy to construct the points $G = + 1$, or $G = - 1$ geometrically by starting from the invariant co-ordinate system $\xi\eta$. The ξ-axis is formed by the points for which $\eta = 0$. On the other hand, the same world-points are characterized in any arbitrary inertial system S by the relation $x = ct$. Hence η must be proportional to $x - ct$. By choosing the unit of η appropriately, we may set

$$\eta = x - ct.$$

In the same way by considering the η-axis we find that we may set

$$\xi = x + ct.$$

Then we have

$$\xi\eta = (x - ct)(x + ct) = x^2 - c^2t^2 = G.$$

$G = \xi\eta$ clearly denotes the content of a rectangle with the sides ξ and η. If we wish to find a world-point for which $G = \eta\xi = 1$, we have only to choose a rectangle of unit area formed by the co-ordinates ξ, η. All these rectangles are clearly exposed to view. They include the square whose side is unity; the others are higher in proportion as they are narrower, and lower in proportion as they are wider, in agreement with the condition $\eta = \frac{1}{\xi}$ (Fig. 115). The points ξ, η clearly form a curve which approaches the ξ- and the η-axis more and more closely. This curve is called an *equilateral hyperbola*. If ξ and η are both negative, $\xi . \eta$ is positive. Hence the construction gives us a second branch, the image of the first, in the opposite quadrant.

For $G = - 1$ the same construction holds as in the other quadrants, for which the co-ordinates ξ and η have different signs.

The four hyperbolæ now form the *calibration curves* which we are seeking and by which the units of length and time are fixed for all systems of reference.

Let the x-axis meet the branch $G = +1$ of the hyperbola at the points P and P′, and the t-axis the branch $G = -1$ of the hyperbola at Q and Q′ (Fig. 116).

Through P we draw a parallel to the t-axis and we assert that this does not cut the right branch $G = +1$ of the calibration curve in a second point but just touches it at P. In other words, we assert that not a single point of this branch of the calibration lies to the left of the straight line but that the whole branch runs to the right of it, that is, all its points have x-co-ordinates that are greater than the distance OP.

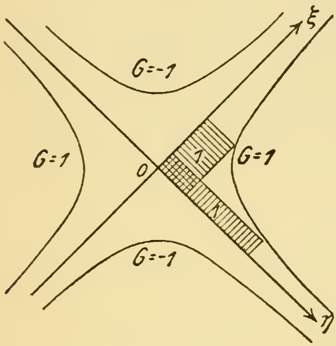


FIG. 115.

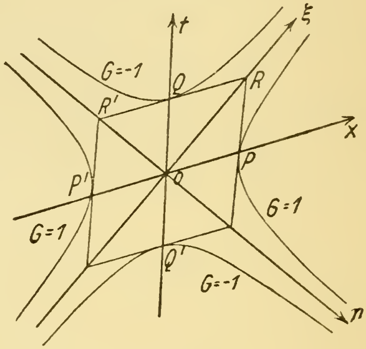


FIG. 116.

This is actually the case, since for every point of the calibration curve $G = x^2 - c^2t^2 = 1$, we have $x^2 = 1 + c^2t^2$. Thus $x^2 = 1$ for the point P of the calibration curve, which at the same time lies on the x-axis $t = 0$, but for every other point on the calibration curve x^2 is greater than 1 by the positive amount c^2t^2. Accordingly $OP = 1$ and for every point of the right branch of the calibration curve x is greater than 1.

In just the same way it follows that the parallel through P′ to the t-axis touches the left branch of the hyperbola $G = 1$ at P′, and that the parallels through Q and Q′ to the x-axis touch the branches $G = -1$ of the hyperbola at Q and Q′. This clearly makes the distance $OQ = \frac{1}{c}$. For the point Q lies on the calibration curve $G = x^2 - c^2t^2 = -1$ and $x = 0$ in the t-axis, thus $c^2t^2 = 1$, or $t = \frac{1}{c}$ is the value of OQ.

The two parallels to the t-axis through P and P′ meet the light-lines ξ, η in the points R and R′. But the parallels through Q and Q′ to the x-axis pass through the same points. For the point R, for example, we have $x = ct$, because it lies on the ξ-axis, and $x = 1$ because it lies on the parallel to the t-axis through P. From this it follows that $t = \frac{1}{c}$, that is, R lies on the parallel to the x-axis through Q.

Now we see that this construction of the x-axis agrees with that previously given (p. 201), involving simultaneous world-points. For the t-axis OQ and the two parallels PR and P′R′ are the world-lines of three points, one of which, O, lies midway between the other two. P, P′. Now if a light-signal be sent out from O in both directions, it will be represented by the light-lines ξ, η, and hence cuts the two external world-lines in R and R′. Consequently, these two world-points are simultaneous, their connecting line is parallel to the x-axis, exactly as given by our new construction.

We condense the result of our reflections into the following short statement:

The axes x *and* t *of a system of reference* S *are so situated with respect to each other that each is parallel to that straight line which is touched by the calibration curve at the point of intersection with the other axis.*

The unit of length is represented by the distance OP. The unit of time is determined by the distance OQ, which does not denote 1 sec. but $\frac{1}{c}$ sec.

Every world-line that meets the branches $G = 1$ of the calibration curve may be taken as the x-axis. The t-axis is then fixed as a parallel to the straight line which touches at P. In the same way the t-axis, too, may be chosen as an arbitrary world-line meeting the curves $G = -1$ of the calibration curves. The corresponding x-axis is uniquely determined by the analogous construction.

These rules take the place of the laws of classical kinematics. There the x-axis was the same for all inertial systems, the unit of length was given on it as fixed, and the unit of time was equal to the section, cut off by a definite straight line parallel to the x-axis from the t-axis which is in general oblique (see p. 66, Fig. 41).

Now, how does it happen that these apparently so different constructions can actually scarcely be distinguished ?

This is due to the enormously great value of the velocity of light c, if we measure the latter in cms. and seconds. For if we wish to represent 1 sec. and 1 cm. in the figure by lines of the *same* length, we must obviously compress the diagram in the direction of t, so that all distances parallel to the t-axis are shortened in the ratio $1:c$. If c were equal to 10, the picture given would be something like that depicted in Fig. 117. The two light-lines would form a very sharp angle representing the freedom of play of the x-axes, and, on the other hand, the angular space of the t-axes would become very great. The greater the value of c the more prominently the quantitative difference between the ranges of free-play of the x- and the t-direction comes into evidence. For the real value of c, namely, $c = 3 \cdot 10^{10}$ cms. per sec. the drawing could not be executed on the paper at all: both light-lines would coincide in practice, and the x-direction which always lies between them would thus be constant. This is exactly what ordinary kinematics assumes. Hence we see that this is a special case or, rather, a limiting case of Einstein's kinematics, namely, that for which the velocity of light is infinitely great.

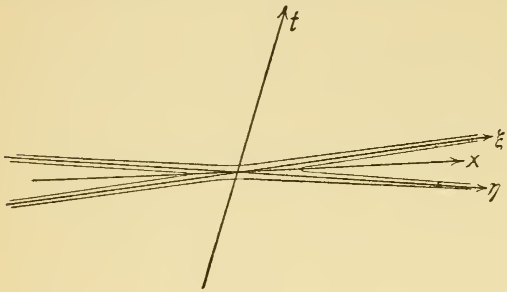


Fig. 117.

4. Moving Measuring-rods and Clocks

We shall now answer the simplest questions of kinematics, which concern the measurement of the length of one and the same measuring-rod and of the duration of one and the same time in different systems of reference.

Let a rod of unit length be placed at the origin of the system S along the x-axis. We enquire what its length is in the system S′. It is at once evident that it will not also be equal to 1. For the observers moving with S′ will, of course, measure the positions of the end-points of the rod simultaneously, that is, simultaneously in the system of reference S′. But this does *not* mean simultaneously in the system of reference S. Thus, even if the position of one end of the rod is read off simultaneously in S and S′, that of the other end will *not* be read off simultaneously with respect to the S-time by the observers of the systems S and S′. In the meantime, rather, the system S′ has moved forward, and the reading of the S′-observers concerns a displaced position of the second end of the rod.

At first sight this matter seems hopelessly complicated. There are opponents of the principle of relativity, simple minds, who, when they have become acquainted with this difficulty in determining the length of a rod, indignantly exclaim : " Of course, everything can be derived if we use false clocks. Here we see to what absurdities blind faith in the magic power of mathematical formulæ leads us," and then they condemn the theory of relativity at one stroke. Our readers will, it is hoped, have grasped that the formulæ are by no means the essential feature, but that we are dealing with purely conceptual relationships, which can be understood quite well without mathematics. Indeed, we might do not only without the formulæ but also without the geometrical figures and present the whole thing in ordinary words, but then this book would become so diffuse and so impossible of design that no one would publish it and no one could be found to read it.

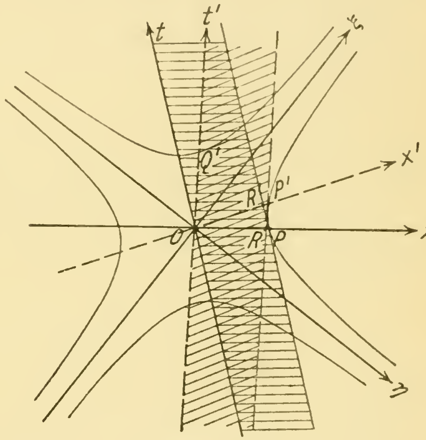


FIG. 118.

We first use our figure in the *xt*-plane to solve the question of determining the length of the rod in the two systems S and S′ (Fig. 118).

The rod is supposed at rest in the system S (x, t). Accordingly the world-line of its initial point is the t-axis and of its end point is the straight line parallel to this at the distance 1 ; the latter touches the calibration curve at the point P. Hence the whole rod is represented for all times by the strip between these two straight lines.

Now its length is to be determined in the system S′ (x', t'), which is moving with respect to S. Thus its t'-axis is inclined to the t-axis. We find the corresponding x'-axis by drawing the tangent at the point of intersection Q of the t'-axis with the calibration curve and then draw the parallel OP′ to this tangent through O. The distance OP′ is the unit of length on the x'-axis. The length of the rod of unit length at rest in the system S as measured in the system S′ is, however, determined by the distance OR′ which the parallel strip representing the rod cuts out of the x'-axis. This is clearly shorter than OP′, thus OR′ is less than 1, and hence the rod appears shortened in the moving system S′.

This is exactly the contraction devised by Fitzgerald and Lorentz to explain Michelson and Morley's experiment. Here it appears as a natural consequence of Einstein's kinematics.

If, conversely, a rod at rest in the system S′ is measured from the system S, it evidently likewise appears contracted and *not* lengthened. For such a rod is represented by the strip which is bounded by the t'-axis and the world-line parallel to it through the point P′. But the latter meets the unit distance OP of the system S in an internal point R, so that OR is smaller than 1.

Thus the contraction is reciprocal, and this is what the theory of relativity demands. Its magnitude is best found with the help of the Lorentz transformation (72).

Let l_0 be the length of the rod in the system of reference S'_0 in which it is at rest; l_0 is called the *statical length* or *proper length* of the rod.

Now, if the length of the rod is to be ascertained as it appears to an observer at rest in the system S, we must set $t = 0$, which expresses the fact that the position of both ends of the rod are read off simultaneously with regard to S. Then it follows from the first equation of the Lorentz transformation (72) that

$$x' = \frac{x}{\sqrt{1 - \frac{v^2}{c^2}}}.$$

Now, for the initial point of the rod $x = 0$, and hence also $x' = 0$. For its final point $x' = l_0$, and if $x = l$ denotes the length of the rod as measured in the system S, we get

$$l = l_0\sqrt{1 - \frac{v^2}{c^2}} \quad . \quad . \quad . \quad (74)$$

This states that the length of the rod in the system S appears shortened in the ratio $\sqrt{1 - \beta^2} : 1$, exactly in agreement with the contraction hypothesis of Fitzgerald and Lorentz (V, 15, p. 187).

The same reflections apply to the determination of an interval of time in two different systems S and S′.

We suppose clocks that go at the same rate to be placed at each of the space-points of the system S. These have a definite position of the hands with respect to S at the same moment. The position $t = 0$ is represented by the world-points of the x-axis, and the position $t = \frac{1}{c}$ by the world-points of the straight line which passes through the point Q and is parallel to the x-axis (Fig. 119).

Suppose a clock, which also indicates $t' = 0$ when $t = 0$, to be placed at the origin of the system S′. We next enquire what position the hands of a clock of the system S has which is at the point where the clock at rest in S′ exactly indicates the time $t' = \frac{1}{c}$. The required value of t is evidently determined by the intersection Q′ of the t'-axis with the calibration curve $G = -1$. On the other hand the position $t = \frac{1}{c}$ of the hands of the clocks at rest in S is represented by the points of the straight line which is drawn through Q parallel to the x-axis. This straight line meets the t'-axis at a point R′, and the figure shows that Q′ lies outside the distance QR′. This, however, denotes that the unit of time of the system S′ appears lengthened in the system S.

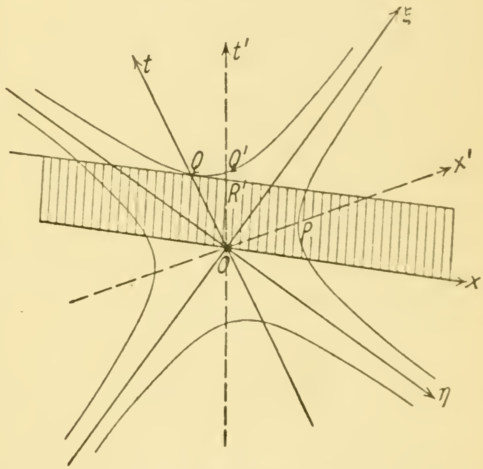


FIG. 119.

To ascertain the amount of the lengthening we set $x' = 0$ in the Lorentz transformation for the clock situated at the origin of S′, that is, $x = vt$. We then get

$$t' = \frac{t - \frac{v}{c^2}x}{\sqrt{1 - \frac{v^2}{c^2}}} = t\sqrt{1 - \frac{v^2}{c^2}}.$$

An interval of time t_0 in the system S′, that is, $t' = t_0$, will accordingly be measured in the system S as

$$t = \frac{t_0}{\sqrt{1 - \frac{v^2}{c^2}}} \quad . \quad . \quad . \quad (75)$$

and thus appears lengthened. The time-dilatation is reciprocal to the contraction in length.

Conversely, of course, the unit of time of a clock at rest in the system S appears increased in the system S′.

Or we may say that, viewed from any one system, the clocks of every other system moving with respect to it appear to be losing time. The course of events in time in the systems in relative motion are slower, that is, all events in a moving system lag behind the corresponding event in the system regarded at rest. We shall return later to the consequences which arise from this fact and which are often regarded as being paradoxical.

The time-datum of a clock in the system of reference in which it is at rest is called the PROPER TIME of the system. This is identical with the "local time" of Lorentz. The progress made by Einstein's theory is not in respect of formal laws but rather of the fundamental view of them. Lorentz made the local time appear as a mathematical auxiliary quantity in contrast to the true absolute time. Einstein established that there is no means of determining this absolute time or of distinguishing it from the infinite number of equivalent local times of the various systems of reference that are in motion. But this signifies that absolute time has no physical reality. Time-data have a significance only relatively to definite systems of reference. This completes the relativization of the conception of time.

5. Appearance and Reality

Now that we have become acquainted with the laws of Einstein's kinematics in the double form of figures and formulæ, we must throw a little light on it from the point of view of the theory of knowledge.

It might be imagined that Einstein's theory furnishes no new knowledge about things of the physical world but is concerned only with definitions of a conventional type which do, indeed, satisfy the facts but might equally well be replaced by others. This suggests itself to us if we think of the starting-point of our reflections, the example of the tug, in which the conventional and arbitrary nature of Einstein's definition of simultaneity forces itself on our attention. As a matter of fact, Einstein's kinematics can be applied in its entirety to ships that move through motionless air if we use sound signals to regulate the clocks. The quantity c would then denote the velocity of sound in all formulæ. Every moving ship would have its own units of length and time according to its velocity, and the Lorentz transformations would hold between the measure systems of the various ships. We should have before us a consistent Einsteinian world on a small scale.

But this world would be consistent only so long as we admit that the units of length and time are to be restricted by no postulate other than the two principles of relativity and the constancy of the velocity of sound or light respectively. Is this the meaning of Einstein's theory?

Certainly not. Rather it is assumed as self-evident that a measuring-rod which is brought into two systems of reference S and S′ under exactly the same physical conditions in each case represents the same length in them, it being assumed that they are affected as little as possible by external forces. A fixed rod that is at rest in the system S and is of length 1 is, of course, also to have the length 1 in the system S′ when it is at rest in S′ provided that the remaining physical conditions (gravitation, position, temperature, electric and magnetic fields, and so forth) are as much as possible the same in S′ as in S. Exactly the same would be postulated for the clocks.

We might call this tacitly made assumption of Einstein's theory the "principle of the physical identity of the units of measure."

As soon as we are conscious of this principle we see that to apply Einstein's kinematics to the case of the ships and to compare clocks with sound signals is incompatible with it. For if the units of length and time are determined according to Einstein's rule with the help of the velocity of sound, they will, of course, by no means be equal to the units of length and time measured with fixed measuring rods and ordinary clocks; for the former are not only different on every moving ship according to its velocity but, moreover, the unit of length in the direction of motion is different from that athwart the ship. Thus Einstein's kinematics would be a possible definition but in this case not even a useful one. The ordinary measuring rods and clocks would without doubt be superior to it.

For the same reason it is only possible with difficulty to illustrate Einstein's kinematics by means of models. These certainly give the relationships between the lengths and the times in the various systems correctly, but they are inconsistent with the principle of the identity of the units of measure; nothing can be done but to choose two different scales of length in two systems S and S′ of the model moving relatively to each other.

According to Einstein the state of affairs is quite different in the real world. In it the new kinematics is to be valid just when the *same* rod and the *same* clock are used first in the system S, and then in the system S′ to fix the lengths and the times. Through this, however, Einstein's theory rises above the standpoint of a mere convention and asserts definite properties of real bodies. This gives it its fundamental importance for the whole physical view of nature.

This important circumstance comes out very clearly if we fix our attention on Römer's method of measuring the velocity of light with the help of Jupiter's moons. The whole solar system moves relatively to the fixed stars. If we imagine a system of reference S rigidly connected with the latter, then the sun and its planets define another system S′. Jupiter and his satellites form a clock (ideally good) with hands. It moves round in a circle so that at one time it arrives at the direction of the relative motion of S with respect to S′, at another in the opposite direction. We can by no means arbitrarily determine the beating of the Jupiter clock in these positions by convention in such a way that the time that the light takes to traverse the diameter of the earth's orbit is the same in all directions, but rather this is so *quite of itself*, thanks to the way the Jupiter clock is arranged. For it just shows the proper time of the solar system S′ and not some absolute time or the foreign time of the system S of the fixed stars. In other words, the time of revolution of the Jupiter moons is constant relatively to the solar system (the velocity of Jupiter himself relative to the solar system being left out of account).

Now it is asserted by some that this view denotes *a transgression of the causal law*. For if one and the same measuring rod, as judged from the system S, has a different length according as it is at rest in S or moving relatively to S then, so these people say, there must be a reason for this change. But Einstein's theory gives no reason; rather it states that the contraction occurs of itself, that it is an accompanying circumstance of the fact of motion. Now, this objection is not justified. It is due to a too limited view of the conception "change." In itself such a conception has no meaning. It denotes nothing absolute, just as data denoting sizes or times have no absolute significance. For we do not mean to say that a body which is moving uniformly rectilinearly with respect to an inertial system S "undergoes a change" although it actually *changes* its *situation* with respect to the system S. It is by no means clear *a priori* what "changes" physics counts as effects for which causes are to be found; rather, this is to be determined by experimental research itself.

The view of Einstein's theory about the nature of the contraction is as follows:

A material rod is physically not a spatial thing but a space-time configuration. Every point of the rod exists at this moment, at the next, and still at the next, and so forth, at every moment of time. The adequate picture of the rod under consideration (one-dimensional in space) is thus not a section of the x-axis but rather a strip of the xt-plane (Fig. 120). The same rod, when at rest in various moving systems S and S′ is represented by various strips. There is NO *a priori* rule as to how these two-dimensional configurations of the xt-plane are to be drawn so that they may represent the physical behaviour of one and the same rod at different velocities correctly. To achieve this a calibration curve in the xt-plane must first be fixed. Classical kinematics draws this differently from Einsteinian kinematics. It cannot be ascertained *a priori* which is correct. In the classical theory both strips have the same width as measured parallel to a fixed x-axis. In Einstein's theory they have the same width as measured in the various x-directions of the systems of reference in relative motion and with different but determinate units. The "contraction" does not affect the strip at all but rather a section cut out of the x-axis. It is, however, only the strip as a manifold of world-points, events, which has physical reality, and not the cross-section. Thus the contraction is only a consequence of our way of regarding things and is not a change of a physical reality. Hence it does not come within the scope of the conceptions of cause and effect.

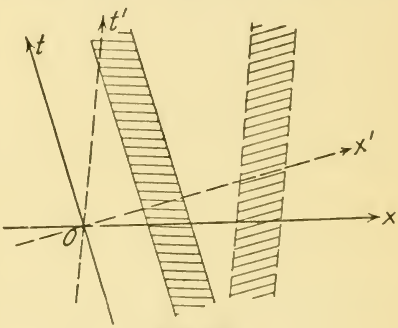


FIG. 120.

The view expounded in the preceding paragraph does away with the notorious controversy as to whether the contraction is "real" or only "apparent." If we slice a cucumber, the slices will be the larger the more obliquely we cut them. It is meaningless to call the sizes of the various oblique slices "apparent," and to call, say, the smallest which we get by slicing perpendicularly to the axis the "real" size.

In exactly the same way a rod in Einstein's theory has various lengths according to the point of view of the observer. One of these lengths, the statical length, is the greatest, but this does not make it more real than the others. The application of the distinction between "apparent" and "real" in this naive sense is no more reasonable than asking what is the real x-co-ordinate of a point x, y when it is not known which xy-co-ordinate system is meant.

Exactly corresponding remarks apply to the relativity of time. An ideal clock has always one and the same rate of beating in the system of reference in which it is at rest. It indicates the " proper time " of the system of reference. Regarded from another system, however, it goes more slowly. In such a system a definite interval of the proper time seems longer. Here, too, it is meaningless to ask what is the " real " duration of an event.

When understood in the right way, Einstein's kinematics contains no obscurities and no inconsistencies. But many of its results appear contrary to our customary forms of thought and to the doctrines of classical physics. When these antitheses occur in a particularly marked way they are often felt to be impossible and paradoxical. In the sequel we shall draw numerous deductions from Einstein's theory which first encountered violent opposition until physicists succeeded in confirming them experimentally. But here we wish to deal with an argument which leads to particularly remarkable results without its seeming possible to test them by experiment. We are referring to the so-called " clock-paradox."

Let us consider an observer A at rest at the origin O of the inertial system S. A second observer B is at first to be at rest at the same point O, and is then to move off with uniform velocity along a straight line, say the x-axis, until he has reached a point C, when he is to turn round and return to O along a straight line with the same velocity.

Let both observers carry with them ideal clocks which indicate their proper-time. The times lost in getting started, in turning round, and in slowing down on arrival at B can be made as short as we please by making the times occupied in moving uniformly there and back sufficiently great. If, say, the rate of the clocks should be influenced by the acceleration, this effect will be comparatively small if the times of the journey are sufficiently great, so that this effect may be neglected. But then the clock of the observer B must have lost time compared with the clock of A after B's return to O. For we know (VI, 4, p. 210) that during the periods of B's uniform motion, which are the determining factors for the result, the proper time lags behind the time of any other inertial system. This is seen particularly vividly in the geometrical picture in the xt-plane (Fig. 121). In this we have for the sake of convenience drawn the axes of the xt-system perpendicularly to each other. The world-line of the point A is the t-axis.

The world-line of the point B is the bent line OUR (drawn as a dotted line) whose corner U lies on the world-line of the turning-point C drawn parallel to the t-axis.

Through U we draw the hyperbola that arises out of the calibration curve G = 1 by the appropriate magnification. Let this meet the t-axis in Q. Then clearly the length OQ of the proper time for the observer A is exactly equal to the length OU of the proper time for the observer B. But the length of the proper time for A until the turning-point R is reached is, as the figure tells us, more than twice as great as OQ, whereas it is twice as great as OU for B. Thus, at the moment of turning round, A's clock is in advance of B's clock. The amount of this advance can easily be calculated from formula (75), in which t_0 is the proper-time of A, and t denotes the time measured in the system B. If we limit ourselves to small velocities of B and regard $\beta = \frac{v}{c}$ as a small number, we may write as an approximation instead of (75) (see the note on p. 182) :

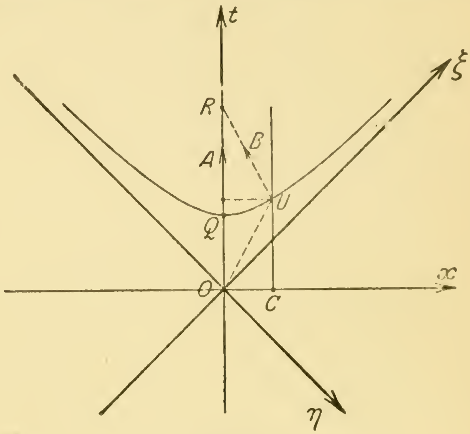


FIG. 121.

$$t = t_0 \left(1 + \tfrac{1}{2}\beta^2\right).$$

Hence the advance of A's clock with respect to B's clock is

$$t - t_0 = \frac{\beta^2}{2} t_0 \quad . \quad . \quad . \quad (76)$$

and this holds for every moment of the motion since the outward and the inward journey take place with the same velocity Hence, in particular, it also holds for the moment of turning, t_0 then denoting the whole time of the voyage according to the proper time of A, and t the time of the voyage according to the proper time of B.

The paradoxical feature of this result lies in the circumstance that *every* inertial event in the system B must take place more slowly than the same event in the system A. All atomic vibrations, indeed, even the course of life itself, must behave just like the clocks. Thus, if A and B were twin brothers, then B must be younger when he returns from his voyage than A. This is truly a strange deduction, which can, however, be eliminated by no artificial quibbling. We must put up with this just as, some centuries ago, others had to endure that their fellow-creatures in the antipodes stood on their heads. Since, as formula (76) shows, it is an effect of the second order, it is scarcely likely that practical consequences will accrue from it.

If we take up arms against this result and call it paradoxical, we simply mean that it is unusual, or "peculiar," and time will help us to conquer this strange feeling. But there are also opponents to the theory of relativity who seek to make of these conclusions an objection against the logical consistency of the theory.

Their argument is as follows. According to the theory of relativity two systems in relative motion are equivalent. We may therefore also regard B as at rest. A then performs a journey in exactly the same way as B previously, but in the opposite direction. We must therefore conclude that when A returns B's clock is in advance of A's. But previously we had come to exactly the opposite conclusion. Now since A's clock cannot be in advance of B's and at the same time B's in advance of A's, this argument discloses an inherent contradiction in the theory—so they conclude superficially. The error in this argument is obvious; the principle of relativity concerns *only* such systems as are moving uniformly and rectilinearly with respect to each other. In the form in which it has been so far developed it is *not* applicable to accelerated systems. But the system B *is* accelerated and it is *not*, therefore, equivalent to A. A is an inertial system, B is not. Later, it is true, we shall see that the general theory of relativity of Einstein also regards systems which are accelerated with respect to each other as equivalent, but in a sense which requires more detailed discussion. When dealing with this more general standpoint we shall return to the "clock paradox" and we shall show that on close examination there are no difficulties in it. For in the above we made the assumption that for sufficiently long journeys the short times of acceleration exert no influence on the beating of the clocks. But this holds *only* when we are judging things from the inertial system A and *not* for the measurement of time in the accelerated system B. According to the principles of the general theory of relativity gravitational fields occur in the latter which affect the beating of the clocks. When this influence is taken into account, it is found that under all circumstances B's clock goes in advance of A's, and thus the apparent contradiction vanishes (see VII, 10, p. 282).

The relativization of the conceptions of length and intervals of time appears difficult to many, but probably only because it is strange. The relativization of the conceptions " below " and " above " which occurred through the discovery of the spherical shape of the earth probably caused contemporaries of that period no less difficulty. In this case, too, the result of research contradicted a view that had its source in direct experience. Similarly, Einstein's relativization of time seems not to be in accord with the experience of time of individuals. For the *feeling* of " now " stretches without limit over the world, linking all being with the ego. The fact that the same moments that the ego experiences as " simultaneous " are to be called " consecutive " by another ego cannot be comprehended in fact by the actual *experience of time*. But exact science has *other* criteria of truth. Since absolute " simultaneity " cannot be ascertained, it has to eradicate this conception out of its system.

6. The Addition of Velocities

We shall now enter more deeply into the laws of Einstein's kinematics. In doing so we shall for the most part restrict ourselves to considering the xt-plane. There is no essential difficulty in generalizing the theorems obtained for the case of the four-dimensional $xyzt$-space and we shall, therefore, merely touch lightly on it.

The light-lines that are characterized by $G = x^2 - c^2t^2 = 0$ divide the xt-plane into four quadrants (Fig. 122). G evidently retains the same values in each quadrant, G being >0 in the two opposite quadrants which contain the hyperbolic branches $G = +1$, and G being <0 in the two opposite quadrants which contain the hyperbolic branches $G = -1$. A straight world-line passing through the origin O can be chosen as the x-axis or the t-axis according as it lies in the quadrants $G > 0$ or $G < 0$. Corresponding to this we distinguish world-lines as " SPACE-LIKE " or " TIME-LIKE."

In any inertial system the x-axis separates the world-points of the " past " ($t < 0$) from those of the future ($t > 0$). But this separation is different for each inertial system, since for another position of the x-axis world-points which previously lay above the x-axis, that is, in the " future," now lie below the x-axis, or in the past, and conversely. Only the events represented by the world-points within the quadrants $G < 0$ are uniquely either " past " or " future " for every inertial system. For such a world-point P we have $t^2 > \frac{x^2}{c^2}$, that is, in every admissible system of reference the time-distance of the two events O and P is greater than the time the light requires to pass from the one point to the other. We can then always introduce an inertial system S such that its t-axis passes through P, that is, in which P represents an event which takes place at the spatial origin. Regarded from another inertial system this inertial system S will move rectilinearly and uniformly in such a way that its origin coincides exactly with the events O and P. Then, obviously, for the event P in the system S, we must have $x = 0$, that is, $G = -c^2t^2 < 0$.

In every inertial system the t-axis marks off the world-points to which events that occur " before " or " behind " (after) the spatial origin on the x-axis. But for a different inertial system with a different t-axis this demarcation will be a different one. It is uniquely determined only for the world-points that lie within the quadrants $G > 0$, whether they lie " before " or " behind " the spatial origin. For such a point P we have $t^2 < \frac{x^2}{c^2}$, that is, in every admissible system of reference the time-interval between two events O and P is less than the time the light takes to pass from one to the other. Then we can introduce an appropriate moving inertial system S, whose x-axis passes through P, in which both events O and P are simultaneous. In this system we evidently have for the event P that $t = 0$, and thus $G = x^2 > 0$.

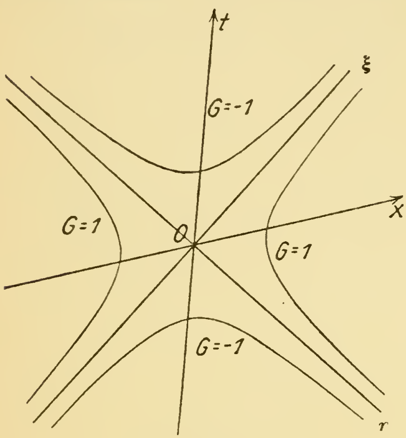


FIG. 122.

From this it follows that the invariant G is for every world-point P a measurable quantity, having a significance which we can appreciate visually ; P either allows itself to be transformed " for the same place " with O ; then $G = -c^2t^2$, where t is the difference of time of the event P with respect to the event O which occurs at the same point of space of the system S ; or else P may be transformed with O " for simultaneous times," and then $G = x^2$, where x is the spatial distance between the two events that occur in the system S.

In every co-ordinate system the light-lines $G = 0$ represent motions which occur with the velocity of light. Accordingly, corresponding to every time-like world-line there is a motion of lesser velocity. Every motion which occurs with a velocity less than that of light can be " transformed to rest " because there is a time-like world-line corresponding to it.

But what holds for motions that occur with a velocity greater than that of light ?

After the preceding remarks it would seem clear that Einstein's theory of relativity must pronounce such motions to be impossible. For the new kinematics would lose all meaning if there were signals which allowed us to control the simultaneity of clocks by means involving a velocity greater than that of light.

Here a difficulty appears to arise.

Let us assume that a system S′ has a velocity v with respect to another system S. Let a moving body K move relatively to S′ with the velocity u'. According to ordinary kinematics the relative velocity of the body K with respect to S is then

$$u = v + u'.$$

Now, if v, as well as u, is greater than half the velocity of light, then $u = v + u'$ is greater than c, and this is to be impossible according to the theory of relativity.

This contradiction is due, of course, to the circumstance that velocities cannot simply be added in the kinematics of the principle of relativity, in which every system of reference has its own units of length and time.

We already see from this that in any two systems moving with respect to each other the velocity of light always has the same value. It is just this fact that we used earlier to derive the Lorentz transformation (VI, 2, p. 198), and the formula established on p. 199 gives us the correct law for the composition of velocities if we introduce $\alpha^2 - 1 = \beta^2 = \frac{v^2}{c^2}$ into it. We find it preferable to derive this rule once again from the Lorentz-transformation (72), p. 200. To do this we divide the expressions for x' and y' (or z') by the expression for t :

$$\frac{x'}{t'} = \frac{x - vt}{t - \frac{v}{c^2}x} \qquad \frac{y'}{t'} = \frac{y\sqrt{1 - \frac{v^2}{c^2}}}{t - \frac{v}{c^2}x}.$$

If we divide the numerators and denominators of the expressions on the right by t, the quotients $u_p = \frac{x}{t}$, $u_s = \frac{y}{t}$ occur, which are obviously the projections or components, measured in the system S, of the velocity of the body K parallel (longitudinal) or perpendicular (transversal) to the direction of motion of the system S′ with respect to S. The quotients $u'_p = \frac{x'}{t'}$, $u'_s = \frac{y'}{t'}$ have the same meaning with respect to the system S′. Accordingly we get the following EINSTEIN ADDITION THEOREM OF VELOCITIES :

$$u'_p = \frac{u_p - v}{1 - \frac{vu_p}{c^2}} \qquad u'_s = u_s \frac{\sqrt{1 - \frac{v^2}{c^2}}}{1 - \frac{vu_p}{c^2}} \quad . \quad (77)$$

which takes the place of the simple formulæ of the old kinematics :

$$u'_p = u_p - v \qquad u'_s = u_s.$$

If, in particular, we are dealing with a light-ray which is travelling in the direction of motion of the system S′ with respect to S, then $u_s = 0$, $u_p = c$. And then formula (77) gives the expected result :

$$u'_p = \frac{c - v}{1 - \frac{v}{c}} = c \qquad u'_s = 0$$

which expresses the theorem of the constancy of the velocity of light. Moreover, we see that for any body moving longitudinally $u'_p < c$, so long as $u_p < c$. For if we replace u_p in the first formula (77) by the greater value c, the numerator is increased and the denominator diminished, so that the fraction becomes greater and we get

$$u'_p < \frac{c - v}{1 - \frac{v}{c}} \qquad \text{or} \qquad u'_p < c.$$

The corresponding result holds, of course, for transversal motion, and, indeed, for motion in any direction.

Hence the velocity of light is, kinematically, a limiting velocity which cannot be exceeded. This assertion of Einstein's theory has encountered much opposition. It seemed an unjustifiable limitation for future discoverers who wished to find motions which occurred with velocities greater than that of light.

We know that the β-rays of radioactive substances are electrons moving nearly with the velocity of light. Why should it not be possible to accelerate them so that they move with a velocity exceeding that of light ?

Einstein's theory, however, asserts that in principle this is not possible, because the inertial resistance or the mass of a body increases the more the more nearly its velocity approaches that of light. We thus arrive at a *new system of dynamics* which is built up on Einstein's kinematics.

7. Einstein's Dynamics

The mechanics of Galilei and Newton is intimately connected with the old kinematics. The classical principle of relativity, in particular, depends on the fact that changes of velocity, accelerations, are invariant with respect to Galilei transformations.

Now we cannot take one kinematics for one part of physical phenomena and the other kinematics for the other part, namely, invariance with respect to Galilei transformations for mechanics, and invariance with respect to Lorentz transformations for electrodynamics.

We know that the former transformations are a limiting case of the latter and are characterized by the constant c having infinitely great values. Accordingly, following Einstein, we shall assume that classical mechanics is not strictly valid at all but rather requires modifying. The laws of the new mechanics must be invariant with respect to Lorentz transformations.

In setting up these laws we must decide which fundamental laws of classical mechanics must be retained and which must be rejected or modified. The fundamental law of dynamics with which we started is the *law of momentum*, expressed by formula (7) (II, 9, p. 31), namely :

$$J = mw.$$

It is obvious that we cannot simply retain it in this form. For, whereas in classical mechanics the change of velocity w for various inertial systems has always the same value (see III, 5, p. 60), this is not the case here, on account of Einstein's Addition Theorem of Velocities (77). Thus the formula (7) has no meaning unless special directions are given for the transformation of the momentum of one system of reference to another, and hence it would not be expedient to start from it to reach the new fundamental law by generalization.

But we may certainly start from *the law of conservation of momentum* (II, 9, p. 32, formula (9)). This concerns the total momentum *carried along* by two bodies, and it states that when the bodies collide this momentum remains preserved no matter how their velocities are changed in the process. Thus the assertion involves nothing but two bodies that act on each other, that suffer a *mutual* impact, without external influences, and hence there is no reference to a third body or co-ordinate system. Accordingly we shall demand that this law of conservation of momentum still remain valid in the new dynamics.

This is, of course, as we shall see presently, impossible if we retain the maxim of classical mechanics that mass is a constant quantity peculiar to each body. Hence we shall assume from the outset that *the mass of one and the same body is a relative quantity*. It is to have different values according to the systems of reference from which it is measured, or, if measured from a definite system of reference, according to the velocity of the moving body. It is clear that the mass with respect to a definite system of reference can depend only on the value of the velocity of the moving body with respect to this system.

Now we consider two systems of reference S and S′ which are moving rectilinearly relatively to each other with the velocity v. Let there be an observer A on S, and an observer B on S′. Let these observers be furnished with two exactly equal spheres. Let the sphere of A have the same mass with respect to the system S as the sphere of B has with respect to S′, so long as the relative motions are the same.

Now, suppose each observer to throw his ball towards the other in a direction at right angles to his motion and to choose the moment of throw so that the spheres meet each other exactly *symmetrically* in flight, that is, so that the line connecting their centres is perpendicular to the direction of motion of S and S′ at the moment of collision.

If we call the longitudinal and transverse component of the first sphere U_{p1} and U_{s1}, and those of the second U_{p2} and U_{s2} we can express what these quantities measured in one of the two systems of reference S or S′ are before and after the collision.

The first sphere is thrown by A transversally with respect to S with a relative velocity U. Hence

$$U_{p1} = 0 \qquad U_{s1} = U \quad . \quad . \quad . \quad (78)$$

In the same way B throws his sphere in the opposite direction with respect to S′ with the same relative velocity U. Hence

$$U_{p}{}'_{2} = 0 \qquad U_{s}{}'_{2} = -U.$$

Now, by the Addition Theorem (77), p. 220, we can in each case transform these quantities so as to refer them to the other system. We shall give only the components in the system S, and we need merely mention that the calculation for the system S′ leads to exactly the same final result, as must be in view of the symmetry of the whole process. By inserting the values of $U_p{}'_2$ and $U_s{}'_2$ in (77) we get

$$U_{p2} = v \qquad U_{s2} = -U\sqrt{1 - \frac{v^2}{c^2}} \quad . \quad (79)$$

If we now wish to calculate the total momentum before the impact we find it advantageous not to try to write as equal the masses of the two equal spheres which are executing different motions. For it at once becomes manifest that they are necessarily different. Thus, if we designate the masses with respect to S before the collision by m_1, m_2, then the total impulse before the collision has the components

$$J_p = m_1 U_{p1} + m_2 U_{p2} = m_2 v$$

$$J_s = m_1 U_{s1} + m_2 U_{s2} = m_1 U - m_2 U\sqrt{1 - \frac{v^2}{c^2}} \qquad (80)$$

Let us next consider the effect of the collision.

Since it is to take place perfectly symmetrically, the longitudinal velocity, as observed from the system S, cannot change as a result of the collision, nor can the longitudinal velocity of the second sphere as observed from the system S′. Indeed, for reasons of symmetry, the observer A must see his sphere perform exactly the same motions as B observes his own to perform. The transverse components of velocity will become changed through the collision. Let the first sphere, as measured from S, assume the velocity $-U'$ which is oppositely directed to its original velocity. Then the second sphere, as observed from S′ must acquire the velocity U′ through the impact which is likewise oppositely directed to its original motion. Hence, after the collision, we have

$$\left.\begin{array}{ll} U_{p_1} = 0 & U_{s_1} = -U' \\ U'_{p_2} = 0 & U'_{s_2} = U' \end{array}\right\} \quad . \quad . \quad (81)$$

and by transforming to the system S by (77) we get from this

$$U_{p_2} = v \qquad U_{s_2} = U'\sqrt{1 - \frac{v^2}{c^2}} \quad . \quad (82)$$

If we designate the masses after the collision by $\overline{m}_1$, $\overline{m}_2$, then the moments of momentum after the collision come out as

$$\left.\begin{aligned} J_p &= \overline{m}_1 U_{p_1} + \overline{m}_2 U_{p_2} = \overline{m}_2 v \\ J_s &= \overline{m}_1 U_{s_1} + \overline{m}_2 U_{s_2} = -\overline{m}_1 U' + \overline{m}_2 U' \sqrt{1 - \frac{v^2}{c^2}} \end{aligned}\right\} \quad (83)$$

If we compare the momenta before and after the collision (80) and (83) we get as the conditions that there should be no change :

$$\left.\begin{aligned} m_2 v &= \overline{m}_2 v \\ m_1 U &= m_2 U \sqrt{1 - \frac{v^2}{c^2}} = -\overline{m}_1 U' + \overline{m}_2 U' \sqrt{1 - \frac{v^2}{c^2}} \end{aligned}\right\} \quad (84)$$

Now, if the mass were constant, that is if $m_1 = m_2 = \overline{m}_1 = \overline{m}_2$, then the first equation would be identically true, but the second would lead to an inconsistency. For then it would follow that

$$(U + U')\left(1 - \sqrt{1 - \frac{v^2}{c^2}}\right) = 0$$

and this is impossible, since U and v are certainly not equal to zero.

Hence we must drop the law of classical mechanics that mass is constant, and we must replace it by the assumption already made above, that the mass of a body with respect to a system S depends on the magnitude of its velocity relative to this system.

We can calculate the value U of the velocity from the components U_p and U_s according to formula (3) (II, 3, p. 24)

$$U = \sqrt{U^2_p + U^2_s}$$

According to this we get for the velocities of the two spheres

$$\left.\begin{aligned} &\text{before the collision: (by (78) and (79))} \quad \begin{cases} u_1 = U \\ u_2 = \sqrt{v^2 + U^2\left(1 - \frac{v^2}{c^2}\right)} \end{cases} \\ &\text{after the collision: (by (81) and (82))} \quad \begin{cases} u_1 = U' \\ u_2 = \sqrt{v^2 + U'^2\left(1 - \frac{v^2}{c^2}\right)} \end{cases} \end{aligned}\right\} \quad (85)$$

Now, the first equation (84) requires that $m_2 = \overline{m}_2$. If mass changes at all with velocity then m_2 can equal $\overline{m}_2$ only if the corresponding velocity u_2 before and after the collision remains unaltered :

$$v^2 + U^2\left(1 - \frac{v^2}{c^2}\right) = v^2 + U'^2\left(1 - \frac{v^2}{c^2}\right).$$

From this, however, it follows that $U = U'$.

Moreover, equations (85) then show that u_1 also remains unaltered by the collision, and hence it follows that $m_1 = \overline{m}_1$.

Accordingly, the second equation (84) may be written as follows :

$$m_1 U = m_2 U\sqrt{1 - \frac{v^2}{c^2}} = m_1 U + m_2 U\sqrt{1 - \frac{v^2}{c^2}}$$

or

$$m_1 - m_2\sqrt{1 - \frac{v^2}{c^2}} = 0.$$

Hence we get

$$m_2 = \frac{m_1}{\sqrt{1 - \frac{v^2}{c^2}}} \quad . \quad . \quad . \quad (86)$$

Now let us suppose the velocity of throw U to be chosen smaller and smaller. Then by (85) we get finally that $u_1 = 0$, $u_2 = v$. So m_1 is the mass which corresponds to the velocity zero and which we call the *statical mass* m_0, whereas m_2 is the mass that corresponds to the velocity v, and we designate it briefly by m alone (without a suffix). Thus, the following holds :

$$m = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}} \quad . \quad . \quad . \quad (87)$$

In this way we have found *how the relativistic mass depends on the velocity.*

After this it is easy to see that through this relation (87) the general equation (86) is fulfilled for any velocity of throw U whatsoever. For, by (85), we have

$$m_1 = \frac{m_0}{\sqrt{1 - \frac{u_1^2}{c^2}}} = \frac{m_0}{\sqrt{1 - \frac{U^2}{c^2}}}$$

$$m_2 = \frac{m_0}{\sqrt{1 - \frac{u_2^2}{c^2}}} = \frac{m_0}{\sqrt{1 - \frac{1}{c^2}\left\{v^2 + U^2\left(1 - \frac{v^2}{c^2}\right)\right\}}}$$

$$= \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}\sqrt{1 - \frac{U^2}{c^2}}}$$

from which the relation (86) follows immediately.

As already mentioned we should arrive at exactly the same result if we were to consider the position from the other system of reference S′.

For the *convected momentum* of a body we get

$$J = mv = \frac{m_0 v}{\sqrt{1 - \frac{v^2}{c^2}}} \quad . \quad . \quad . \quad (88)$$

From this we can pass on to the law of motion for forces that act continuously. In doing so we must use the formulation of classical mechanics (II, 10, p. 33) which is based on convected momentum. It may clearly at once be applied to the new dynamics, but the law for the longitudinal and the transverse component must be formulated separately thus:

A force K produces a change in the convected momentum, and this is such that the change of the longitudinal or, respectively, transverse component of momentum per unit of time is equal to the corresponding component of the force.

The equations of motion may then easily be set up.

If we make a small longitudinal addition w_p to the velocity v, simple calculation * tells us that the longitudinal change in

*For if $u_p = v + w_p \qquad u_s = w_s$

are the components of velocity after the change, then the corresponding components of momentum are

$$J_p = \frac{m_0(v + w_p)}{\sqrt{1 - \frac{v^2}{c^2}}} \qquad J_s = \frac{m_0 w_s}{\sqrt{1 - \frac{v^2}{c^2}}}$$

in which

$$u = \sqrt{u_p^2 + u_s^2} = \sqrt{(v + w_p)^2 + w_s^2}$$

is the magnitude of the changed velocity. But we may set the latter approximately equal to the component u_p, for

$$u = \sqrt{v^2 + 2vw_p + w_p^2 + w_s^2}$$

and if we neglect the squares of w_p and w_s

$$u = \sqrt{v^2 + 2vw_p} = v\sqrt{1 + 2\frac{w_p}{v}}.$$

Next, we again apply the process used above (see Note on p. 182) to derive approximate formulæ. For small values of x

$$(1 + x)^2 = 1 + 2x + x^2 = 1 + 2x \text{ (approximately)}$$

and hence

$$\sqrt{1 + 2x} = 1 + x \text{ (approximately)}.$$

Thus, to a sufficient degree of approximation:

$$\frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{1}{\sqrt{1 - \frac{1}{c^2}(v^2 + 2vw_p)}} = \frac{1}{\alpha\sqrt{1 - \frac{2vw_p}{\alpha^2 c^2}}}$$

J is

$$\frac{m_0 w_p}{\left(\sqrt{1 - \frac{v^2}{c^2}}\right)^3}$$

but if we make a small transversal addition w_s to the v, then the transverse change of J becomes

$$\frac{m_0 w_s}{\sqrt{1 - \frac{v^2}{c^2}}}.$$

These expressions are to be divided by the small time t during which the change takes place: we thus get the *components of acceleration:*

$$b_p = \frac{w_p}{t} \qquad b_s = \frac{w_s}{t}$$

in which we have used the abbreviation α introduced earlier (VI, 2, formula (71), p. 199). According to the approximate formulæ used earlier (Note on p. 182) we have

$$\frac{1}{\sqrt{1 - x}} = 1 + \tfrac{1}{2}x$$

and thus we get

$$\frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{1}{\alpha}\left(1 + \frac{vw_p}{\alpha^2 c^2}\right).$$

Now we get for the components of momentum or impulse after the change, if we neglect the quadratic members in w_p and w_s:

$$J_p = m_0(v + w_p)\frac{1}{\alpha}\left(1 + \frac{vw_p}{\alpha^2 c^2}\right) = \frac{m_0}{\alpha}\left\{v + w_p\left(1 + \frac{v^2}{\alpha^2 c^2}\right)\right\}$$

$$= \frac{m_0}{\alpha}\left(v + \frac{w_p}{\alpha^2}\right)$$

and

$$J_s = m_0 w_s \frac{1}{\alpha}\left(1 + \frac{vw_p}{\alpha^2 c^2}\right)$$

$$= \frac{m_0}{\alpha} w_s.$$

From these we must subtract the original momenta

$$J^0_p = \frac{m_0 v}{\alpha} \qquad J^0_s = 0$$

and we get for the changes of momentum

$$J_p - J^0_p = \frac{m_0 w_p}{\alpha^2} \qquad J_s - J^0_s = \frac{m_0 w_s}{\alpha}$$

which agree with the formulæ in the text.

and for the *components of force* we get the expressions

$$K_p = \frac{m_0 b_p}{\left(\sqrt{1 - \frac{v^2}{c^2}}\right)^3} \qquad K_s = \frac{m_0 b_s}{\sqrt{1 - \frac{v^2}{c^2}}} \quad . \quad (89)$$

The relationship between the force and the acceleration generated is thus different according as the force acts in the direction of the acceleration that is already present or in a direction perpendicular to this.

It is usual to bring these formulæ into a form in which they resemble the fundamental law of classical dynamics (II, 10, formula (10), p. 33) as much as possible. For this purpose we set

$$m_p = \frac{m_0}{\left(\sqrt{1 - \frac{v^2}{c^2}}\right)^3} \qquad m_s = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}} \quad . \quad (90)$$

and we call these quantities *longitudinal* and *transverse mass.* The latter is identical with the quantity m, simply called *relativistic mass* above in formula (87).

Then we may write in place of (89)

$$K_p = m_p b_p \qquad K_s = m_s b_s \quad . \quad . \quad (91)$$

which agrees in form with the fundamental classical law.

We see how necessary it is to define the conception of mass from the very beginning exclusively by the inertial resistance. Otherwise it would not be possible to apply it in relativistic mechanics, since a different expression "mass" comes into account for the convected momentum, for the longitudinal and the transverse force, and, moreover, these masses are not characteristic constants of the body, but depend on its velocity. Thus the conception of mass in Einstein's dynamics differs very widely from that to which we are accustomed, and in which mass denotes quantity of matter in some way. In a certain sense the statical mass m_0 is a measure of the Einsteinian mass, but again, unlike the mass of ordinary mechanics it is not, in an arbitrary system of reference, equal to the ratio of momentum to velocity or of force to acceleration.

A glance at the formulæ (87) and (90) for the mass tells us that the values of the relativistic mass m (m_p or m_s respectively) become greater the more the velocity v of the moving body approaches the velocity of light. For $v = c$ the mass becomes infinitely great.

From this it follows that it is impossible to make a body move with a velocity greater than that of light by applying finite forces: its inertial resistance grows to an infinite extent and prevents the velocity of light from being reached.

Here we see how Einstein's theory becomes rounded off to a harmonious whole. The assumption that there is a limiting velocity that cannot be exceeded, which seems almost paradoxical, is itself required by the physical laws in their new form.

Formula (87) giving the dependence of the mass on the velocity is the same as that already found by Lorentz from electrodynamic calculations for his flattened electron. In it m_0 was expressed in terms of the electrostatic energy U of the stationary electron just as in Abraham's theory (V, 13, p. 177, formula (69)) namely, by

$$m_0 = \frac{4}{3}\frac{U}{c^2}.$$

We now see that Lorentz' formula for the mass has a much more general significance than is at first apparent. It must hold for every kind of mass, no matter whether it is of electrodynamic origin or not.

Recent researches into the deflections of cathode rays seem to indicate that Lorentz' formula is more correct than Abraham's. A surprising confirmation of the relativistic formula for the mass has been obtained in a branch of physics which seems to be quite foreign to the theory of relativity, namely, in the *spectroscopy* of optical and Röntgen rays.

We cannot do more than just touch on these wonderful relationships. The luminescence of atoms comes about through electrons within the atomic configuration executing oscillatory motions and producing electromagnetic waves which are propagated in all directions. The older theory carried out the calculations of these phenomena with the help of Maxwell's field equations, but latterly it has been found necessary to give up the strict validity of these equations in the interior of the atom and to assume other laws, which were introduced by Max Planck (in 1900) for the first time in the theory of heat radiation. The latter constitute the so-called *quantum theory*. Niels Bohr (in 1913) applied it to explain spectra and achieved great success. Without entering into details we remark only that in rapid motions of the electrons the mass must be increased according to the theory of relativity, and this will exert an influence on the spectra. Sommerfeld (in 1915) was actually able to show that in consequence of the variability of mass the spectral lines have a complicated structure. Each line in reality consists of a whole system of intense lines with neighbouring fine lines. In the visible spectra, which are emitted by the outer electrons of the atom, this group of lines is very narrow, it constitutes a "*fine structure.*" But in the case of Röntgen spectra that come from the interior of the atom it is a coarse structure in which the resolution of the lines is magnified millions of times. The fine structure calculated by Sommerfeld for the lines of the hydrogen and the helium spectrum has been observed by Paschen (1916). In the case of the Röntgen spectra, too, these hypotheses of Sommerfeld have proved trustworthy. They are so exact that they allow us to discriminate between the formulæ for the mass given by Abraham and Lorentz, which is a quantity of the second order in β. Sommerfeld's pupil, Glitscher, was able to show, in 1917, that Abraham's formula is not compatible with observations of the helium spectrum, but that Lorentz' formula is in agreement with them. We are thus justified in speaking of a *spectroscopic confirmation* of Einstein's theory of relativity.

Since, by formula (87), every mass depends on the velocity, the proof of the electromagnetic nature of the mass of the electron lapses, and with it the relationship between statical mass and electrostatic energy. With Lorentz' theory of the stationary ether it was possible to try to trace inertial mass back to the peculiar property of persistence of the electromagnetic field. If Einstein's theory of relativity had had to give up this great plan everyone who appreciates uniformity would have regarded it as a serious defect. The new dynamics has not failed here, however, but has allowed us to penetrate into the deepest recesses of the nature of inertial mass.

8. The Inertia of Energy

For all practical purposes, and also for the case of the most rapid electrons, it is sufficient to write down the formula (87) for the mass as far as members no higher than the second order. Now, as we have seen (p. 182, footnote), this approximation gives us

$$\frac{1}{\sqrt{1-\beta^2}} = 1 + \frac{1}{2}\beta^2.$$

Hence we get

$$m = m_0\left(1 + \frac{1}{2}\frac{v^2}{c^2}\right).$$

In ordinary mechanics the kinetic energy (II, 14, p. 44) is defined by $T = m_0 \frac{v^2}{2}$. From our formula the expression that follows for it is

$$T = c^2 (m - m_0).$$

It can be shown that this definition of kinetic energy is rigorously valid even if the members of order higher than the second are not neglected.

The energy law (II, 14, formula (16), p. 44) demands that the time-change (that is, change with respect to the time) of energy $E = T + U$ be equal to zero during the whole of the motion. In the latter expression the classical value $T = \frac{m}{2}v^2$ must be replaced by the relativistic value

$$T = c^2(m - m_0) = c^2 m_0 \left(\frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} - 1 \right) \quad . \quad (92)$$

If we form the time-change of this we get by a calculation similar to that made above (footnote, p. 226) for longitudinal acceleration * :

$$\text{Time-change of } T = \frac{m_0 v b_p}{\left(\sqrt{1 - \frac{v^2}{c^2}}\right)^3} = K_p v \quad . \quad (93)$$

in which the longitudinal component of force has been introduced in accordance with (89), p. 228. But the right-hand side is the negative time-change of the potential energy U. For during a sufficiently small interval of time t the force may be regarded as approximately constant, and we may calculate as if we were dealing with a gravitational force whose potential energy (II, 14, formula (15), p. 42) is equal to Gx; we took

* It was shown there that if the velocity v is replaced by the changed value of u with the components $u_p = v + w_p$, $u_s = w_s$, the expression

$$\frac{1}{a} = \frac{1}{\sqrt{1 - \frac{u^2}{c^2}}} \text{ passes over into } \frac{1}{\sqrt{1 - \frac{u^2}{c^2}}} = \frac{1}{a}\left(1 + \frac{v w_p}{a^2 c^2}\right).$$

Thus its change is

$$\frac{1}{\sqrt{1 - \frac{u^2}{c^2}}} - \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{v w_p}{a^3 c^2}$$

and the formula in the text follows at once from this.

the direction of x as opposite to that of gravity, so that we had to set $G = -K_p$. The time-change of the potential energy then becomes

$$G\frac{x}{t} = -K_p v.$$

Accordingly, equation (93) actually expresses that the quantity $E = T + U$ is constant in time, T denoting the expression (92).

If we write the formula (92) thus

$$m = m_0 + \frac{T}{c^2}$$

it states that the mass differs from its value at rest by the kinetic energy divided by the square of the velocity of light.

This formulation suggests to us that the statical mass m_0 is related in the same way to the energy content of the resting body, or that the universal relationship

$$m = \frac{E}{c^2} \quad . \quad . \quad . \quad . \quad (94)$$

holds between mass and energy in all cases. Einstein has called this *law of the inertia of energy* the most important result of the theory of relativity. For it signifies that the two fundamental conceptions of mass and energy are identical and thus gives us a clearer vision of the structure of matter. Before dealing with this we shall give Einstein's simple proof of formula (94).

This is based on the fact that radiation exerts a pressure. From Maxwell's field equations, supplemented by a theorem first deduced by Poynting (1884), it follows that a light-wave which falls on an absorbing body exerts a pressure on it. And it is found that the momentum which is imposed on the absorbing surface by a short flash of light is equal to $\frac{E}{c}$. This result was experimentally confirmed by Lebedew (1890) and later again with great accuracy by Nichols and Hull (1901). Exactly the same pressure is experienced by a body which emits light, just as a gun experiences a recoil when a shot is fired.

We next imagine a hollow body, say a long tube, and at the ends of it two exactly equal bodies A and B of the same material which, according to the ordinary ideas, have the same mass (Fig. 123). But the body A is to have an excess of energy E over that of B, say in the form of heat, and there is to be an arrangement (hollow mirror or something similar) by which this energy E can be sent in the form of radiation to B. Let the spatial extent of this flash of light be small compared with the length l of the tube (Fig. 123).

Then A experiences the recoil $\frac{E}{c}$. Thus the whole tube, whose total mass we take as M, acquires a velocity v directed backwards and determined by the equation of momentum

$$Mv = \frac{E}{c}.$$

This motion continues until the flash arrives at B and is there absorbed. Then B experiences the same blow forwards and hence the whole system comes to rest. The displacement which it undergoes during the time of travel t of the flash is $x = vt$, where v is to be taken from the above equation, thus

FIG. 123.

$$x = \frac{Et}{Mc}.$$

The time of travel, however, is determined (except for a small error of higher order) by $l = ct$. Hence the displacement is

$$x = \frac{El}{Mc^2}.$$

Now the bodies A and B may be exchanged and this may be done without using external influences. Let us suppose that two men are situated in the tube, who put A in the place of B and B in the place of A and then themselves return to their original positions. According to ordinary mechanics the tube as a whole must suffer no displacement, for changes of position can be effected only by external forces.

If this exchange were to be carried out, everything in the interior of the tube would be as at the beginning, the energy E would again be at the same place as before, and the distribution of mass would be exactly the same. But the whole tube would be displaced a distance x with respect to its initial position by the light-impulse. This, of course, contradicts all the fundamental canons of mechanics. We could repeat the process and thus impart any arbitrary change of position to the system without applying external forces. This is, however, an impossibility. The only escape from the difficulty is to assume that when the bodies A and B are exchanged these two bodies are not mechanically equivalent but that B has a mass greater by m than A in consequence of its excess of energy E. But then the symmetry during the exchange is not maintained, and the mass m is then displaced from right to left by a distance l. At the same time the whole tube is displaced a distance x in the reverse direction. This distance is determined by the circumstance that the process occurs without the intervention of external influences. The total momentum, consisting of that of the tube $M\frac{x}{t}$ and of that of the transported mass $-m\frac{l}{t}$ is thus zero.

$$Mx - ml = 0$$

from which it follows that

$$x = \frac{ml}{M}.$$

Now this displacement must exactly counterbalance that produced by the light-impulse, hence we must have

$$x = \frac{ml}{M} = \frac{El}{Mc^2}.$$

This allows us to calculate m and we get

$$m = \frac{E}{c^2}.$$

This is the amount of inertial mass that must be ascribed to the energy E in order that the maxim of mechanics remain valid which states that no changes of position can occur without the action of external forces.

Since every form of energy is finally transformable into radiation by some process or other, this law must be universally valid. Accordingly *mass is nothing other than a form of appearance of energy*. Matter itself loses its primary character as an indestructible substance and is nothing more than points of concentrated energy. Wherever electric and magnetic fields or other effects lead to intense accumulations of energy the phenomenon of inertial mass presents itself. Electrons and atoms are examples of such places at which there are enormous concentrations of energy.

We can touch on only a few of the numerous important consequences of this theorem.

Concerning firstly the mass of the electron, formula (69),

p. 177, shows that for the statical mass $m_0 = \frac{4}{3}\frac{U}{c^2}$ the electrostatic energy U cannot be the total energy E of the electron at rest. There must be present another part of energy V, $E = U + V$, such that

$$m_0 = \frac{4}{3}\frac{U}{c^2} = \frac{E}{c^2} = \frac{U + V}{c^2}.$$

From this it follows that $V = \frac{4}{3}U - U = \frac{1}{3}U = \frac{1}{4}E$. Thus the total energy is three-quarters electrostatic and one-quarter of a different kind. This part must be due to the cohesive forces which hold the electron together by counterbalancing the electrostatic attraction. Ingenious theories on this point have already been developed by Mie, Hilbert, and Einstein, but the results are not yet sufficiently satisfactory to enable us to describe them. The hypotheses of Einstein seem to be the most promising, and we shall return to consider them briefly when we come to deal with the general theory of relativity.

On the other hand, the law of the inertia of energy is even now of very great importance for the problem of the structure of the material atom.

We have already mentioned (p. 171) that every atom consists of a positive part, which is indissolubly connected with the inertial mass, and of a number of negative electrons. Experiments by Rutherford (1913) and his collaborators on the scattering of the positive rays, the so-called α-rays, emitted by radioactive substances, proved that the positive constituents of atoms, which are nowadays called " nuclei " or " protons," are extraordinarily small, indeed much smaller than electrons, whose radii (p. 178) have been estimated at $2 . 10^{13}$ cms. Now if the mass of the nucleus, like that of the electron, is for the main part (three-quarters) electromagnetic by nature, then there should be between it and the radius a a formula similar to that applied earlier (p. 178) to electrons, namely, $m_0 = \frac{2}{3}\frac{e^2}{ac^2}$, but perhaps with a different numerical factor. Thus the masses would be inversely proportional to the radii :

$$\frac{\text{Radius of the electron}}{\text{Radius of the nucleus}} = \frac{\text{Mass of the nucleus}}{\text{Mass of the electron}}.$$

But we know that the hydrogen atom is 2000 times smaller than the electron. From this it follows that the radius of the hydrogen nucleus is about 2000 times smaller than that of the electron, and this is in good agreement with the results.

Thus we can apply the law of the inertia of energy successfully to the masses of atoms or nuclei.

Radioactive atoms are catalogued, as we know, according to the type of rays they emit. 1. α-rays. These are positively charged particles which have been shown to be helium nuclei. 2. β-rays, which are electrons. 3. γ-rays, which are electromagnetic rays of the nature of Röntgen rays. In these processes of emission the atom loses not only direct mass, but also energy to a considerable amount. But according to the law of the inertia of energy the loss of energy also entails a loss of mass. Unfortunately this is so small that it has not yet been possible to determine it experimentally.

Fundamentally, however, the discovery that when an atom is disintegrated the sum of the masses of the component parts is not equal to the mass of the original atom is of great importance. It has long been an object of research to resolve all atoms into simpler primary constituents. Prout (1815) set up the hypothesis that these primary constituents are the *hydrogen atoms.* He supported this idea by pointing out that the weights of many atoms are whole multiples of the weight of the hydrogen atom. But exact measurements of the atomic weights have not confirmed this assertion and this has served to discountenance Prout's hypothesis. But nowadays it has been taken up again with success. For according to the law of the inertia of energy the mass of an atomic nucleus composed of n hydrogen nuclei is not simply equal to n times the mass of the hydrogen nucleus, but differs from it by an amount of energy necessary to combine these nuclei. Recently this view has received strong support through the discovery by Rutherford (1919) that hydrogen nuclei can be forcibly detached from nitrogen nuclei by a bombardment with α-rays. It is true that the law of the inertia of energy can account for only small deviations of the ratio of the atomic weights from integral numbers. But there is still another cause which produces the great differences, the fact of *isotopes.* Many elements are mixtures of atoms having equally charged nuclei and a similar arrangement of electrons but they have different nuclear mass. These cannot be separated chemically, although this has been carried out physically. The existence of isotopes was first proved in the case of radio-active substances, and recently by Aston (1920) in the case of many other elements. But we cannot here enter into this interesting subject.

This survey of the problem of modern atomic theory shows us very distinctly that Einstein's theory of relativity is no product of fantastic speculation, but rather a guiding thread in the most important region of physical research. The unveiling of the secret of the world of atoms is an aim which exerts a directive influence on the mental growth of humanity, and it exceeds in grandeur and importance all other problems of natural science, perhaps even the problem of the structure of the universe. For every step towards this goal gives us not only new weapons in the struggle for existence but furnishes us with knowledge of the most intimate relationships of the natural world, and teaches us to distinguish between the deception of the senses and the truth of the eternal laws of the universe.

9. The Optics of Moving Bodies

Now that we have drawn the most important inferences from our modified mechanics it is time to return to those problems from which Einstein's theory of relativity emanated, namely, the electrodynamics and optics of moving bodies. The fundamental laws of these regions of physics are condensed in Maxwell's field equations, and even Lorentz had recognized that these are invariant for empty space ($\epsilon = 1, \mu = 1, \sigma = 0$) with respect to Lorentz transformations. The exact invariant field equations for moving bodies have been set up by Minkowski (1907). They differ from the Lorentz formulæ of the theory of electrons only in minor terms which cannot be tested by observation, but have in common with these the partial convection of the dielectric polarization, and hence account in full agreement with observation for all electromagnetic and optical phenomena involving moving bodies. We recall, in particular, the experiments of Röntgen, Eichenwald, and Wilson (V, 11, p. 166), yet we shall not discuss them further as this would require elaborate mathematical calculation. But the optics of moving bodies may be treated in quite an elementary way and we shall describe them here as one of the most beautiful applications of Einstein's theory.

According to Einstein's theory of relativity there is no ether but only bodies moving relatively to each other, and so it is self-evident that all optical phenomena in which the source of light, the substances traversed by radiation, and the observer are at rest in one and the same inertial system are the same for all inertial systems. Thus, this also explains the Michelson-Morley experiment, which gave rise to the theory. The question now is merely to determine whether the phenomena which occur when the source of light, the medium traversed by radiation, and the observer are in relative motion are correctly represented by the theory.

Let us imagine a light wave in a material body which is at rest in the system of reference S. Let its velocity be $c_1 = \frac{c}{n}$ (n being the index of refraction), its vibration number ν, and its direction relative to the system S definitely fixed. We enquire as to how these three characteristics of the wave are judged by an observer who is at rest in a system of reference S′ which is moving with the velocity v parallel to the x-direction of the system S.

We treat this question according to the same method as that which we applied earlier (IV, 7, p. 103) except that now we use Lorentz transformations as our basis of reasoning in place of Galilei transformations. We showed there that the wave-number

$$\nu\left(t - \frac{s}{c}\right)$$

is an invariant. For it denotes the number of waves which have left the zero point or origin after the moment $t = 0$ and up to the moment t have reached the point P, during which time they advance a distance s (Fig. 69, p. 103). This invariance now holds, of course, for Lorentz transformations.

We next consider waves that advance parallel to the x-direction. Then the x-co-ordinate of the point P must be inserted for s, and we get

$$\nu\left(t - \frac{x}{c_1}\right) = \nu'\left(t' - \frac{x'}{c'_1}\right)$$

where ν, ν', and c_1, c'_1 are the frequencies and velocities of the wave relative to the systems S and S′. If on the right we insert the expressions for x' and t' given by the Lorentz transformation (72), p. 200, we get,

$$\nu\left(t - \frac{x}{c_1}\right) = \frac{\nu'}{\alpha}\left(t - \frac{v}{c^2}x - \frac{x - vt}{c'_1}\right)$$

where $\alpha = \sqrt{1 - \beta^2} = \sqrt{1 - \frac{v^2}{c^2}}$. If we now first set $x = 1$, $t = 0$, and then $t = 1$, $x = 0$, we get

$$\left.\begin{aligned} \frac{\nu}{c_1} &= \frac{\nu'}{\alpha}\left(\frac{v}{c^2} + \frac{1}{c'_1}\right) \\ \nu &= \frac{\nu'}{\alpha}\left(1 + \frac{v}{c'_1}\right) \end{aligned}\right\} \quad . \quad . \quad . \quad (95)$$

If we divide the second equation by the first we get

$$c_1 = \frac{1 + \frac{v}{c'_1}}{\frac{v}{c^2} + \frac{1}{c'_1}} = \frac{c'_1 + v}{1 + \frac{vc'_1}{c^2}}$$

If, conversely, we solve for c'_1 we get the *strict convection formula*

$$c'_1 = \frac{c_1 - v}{1 - \frac{vc_1}{c^2}}.$$

This agrees exactly with Einstein's addition theorem of velocities for longitudinal motion [first formula (77), p. 220], if we replace u_p in it by c_1, and u'_p by c'_1. The same rule which holds for calculating the velocities of material bodies relative to various systems of reference may also be applied to the velocity of light.

If members of order higher than the second in $\beta = \frac{v}{c}$ be neglected, the law, however, becomes identical with Fresnel's convection formula (43), p. 117. For, with this approximation, we may write

$$\frac{1}{1 - \frac{vc_1}{c^2}} = \frac{1}{1 - \frac{\beta}{n}} = 1 + \frac{\beta}{n} = 1 + \frac{v}{nc}.$$

Thus

$$c'_1 = (c_1 - v)\left(1 + \frac{v}{nc}\right)$$

$$= c_1 - v + \frac{vc_1}{nc} - \frac{v^2}{nc},$$

and if we omit the last term of the second order and set $\frac{c_1}{c} = \frac{1}{n}$, we get

$$c'_1 = c_1 - v\left(1 - \frac{1}{n^2}\right).$$

This is precisely *Fresnel's convection formula.*

The second of the formula (95) represents Doppler's principle. This is usually applied to a vacuum, so that $c_1 = c$; then, as we know, it follows from the addition theorem of velocities (p. 220) that $c'_1 = c$. And the second of formulæ (95) gives us

$$\nu' = \nu\frac{\alpha}{1 + \frac{v}{c}} = \nu\frac{\sqrt{1 - \beta^2}}{1 + \beta}.$$

But now $1 - \beta^2 = (1 - \beta)(1 + \beta)$, hence we may write

$$\nu' = \nu\frac{\sqrt{(1 - \beta)(1 + \beta)}}{1 + \beta} = \nu\sqrt{\frac{1 - \beta}{1 + \beta}}.$$

Thus the *strict formula for the Doppler effect* assumes the symmetrical form

$$\nu'\sqrt{1 + \frac{v}{c}} = \nu\sqrt{1 - \frac{v}{c}} \quad . \quad . \quad . \quad (96)$$

which expresses in an evident way the equivalence of the systems of reference S and S'. If we neglect the terms of order higher than the second, we may replace $\sqrt{1 + \beta}$ by $1 + \frac{1}{2}\beta$, and $\sqrt{1 - \beta}$ by $1 - \frac{1}{2}\beta$. Accordingly we get

$$\nu'(1 + \tfrac{1}{2}\beta) = \nu(1 - \tfrac{1}{2}\beta)$$

$$\nu' = \nu\frac{1 - \frac{1}{2}\beta}{1 + \frac{1}{2}\beta}.$$

Now, to the same degree of accuracy,

$$\frac{1}{1 + \frac{1}{2}\beta} = 1 - \tfrac{1}{2}\beta,$$

and thus

$$\nu' = \nu(1 - \tfrac{1}{2}\beta)^2 = \nu(1 - \beta + \tfrac{1}{4}\beta^2)$$

and if we neglect β^2

$$\nu' = \nu\left(1 - \frac{v}{c}\right)$$

in complete agreement with formula (40), p. 107.

To derive the *aberration* of light by the same method we must consider a train of waves which propagates itself perpendicular to the direction of motion x of the systems S and S' with respect to each other. We must state whether the perpendicular direction with respect to S or S' is meant, for if the ray is perpendicular to the x-axis relative to S, it is not so relative to S'. Then we must set $s' = y'$, and we get for a vacuum ($c_1 = c_1' = c$)

$$\nu\left(t - \frac{s}{c}\right) = \nu'\left(t' - \frac{y'}{c}\right).$$

If we here insert the values given by the Lorentz transformation (72) p. 200, we get

$$\nu\left(t - \frac{s}{c}\right) = \nu'\left(\frac{t - \frac{v}{c^2}x}{\alpha} - \frac{y}{c}\right).$$

From this it follows, firstly, for $x = 0$, $y = 0$, $s = 0$, $t = 1$,

$$\nu = \frac{\nu'}{\alpha},$$

and then, for $t = 0$,

$$\frac{\nu}{c}s = \nu'\left(\frac{vx}{c^2\alpha} + \frac{y}{c}\right) = \frac{\nu'}{\alpha c}(\beta x + \alpha y).$$

Thus

$$s = \alpha y + \beta x.$$

If the wave-plane relative to the system of reference S were perpendicular to the y-axis, then we should have $s = y$. Since this is not the case, it must be deflected (Fig. 124). Suppose x, y are the co-ordinates of any point P in the wave-plane. If, in particular, we choose for P the point of intersection A with the x-axis, we must set $x = a$, $y = 0$, and thus $s = \beta a$. In the same way we must set $x = 0$, $y = b$ for the point of intersection B with the y-axis, thus $s = \alpha b$.

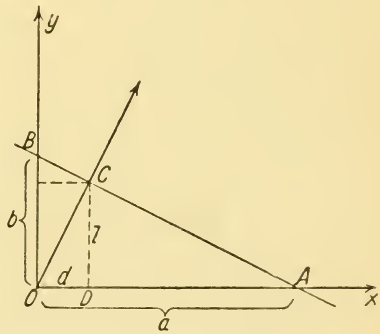


FIG. 124.

Hence we get

$$s = \beta a = \alpha b \quad \text{or} \quad \frac{b}{a} = \frac{\beta}{\alpha}.$$

This ratio $\frac{b}{a}$ is clearly a measure of the deflection of the wave-front. It is easy to see that it agrees with the elementary definition of the aberration constant according to the emission theory (IV, 3, p. 82). For the perpendicular OC dropped from the origin on to the wave-plane is the direction of propagation. If D is the projection of C on the x-axis, then OD $= d$ is the displacement to which a telescope of length DC $= l$

placed parallel to the y-axis must be subject during the time required by the light to traverse the tube, in order that a ray that strikes the middle of the object glass at C should reach exactly the middle of the eye-piece at O. Thus $\frac{d}{l}$ is the aberration constant. From the similarity of the triangles OCD, BAO we get the proportion

$$\frac{d}{l} = \frac{b}{a} = \frac{\beta}{\alpha} = \frac{\beta}{\sqrt{1 - \beta^2}}.$$

This is the exact *aberration formula.* If we neglect β^2 in comparison with 1, it simplifies into the elementary formula

$$\frac{d}{l} = \beta = \frac{v}{c}.$$

This result is particularly remarkable because all the ether theories have considerable difficulties to overcome in explaining aberration. In making use of the Galilei transformation we obtain no deflection at all of the wave-plane and the wave-direction (IV, 10, p. 121), and to explain aberration we must introduce the conception " ray " which need not agree in moving systems with the direction of propagation. In Einstein's theory this is not necessary. In every inertial system S the direction of the ray, that is, the direction along which the energy is transported coincides with the perpendicular on the wave-planes, nevertheless the aberration comes out in the same simple way as the Doppler effect and Fresnel's convection coefficient from the conception of a wave with the help of the Lorentz transformation.

This method of deriving the fundamental laws of the optics of moving bodies shows very strikingly the superiority of Einstein's theory of relativity above all other theories.

10. Minkowski's Absolute World

The essence of the new kinematics consists in the inseparability of space and time. The world is a four-dimensional manifold, its element is the world-point. Space and time are forms of arrangement of the world-points, and this arrangement is, to a certain extent, arbitrary. Minkowski has expressed this view in the words : " From now onwards space and time are to sink to shadows and only a sort of union of both retain self-dependence." And he has worked out this idea logically by developing kinematics as four-dimensional geometry. We have made use of his method of description throughout, omitting the y- and z-axis only for the sake of simplicity and working in the xt-plane. If we throw a glance at the geometry in the xt-plane from the mathematical point of view, we see that we are not dealing with ordinary Euclidean geometry. For in this all straight lines that radiate out from the origin are equivalent, the unit of length on them is the same, and the calibration curve is thus a circle (Fig. 125). But in the xt-plane the space-like and the time-like straight lines are not equivalent. There is a different unit of length on each, and the calibration curve consists of the hyperbolæ

$$G = x^2 - c^2t^2 = \pm 1.$$

In Euclidean geometry we can construct an infinite number of rectangular co-ordinate systems with the same origin O, which emerge from each other by rotation. In the xt-plane there are likewise an infinite number of equivalent co-ordinate systems, for which the one axis can be chosen at will within a certain angular region.

In Euclidean geometry the distance s of a point P with the co-ordinates x, y from the origin is an invariant with respect to rotations of the co-ordinate system (see III, 7, formula (28), p. 64). By Pythagoras' theorem we have (Fig. 125) in the xy-system

$$s^2 = x^2 + y^2$$

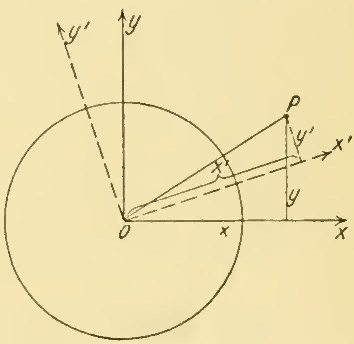


FIG. 125.

and in any system $x'y'$ we likewise have $s^2 = x'^2 + y'^2$. The calibration curve, the circle of radius unity, is represented by $s = 1$. Hence we shall regard s, or s^2, as the *ground invariant* of Euclidean geometry.

In the xt-plane the ground invariant is

$$G = x^2 - c^2t^2$$

and the calibration curve is

$$G = \pm 1.$$

Now Minkowski observed that a parallelism presents itself here which throws a light on the mathematical structure of the four-dimensional world (or the xt-plane, respectively). For if we set $- c^2t^2 = u^2$, we clearly get

$$G = x^2 + u^2 = s^2$$

and this may be regarded as the ground invariant s^2 of a Euclidean geometry having the rectangular co-ordinates x, u.

It is true that we cannot extract the square root of the negative quantity $-c^2t^2$, and u itself cannot be calculated from the time t. But mathematicians have long been accustomed to overcoming such difficulties by making a bold stroke. The "imaginary" quantity $\sqrt{-1} \equiv i$ has been firmly established in mathematics since the time of Gauss. We cannot here enter into the question of the basis on which the doctrine of imaginary numbers is strictly founded. These numbers are essentially no more "imaginary" than a fraction such as $\frac{2}{3}$, for numbers "with which we number things or count" properly comprise only the natural integral numbers, 1, 2, 3, 4, . . . And 2 is not divisible by 3 so that $\frac{2}{3}$ is an operation that can be carried out just as little as $\sqrt{-1}$. Fractions such as $\frac{2}{3}$ signify an extension of the natural conception of numbers, and they have become familiar through our school teaching and custom, and excite no feeling of strangeness. A similar extension of the conception of number is given by imaginary numbers, which are just as little unusual to mathematicians as fractions. All formulæ that contain imaginary numbers have just as definite a meaning as those formed from ordinary "real" numbers, and the inferences drawn from them are just as convincing.

If we here use the symbol $\sqrt{-1} \equiv i$, we may write

$$u = ict.$$

The non-Euclidean geometry of the xt-plane is thus formally identical with Euclidean geometry in the xu-plane, and u-values correspond only to real times t.

This theorem is of inestimable advantage for the mathematical treatment of the theory of relativity. For in the case of numerous operations and calculations the real issue concerns, not the reality of the quantities considered, but the algebraic relations that exist between them and that hold just as well for imaginary numbers as for real numbers. Hence we can apply the laws known from Euclidean geometry to the four-dimensional world. Minkowski replaces

$$x \quad y \quad z \quad ict$$

by

$$x \quad y \quad z \quad u$$

and then operates with these four co-ordinates in a fully symmetrical way. The ground-invariant then clearly becomes

$$G = s^2 = x^2 + y^2 + z^2 + u^2.$$

The special position occupied by the time thus vanishes from all formulæ, and this to a very considerable degree facilitates the calculations and allows us to survey them readily as a whole. In the final result we have again to replace u by ict, and then only such equations retain a physical meaning as are formed exclusively from real numbers.

Non-mathematicians will not understand much of these arguments and will, perhaps, become indignant at the "mystic equation" $3 \cdot 10^{10}$ cms. $= \sqrt{-1}$ sec., formed half in jest by Minkowski, and will be inclined to support the critics of the theory of relativity, to whom the equivalence of time with the spatial dimensions appears sheer nonsense.

We hope that our method of representation in which the formal method of Minkowski appears only at the conclusion will be able to nullify such objections. In the xt-plane t is clearly by no means interchangeable with the dimension of length x. The light-lines ξ and η are the insuperable barriers between the time-like and the space-like world-lines. Thus Minkowski's transformation $u = ict$ is to be valued only as a mathematical artifice which puts in the right light certain formal analogies between the space-co-ordinates and the time, without, however, allowing them to be interchanged

But this artifice has led to important disclosures. Without it Einstein's general theory of relativity cannot be imagined. The important point in it is the analogy between the ground-invariant G and the square of a distance. In future we shall call

$$s = \sqrt{G} = \sqrt{x^2 + y^2 + z^2 + u^2} = \sqrt{x^2 + y^2 + z^2 - c^2t^2}$$

the "four-dimensional distance," but we must remember that the expression is used in an applied sense.

After our earlier discussion of the invariant G the real meaning of the quantity s is easy to understand. Let us confine our attention to the xt-plane, then

$$s = \sqrt{G} = \sqrt{x^2 + u^2} = \sqrt{x^2 - c^2t^2}.$$

Now for every space-like world-line G is positive, and thus s, as the square root of a positive number, is a real quantity. We can then make the world-point x, t simultaneous with the origin by choosing a suitable system of reference S. We then have $t = 0$, and $s = \sqrt{x^2} = x$ as the spatial distance of the world-point from the origin.

For every time-like world-line G is negative, and hence s is imaginary. Then there is a co-ordinate system in which $x = 0$, and hence $s = \sqrt{-c^2t^2} = ict$. Thus, in any case s

has a simple meaning and is to be regarded as a measurable quantity.

We here close our account of Einstein's special theory of relativity. Its results may be condensed into the following statements :

Not only the laws of mechanics but those of all physical events, in particular of electromagnetic phenomena, are completely identical in an infinite number of systems of reference which are moving uniformly with translation relatively to each other and which are called inertial systems. In each of these systems a particular measure holds for the lengths and the times, and these measures are connected with each other by the Lorentz transformations.

Systems of reference which move with acceleration relatively to each other are no more than in mechanics identical with inertial systems. If we refer physical laws to such accelerated systems, they become different. In mechanics centrifugal forces manifest themselves, and in electrodynamics there are analogous effects, the study of which would take us too far. Thus Einstein's special theory of relativity does *not* do away with Newton's absolute space in the restricted sense which we attached to this expression earlier (III, 6, p. 61). In a certain sense it puts *the whole of physics*, including electrodynamics, into the same state as that in which mechanics has been since the time of Newton. The far-reaching questions of absolute space which troubled us there are not yet solved. We have scarcely advanced a step further, indeed, though by extending the physical complex beyond the vista of mechanics the problem has become considerably more difficult.

We shall now see how Einstein has overcome these obstacles.

CHAPTER VII

EINSTEIN'S THEORY OF RELATIVITY

1. Relativity in the Case of Arbitrary Motions

In dealing with classical mechanics we discussed in detail the reasons that led Newton to set up the conceptions of absolute space and absolute time. But at the same time we emphasized the objections which can be raised against these abstractions from the point of view of the theory of knowledge.

Newton supported his assumption of absolute space on the existence of inertial resistances and centrifugal forces. It is clear that these cannot depend on inter-actions between bodies since they occur in the same way, independently of the local distribution of masses, in the whole universe as far as observation can reach. Hence Newton concludes that they depend on *absolute* accelerations. In this way absolute space is introduced as the fictitious cause of physical phenomena.

The unsatisfactory features of this theory may be recognized from the following example.

Suppose two fluid bodies S_1 and S_2 of the same material and size to be present in astronomic space and at such a distance from each other that ordinary gravitational effects of the one on the other are inappreciably small (Fig. 126). Each of these bodies is to be in equilibrium under the action of the gravitation of its parts on each other and the remaining physical forces, so that no relative motions of its parts with respect to each other occur. But the two bodies are to execute a relative motion of rotation with constant velocity about the line connecting their middle points. This signifies that an observer on the one body S_1 notes a uniform rotation of the other body S_2 with regard to his own point of vantage, and vice versa. Now each of these bodies is to be measured by observers that are at rest with respect to each other. Suppose it is found that S_1 is a sphere and S_2 is a flattened ellipsoid of rotation.

Newtonian mechanics would infer from the different behaviour of the two bodies that S_1 is at rest in absolute space, but that S_2 executes an absolute rotation. The flattening of S_2 is then due to the centrifugal forces.

This example shows us clearly that absolute space is introduced as the (fictitious) cause. For S_1 cannot be responsible for the flattening of S_2, since the two bodies are in exactly the same condition relatively to each other and therefore cannot deform each other differently.

To take space as a cause does not satisfy the requirements of logic with regard to causality. For as we know *no* other expression of its existence than centrifugal forces, we can support the hypothesis of absolute space by nothing beyond the fact for the explanation of which it was introduced. Sound epistemological criticism refuses to accept such made-to-order hypotheses. They are too ready to hand and they disregard all bounds that scientific research seeks to interpose between its results and the wild dreams of fancy. If the sheet of paper on which I have just written suddenly flies up from the table I should be free to make the hypothesis that the spectre of Newton had spirited it away. But common sense prevents me making this hypothesis and leads me to think of the draught which arises because the window is open and someone is entering by the door. Even if I do not feel the draught myself, this hypothesis is reasonable because it brings the phenomenon which is to be explained into relationship with other observable events. This critical choice of admissible causes distinguishes the logical view of the world based on cause and effect, which includes physical research, from mysticism, spiritism, and similar manifestations of unbridled fancy.

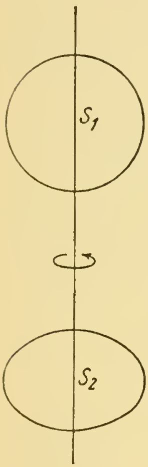


FIG. 126.

But absolute space is almost spiritualistic in character. If we ask " what is the cause of centrifugal forces ? " the answer is : " absolute space." If, however, we ask what is absolute space and in what other way does it express itself, no one can furnish an answer other than that absolute space is the cause of centrifugal forces but has no further properties. This presentation shows with sufficient clearness that space as the cause of physical occurrences must be eliminated from the world-picture.

It is, perhaps, not superfluous to mention that this opinion of absolute space is in no wise affected by the introduction of electromagnetic phenomena. With them effects occur in rotating co-ordinate systems which are analogous to the centrifugal forces of mechanics. But, of course, this does not give new and independent proofs of the existence of absolute space, for, as we know, the theorem of the inertia of energy amalgamates mechanics and electrodynamics to a complete whole. It is merely more convenient for us to operate with the conceptions of mechanics alone.

Let us now again consider the two bodies S_1 and S_2. If space is not accepted as the cause of their different behaviour we must look for other and more convincing causes.

Let it be supposed that there are no other material bodies at all outside the two bodies S_1 and S_2. The different behaviour of S_1 and S_2 would then be really inexplicable. But is this behaviour, then, an empirical fact? There is no doubt that it is *not*. We have never been able to gather experience of two bodies that are poised alone in the universe. The assumption that two real bodies S_1 and S_2 behave differently under these circumstances is supported by *no evidence* at all. Rather, we must demand of a satisfactory mechanics that it exclude this assumption. But if we observe in the case of two real bodies S_1 and S_2 the different behaviour above described (we are acquainted with planets that are more or less flattened) we can take as the reason of this only *distant masses*. In the real world such masses are actually present, namely, the countless legion of stars. Whatever stellar body we select, it is surrounded by innumerable others which are enormously distant from it and which move so slowly relatively to each other that, as a whole, they exert the effect of a solid mass containing a cavity in which the body under consideration is situated.

These distant masses must be the cause of the centrifugal forces. All our experiences are in agreement with this. For the system of reference of astronomy with respect to which the rotations of heavenly bodies are determined has been chosen so that it is at rest relatively to the stellar system as a whole, or, more accurately, that the apparent motions of the fixed stars relative to the system of reference are quite irregular and have no favoured direction. The flattening of a planet is the greater, the greater its velocity of rotation with respect to this system of reference which is attached to the different masses.

Accordingly we shall demand that the laws of mechanics and, indeed, of physics in general involve only the relative positions and motions of bodies. No system of reference may be favoured *a priori* as was the case with the inertial systems of Newtonian mechanics and of Einstein's special theory of relativity, or otherwise absolute accelerations with respect to these favoured systems of reference, and not only the relative motions of bodies, would enter into physical laws.

We thus arrive at the postulate that the true laws of physics must hold in exactly the same way in systems of reference that are moving arbitrarily. This denotes a considerable extension of the principle of relativity.

2. The Principle of Equivalence

Fulfilment of this postulate requires an entirely new formulation of the law of inertia since this is what gives inertial systems their favoured position. The inertia of a body is no longer to be regarded as an effect of absolute space but rather as one due to other bodies.

Now we know of only one kind of inter-action between *all* material bodies, namely, gravitation. Further, we know that experiment has exhibited a remarkable relationship between gravitation and inertia, which is expressed in the law of the equality of gravitational and inertial mass (II, 12, p. 37). Thus the two phenomena of inertia and attraction which are so different in Newton's formulation must have a common root.

This is the great discovery of Einstein which has transformed the general principle of relativity from a postulate of the theory of knowledge to a law of exact science.

We may characterize the object of the following investigation thus. In ordinary mechanics the motion of a heavy body (on which no electromagnetic nor other forces act) is determined by two causes: (1) its inertia towards accelerations with respect to absolute space; (2) the gravitation of the remaining masses. A formulation of the law of motion is now to be found in which inertia and gravitation amalgamate to a conception of higher order in such a way that the motion is determined only by the distribution of the remaining masses in the universe. But before we set up the new law we must follow a somewhat longer road to overcome certain conceptual difficulties.

We discussed the law of the equality of gravitational and inertial mass in detail earlier. For events on the earth it states that all bodies fall equally quickly; for motions of the heavenly bodies it expresses that the acceleration is independent of the mass of the moving body. We have already mentioned that according to measurements of Eötvös this law is valid to an extraordinary degree of accuracy, but that, in spite of this, it is not reckoned among the fundamental laws in classical mechanics but rather is accepted, so to speak, as an accidental gift of Nature.

This is now to become different. This law plays the fundamental part not only in mechanics but, indeed, in the whole of physics. We must, therefore, illustrate it so that its essential content comes out quite clearly. We advise the reader to make the following simple experiment. Let him take two light but differently heavy objects, say a coin and a piece of indiarubber, and place them on the palm of his hand. He then experiences the weights of the two bodies as pressures on his hand, and finds them different. Now let him move his hand rapidly downwards, he experiences a diminution in the pressure of both bodies. If this motion is continued more and more rapidly a moment will finally come when the bodies will release themselves from his palm and will lag behind in the motion. This will clearly occur as soon as the hand is drawn down more rapidly than the bodies can fall freely. Now, since they fall equally quickly in spite of their different weights they always remain together at the same height even when they are no longer in contact with the hand.

Let us imagine little imps living on the surface of the hand who know nothing of the outer world. How would they judge this whole process? It is easy to imagine oneself in the position of such little observers moving with the hand, if we make the experiment and pay attention to the changing pressures and motions of the bodies with respect to the hand. When the hand is at rest the imps will establish that the two bodies have different weights. When the hand sinks they will note a decrease of weight of the bodies. They will look for a cause and will observe that their point of vantage, the hand, sinks relatively to the surrounding bodies, the walls of the room. But we may also imprison the imp and the two test bodies in a closed box and pull this box downwards with the hand. The observers in the box then observe nothing which will allow them to establish the motion of the box. They can simply note the fact that the weight of all bodies in the box decreases at the same rate. If the hand is now moved so rapidly that the objects cannot follow but fall freely, the observers in the box will notice to their astonishment that the objects which were just before considerably heavy now fly upwards. They acquire a negative weight, or rather, gravitation no longer acts downwards but upwards. Moreover, in spite of their different weights the two bodies fall equally quickly upwards. The people in the box can account for these observations in two ways. Either, they think that the gravitational field continues to act unaltered but that the box is accelerated in the direction of the field, or they assume that the masses which previously exerted an attractive force below the box have disappeared, and that in their place new masses have appeared above the box so that the direction of action of gravitation has been reversed. We then enquire whether there is any means of distinguishing by experiments within the box between these two possibilities ?

And we are bound to answer that physics knows of *no* such means. Actually, the effect of gravitation can in no wise be distinguished from the effect of acceleration ; each is fully equivalent to the other. This is essentially due to the circumstance that all bodies fall *equally* quickly. If this were not the case we could at once distinguish whether an accelerated motion of bodies of different weight is produced by the attraction of outside masses or is an illusion arising from the acceleration of the observer's point of support. For in the first case the bodies of different weight move with different velocities whereas in the latter case the relative acceleration of all freely moving bodies with respect to the observer is equally great and they fall equally quickly in spite of their different weights.

This *principle of equivalence* of Einstein is thus one of those theorems which we have particularly emphasized in this book, namely, such as assert that a certain physical statement cannot be established or that two conceptions cannot be distinguished. Physics refuses to accept such conceptions and theorems and replaces them by new ones. For only ascertainable facts have physical reality.

Classical mechanics distinguishes between the motion of a body that is left to itself and is subject to no forces, inertial motion, and the motion of a body under the action of gravitation. The former is rectilinear and uniform in an inertial system ; the latter occurs in curvilinear paths and is non-uniform. According to the principle of equivalence this distinction must be dropped. For by merely passing over to an accelerated system of reference we can transform the rectilinear uniform motion of inertia into a curved, accelerated, motion, which cannot be distinguished from one produced by gravitation. And the converse holds too, at least for limited portions of the motion, as will be explained more fully later. From now onwards we shall call every motion of a body on which no forces of an electrical, magnetic, or other origin act, but is only under the influence of gravitating masses, *an inertial motion*. This term is thus to have a more general significance than earlier. The theorem that the inertial motion relative to the inertial system is uniform and rectilinear, namely, the ordinary law of inertia, now comes to an end. Rather our problem now is to state the law of inertial motion in the generalized sense.

The solution of this problem releases us from absolute space and at the same time furnishes us with a *theory of gravitation* which thereby becomes linked up much more intimately with the principles of mechanics than Newton's theory.

We shall supplement these remarks by adding a few calculations. We have shown earlier (III, 8, p. 68) that the equations of motion of mechanics referred to a system S which has the constant acceleration k with respect to the inertial systems may be written in the form

$$mb = K'$$

where K′ denotes the sum of the true force K and the inertial force $-mk$, i.e.,

$$K' = K - mk.$$

Now, if K is the force of gravitation, then $K = mg$, thus

$$K' = m(g - k).$$

By choosing the acceleration k of the system of reference S appropriately we can make the difference $g - k$ assume any arbitrary positive or negative value, or zero. If, in analogy with electrodynamics, we call the force on unit mass the "*intensity of field*" of gravitation, and the space in which it acts, the *gravitational field*, we may say that by choosing the accelerated system appropriately, we can produce a constant gravitational field, reduce one that is given, annul, intensify or reverse it.

It is clear that within a sufficiently small portion of space and during a small interval of time any arbitrary gravitational field may be regarded as approximately constant. Hence we can always find an accelerated system of reference relative to which there is no gravitational field in the limited space-time region.

We next ask whether it is not possible to eliminate every gravitational field in its whole extent and for all times by merely choosing an appropriate system of reference, that is, whether gravitation may, to a certain extent, be regarded as "apparent." But this is clearly not the case. The field of the earth, for example, cannot be fully eliminated. For it is directed towards the centre, thus the acceleration would have to point away from it (the centre) ; but this is not possible. Even if we were to admit (and this we shall have to do) that the system of reference is not rigid but extends with acceleration about the centre, this motion would not have been possible for any arbitrary length of time but would have to have begun at some moment at the centre. By rotating the system of reference about an axis we get an inertial force directed away from this axis (III, 9, p. 71, formula (31)), namely, the centrifugal force

$$mk = m\frac{4\pi^2 r}{T^2}.$$

This compensates the gravitational field of the earth only at a certain distance r, namely, at that of the radius of the moon's orbit, supposed circular, with the time of revolution T.

Thus there are "true" gravitational fields, yet the sense of this word in the general theory of relativity is different from that in classical mechanics. For we can always eliminate an arbitrary sufficiently small part of the field by choosing the system of reference appropriately. We shall define the conception of gravitational fields more accurately later.

There are, of course, certain gravitational fields which can be eliminated to their full extent by a suitable choice of the system of reference. To find such we need only start from a system of reference in which a part of space is fieldless and then introduce a system of reference which is accelerated in some way. Relative to this there is then a gravitational field. It vanishes as soon as we return to the original system of reference. The centrifugal field $k = \frac{4\pi^2 r}{T^2}$ is of this kind. The question as to the conditions under which a gravitational field can be made to vanish in its whole extent can be answered, of course, only by the finished theory.

3. The Failure of Euclidean Geometry

But before we proceed we must overcome a difficulty which calls for a considerable effort.

We are accustomed to represent motions in the Minkowski world as world-lines. The framework of this four-dimensional geometry was furnished by the world-lines of light-rays and the orbits of inertial masses moving under no forces. In the old theory these world-lines are straight with respect to the inertial systems. But if we allow the general theory of relativity to be valid, accelerated systems are equivalent, and in them the world-lines that were previously straight are now curved (III, 1, p. 49, Fig. 32). And, in place of these, other world-lines become straight. Moreover, this is also true of the orbits in space. The conceptions straight and curved become relativized, so far as they are referred to the orbits of the light-rays and of freely moving bodies.

Through this the whole structure of Euclidean geometry is caused to totter. For this rests essentially (cf. III, 1, p. 48) on the classical law of inertia, which determines the straight lines.

It might now be thought that this difficulty could be surmounted by using only rigid measuring-rods to define such geometrical elements as straight line, plane, and so forth. But not even that is possible, as Einstein shows in the following way.

We start out from a space-time region in which no gravitational field exists during a certain time relative to an appropriately selected system of reference S.

Next, we consider a body which rotates in this region with a constant velocity of rotation, say a plane circular disc (Fig. 127) rotating on its axis at right angles to its own plane. We introduce a system of reference S′ which is rigidly fixed to this disc. A gravitational field directed outwards then exists in S′, and it is given by the centrifugal acceleration $k = \frac{4\pi^2 r}{T^2}$.

Now, an observer situated on S′ wishes to measure out the disc. To do so, he uses a rod of definite length as his unit, which must thus be at rest relatively to S′. An observer in the system of reference S uses exactly the same rod as *his* unit of length, and in this process it must, of course, be at rest relatively to S.

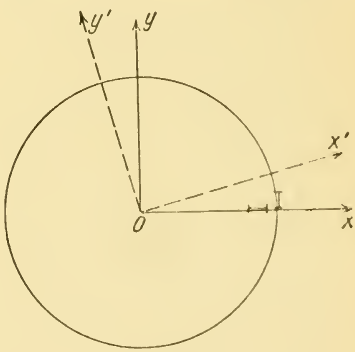


FIG. 127.

We shall now have to assume that the results of the special principle of relativity hold so long as we restrict ourselves to portions of space and time in which the motion can be regarded as uniform. To make this possible, we assume that the unit rod is small compared with the radius of the disc.

If the observer in S′ applies his rod in the direction of a radius of the disc, the observer in S will notice that the length of the moving rod relative to S remains unaltered and equal to 1. For the motion of the rod is perpendicular to the direction of its length. If the observer in S′ applies the rod to the periphery of the disc, then, by the special theory of relativity, it will appear shortened to the observer in S. If it be assumed that 100 little rods have to be applied end to end in order to reach from one end of the diameter to the other, the observer in S would require $\pi = 3{\cdot}14\ldots$ times 100, i.e. about 314 rods, which are at rest relatively to S, to measure out the periphery;

but the observer in S′ would not find this number of rods sufficient. For the rods that are at rest in S′ appear shortened as regarded from S, and it requires more than 314 of them to go once completely round the periphery.

Accordingly the observer in S′ would assert that the ratio of the circumference of the circle to the diameter is not $\pi = 3{\cdot}14 \ldots$ but more; and this is a contradiction of Euclidean geometry.

An exactly corresponding result holds for the measurement of times. If we bring one of two similarly constructed clocks to the centre, the other to the rim of the disc at rest relatively to the latter, the second clock, as regarded from the system S, goes more slowly, because it is moving relatively to S.

An observer situated in the middle of the disc would necessarily establish the same result. It is thus impossible to arrive at a reasonable definition of time with the help of clocks which are at rest relatively to the system of reference, if this system of reference is rotating, i.e. is being accelerated, or, what signifies the same according to the principle of equivalence, if a gravitational field exists in it.

In a gravitational field a rod is longer or shorter, a clock goes more quickly or more slowly, according to the position at which the measuring apparatus is situated.

This entails that the foundation of the space-time world, on which rest all the reflections we have so far made, collapses. We are again compelled to generalize the conceptions of space and time, but this time in a much more radical way, far exceeding the previous efforts in range.

It is clearly meaningless to define co-ordinates and time x, y, z, t in the ordinary way. For then the fundamental geometrical conceptions, straight line, plane, circle, and so forth, are regarded as immediately given, and the validity of Euclidean geometry in space or of Minkowski's generalization to the space-time world, is assumed.

Hence the problem arises to describe the four-dimensional world and its laws, without basing it on a definite geometry *a priori*.

It seems now as if the ground beneath us is giving way. Everything is tottering, straight is curved, and curved is straight. But the difficulty of this undertaking did not intimidate Einstein. Mathematicians had already accomplished important preparatory work. Gauss (1827) had sketched out the theory of curved surfaces in the form of a general two-dimensional geometry, and Riemann (1854) had founded the doctrine of space of continuous manifolds of any number of dimensions. We cannot here show how these mathematical instruments are applied, although a deeper understanding of the general principle of relativity is impossible without them. The reader must not, therefore, expect complete elucidation of Einstein's doctrine from the following discussion. He will find only pictures and analogies, which are always poor substitutes for exact conceptions. But if these indications stimulate the reader to further study, their purpose will have been fulfilled.

4. Geometry on Curved Surfaces

The problem of outlining a geometry without the framework of straight lines and their Euclidean connecting laws being given *a priori* is by no means so unusual as it may appear at first sight. Let us suppose that a surveyor has the task of measuring out a hilly piece of land quite covered by a dense wood, and that he has to sketch a map of it. From each point he can see only a quite limited part of the surroundings. Theodolites are useless to him; he has essentially to resort to the measuring chain. This enables him to measure out small triangles or quadrangles, whose corners are fixed by thrusting graduated poles into the ground; and by linking such directly measurable figures up with each other he can gradually advance to more distant parts of the wood, which are not directly visible.

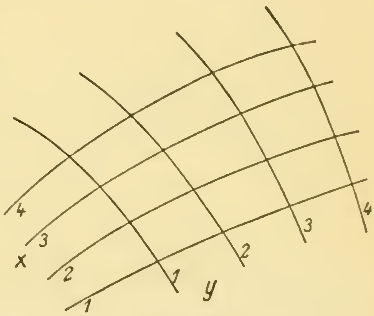


Fig. 128.

Expressed abstractly, the surveyor may apply the methods of ordinary Euclidean geometry to small regions. But these methods cannot be applied to the piece of land as a whole. It may be investigated geometrically only step by step, by proceeding from one place to the next. Nay more, Euclidean geometry is not strictly valid in hilly territory; there are *no* straight lines in it at all. Short pieces of line of the length of the measuring chain may be regarded as straight. But there is no straight connecting line along the ground from valley to valley, and from hill to hill. Euclidean geometry thus, in a certain sense, holds only in small or infinitesimal regions; but in greater regions a more general doctrine of space or rather of surfaces holds.

If the surveyor wishes to proceed systematically he will first cover the ground in the wood with a network of lines which are marked by poles or specified trees. He requires two families of lines which intersect (Fig. 128). The lines will be chosen as smoothly and continuously curved as possible, and in each family they will bear consecutive numbers. We take x as the symbol for any member of the one family, and y as that for any member of the other.

Each point of intersection has, then, two numbers x, y, say $x = 3$, $y = 5$. Intervening points may be characterized by fractional values of x and y.

This method of fixing the points of a curved surface was first applied by Gauss; x and y are, therefore, called *Gaussian co-ordinates.*

The essential feature involved is that the numbers x and y denote neither lengths, nor angles, nor other measurable geometrical quantities, but merely numbers, exactly as in the American system of numbering streets and houses.

The task of introducing a measure into this numbering of the points on the land falls to the lot of the surveyor. His measuring chain comprises about the region of one mesh of the network of Gaussian co-ordinates.

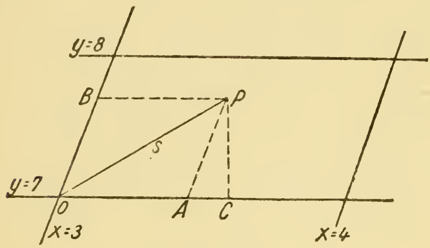


Fig. 129.

The surveyor will now proceed to measure out mesh for mesh. Each of these may be regarded as a small parallelogram and is defined when the lengths of two adjacent sides and an angle are known. The surveyor has to measure these and plot them in his map for each mesh. When this has been done for all meshes he clearly has a complete knowledge of the geometry of the land in his map.

In place of the 3 data for each mesh (2 sides and 1 angle) it is usual to apply a different method in determining the measure, which has the advantage of greater symmetry.

Let us consider a mesh, a parallelogram, whose sides correspond to two consecutive integers (say $x = 3$, $x = 4$, and $y = 7$, $y = 8$) (Fig. 129). Let P be any point within this mesh, and S its distance from the corner-point O with the smaller numbers. This will be measured out by the measuring chain. We draw the parallels to the net lines through P, and they intersect the net lines in A and B. Further, let C be the foot of the perpendicular dropped from P on to the x-co-ordinate.

The points A and B then also have numbers, or Gaussian co-ordinates, in the net. A is determined, say, by measuring the side of the parallelogram on which A lies and the distance AO, and by regarding the ratio of these two lengths as the increase of the x-co-ordinate of A towards O. We shall denote this increase itself by x, choosing O as the origin of the Gaussian co-ordinates. In the same way we determine the Gaussian co-ordinate y of B as the ratio in which B cuts the corresponding side of the parallelogram. x and y are then clearly the Gaussian co-ordinates of P.

The true length of OA is, of course, not x but, say, ax where a is a definite number to be determined by measurement. In the same way the true length of OB is not y but by. If we move the point P about, its Gaussian co-ordinates change, but the numbers a and b which give the ratio of the Gaussian co-ordinates to the true lengths remain unchanged.

We next express the distance OP $= s$ with the help of the right-angled triangle OPC according to Pythagoras' theorem. We have

$$s^2 = \mathrm{OP}^2 = \mathrm{OC}^2 + \mathrm{CP}^2.$$

Now OC = OA + AC, thus

$$s^2 = \mathrm{OA}^2 + 2\mathrm{OA} \,.\, \mathrm{AC} + \mathrm{AC}^2 + \mathrm{CP}^2.$$

On the other hand, in the right-angled triangle APC, we have

$$\mathrm{AC}^2 + \mathrm{CP}^2 = \mathrm{AP}^2.$$

Hence

$$s^2 = \mathrm{OA}^2 + 2\mathrm{OA} \,.\, \mathrm{AC} + \mathrm{AP}^2.$$

Here OA $= ax$, AP = OB $= by$. Further, AC is the projection of AP $= b \,.\, y$, and thus bears a fixed ratio to it, say AC $= cy$ Hence we get

$$s^2 = a^2x^2 + 2acxy + b^2y^2.$$

Here a, b, c are fixed ratio numbers. It is usual to designate the three factors of this equation differently and to set

$$s^2 = g_{11}x^2 + 2g_{12}xy + g_{22}y^2 \qquad . \qquad . \quad (97)$$

This equation may be called the *generalized Pythagorean Theorem* for Gaussian co-ordinates.

The three quantities g_{11}, g_{12}, g_{22} may serve, just like the sides and angle, to determine the actual conditions of size of the parallelogram. We therefore call them the *factors of the measure determination.* They have different values from mesh to mesh, which must be inserted in the map or given as " functions " with the help of analytical mechanics. But if they are known for every mesh, then, by formula (97), the true distance of an arbitrary point P within an arbitrary mesh from the origin can be calculated, so long as the numbers or the Gaussian co-ordinates x, y of P are given.

The factors of measure-determination thus represent the whole geometry on the surface.

It will be objected that this assertion cannot be right. For the network of Gaussian co-ordinates was chosen quite arbitrarily and so this arbitrary selection also applies to g_{11}, g_{12}, g_{22}. That is quite true. Another network could be chosen, and we should obtain for the distance between the same points O, P an expression built up just like (97) but with different factors $g'_{11}, g'_{12}, g'_{22}$. Yet there are, of course, rules for calculating g_{11}, g_{12}, g_{22}, transformation formulæ of a kind similar to those with which we became acquainted earlier.

Every real geometrical fact on the surface must clearly be expressed by such formulæ as remain unaltered for a change of the Gaussian co-ordinates, that is, are invariant. This makes the geometry of surfaces a theory of invariants of a very general type. For although the lines of the co-ordinate net are quite arbitrary, they must be so chosen that they are continuously curved and cover the surface singly and without gaps.

Now what are the geometrical problems that the surveyor has to solve as soon as he has obtained the measure-determination ?

There are no straight lines on the curved surface, but there are *straightest lines ;* these are at the same time those which form the shortest connection between two points. Their mathematical name is "*geodetic lines,*" and they are characterized mathematically thus : divide an arbitrary line on the surface into small, measurable sections of lengths $s_1, s_2, s_3, \ldots$; then the sum

$$s_1 + s_2 + s_3 + \ldots$$

for the geodetic line between two points P_1, P_2 is less than for any other line between them (Fig. 130). The $s_1, s_2, \ldots$ may be determined in this by mere arithmetic from the generalized Pythagorean Theorem (97), if the g_{11}, g_{12}, g_{22} are known.

On a spherical surface it is known that the "greatest" circles on the sphere are the shortest lines. They are cut out by the planes that pass through the centre. On other surfaces, they are often very complicated curves ; and yet they are the simplest curves which form the framework of geometry, just as straight lines form the framework of Euclidean geometry of the plane.

Geodetic lines are, of course, represented by invariant formulæ. They are real geometric properties of the surface.

All the higher invariants can be derived from these invariants. But we cannot here enter further into this question.

Another fundamental property of a surface is its *curvature*. It is generally defined with the help of the third dimension of space. The curvature of a sphere, for example, is measured with the help of the sphere's radius, that is, of a distance which lies outside the spherical surface. Our surveyor in the woody regions will not be able to apply these means. He cannot move out of his surface, so he has to try to find out the curvature conditions with his measuring chain alone. Gauss proved systematically that this is actually possible. We make this clear to ourselves by the following reflections.

The surveyor measures out twelve equally long wires with his measuring chain, and with them he forms the regular hexagon and its radii as shown in Fig. 131. According to a well-known theorem of ordinary geometry of the plane it is actually possible to have the twelve wires in one plane all stretched tight simultaneously. This is really very remarkable, for when five or six

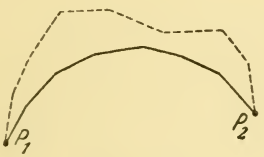


FIG. 130.

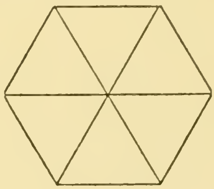


FIG. 131.

equiangular triangles are stretched out then the last wire must fit into its position accurately of itself. We learn at school that it does so, and what is learned at school is not usually much reflected on later. And yet it is very astonishing that the gap is filled in by a wire of exactly the same length as the other sides.

Actually this succeeds only in the plane. If we attempt the same thing on a curved surface in such a way that the centre and the six corners rest on it, the hexagon does not close. On the summits of hills and in the depths of valleys the last wire is too long, in passes (saddle-shaped curved surfaces) it is too short.

We advise the reader to try this for himself with twelve pieces of string and a cushion.

But this gives us a criterion as to how to find the curvature of surfaces without leaving the surface. If the hexagon is complete in it, then the surface is plane; if not, then it is curved. We shall not derive the measure of curvature. The indications given are sufficient to make it plausible that such a measure can be defined rigorously. It clearly depends on how the factors of measure-determination change from place to place. As Gauss has proved, the *measure of curvature* can be expressed in terms of the g_{11}, g_{12}, g_{22}, and it is an invariant of the plane.

Gauss' theory of surfaces is a method of geometry to which we can apply the expression *contact theory*, borrowed from physics. Not the laws of a surface on a large scale are given primarily, but their differential properties, the coefficients of the measure-determination and the invariants formed from them, above all, the measure of curvature. The form of the surface and its geometrical properties as a whole can then be determined subsequently, by processes of calculation, which are very similar to the solution of the differential equations of physics. In contrast with this, Euclidean geometry is a typical theory of action at a distance. This is why the new physics which is entirely built up on the conceptions of contact action, of the field, finds the Euclidean scheme insufficient and has to pursue new paths after the manner of Gauss.

5. The Two-dimensional Continuum

Let us suppose that our surveyor is operating with a wire hexagon to establish the curvature of the ground, and that he takes no account of the fact that there is a clear space in the wood in the middle of the hexagon which allows the sun to shine on to the ends of the wires that meet there. These wires will stretch a little owing to their being heated. Hence the six radial wires will be longer than the six outer wires, so that the latter will not join up. Hence if the ground is flat in reality, the surveyor will believe that he is standing on a flat hill (or in the hollow of a valley). If he is conscientious he will repeat the measurement with wires of another material. These will expand under the influence of the sun's heat more or less than those used before. This will draw his attention to the error and will lead him to correct it.

Now let us assume that the increase of length produced by the heating is the same for all the available materials of which the wires can be made. The error will then never come to light. Plains will be regarded as mountains and some mountains will be regarded as plains. Or let us imagine that some physical forces as yet unknown to us exert some influence on the lengths of rods and wires, but to the same extent in all cases. Then the geometry which the surveyor would determine with his measuring chain and wire polygons would turn out quite differently from the true geometry of the surface.

But so long as he operates in this geometry and has no possibility of adopting the higher standpoint of using the third dimension he will be firmly convinced that he has determined the correct geometry of the surface.

These reflections show us that the conception of geometry in a surface, or, as Gauss denominates it, "geometria intrinsica," has nothing to do with the form of the surface as it appears to an observer who has the third dimension of space at his disposal. Once the unit of length has been given by a measuring chain or a ruled scale the geometry in the surface is fully established *relatively* to this measure-determination, no matter what changes the measures undergo in reality during the process of measurement. These changes do not exist for a creature which is confined to the surface, so long as they affect all substances in the same way. Hence this creature will find curvatures where there are in reality none, and conversely. But this "in reality" becomes meaningless so far as surface creatures are concerned for they have no conception at all of a third dimension; just as we human beings have no idea of a fourth dimension of space. It is, therefore, also meaningless for these creatures to denominate their world as "a surface," which is embedded in a three-dimensional space; rather it is a "two-dimensional continuum." This continuum has a definite geometry, definite shortest or geodetic lines, and also a definite "measure of curvature" at every point. But the surface creatures will associate by no means the same idea with the latter phrase as we do with the intuitive conception of the curvature of a surface, rather they will only mean that the wire hexagon remains more or less open or closed and nothing more.

If the reader succeeds in experiencing in himself the feelings of this surface creature and in imagining the world as it appears to this creature the next stage of abstraction will present no difficulty.

Exactly the same thing might happen to us as human beings in our three-dimensional world. Perhaps this is embedded in a four-dimensional space in precisely the same way as a surface is embedded in our three-dimensional space; and unknown forces may change all lengths in certain regions of space without our ever being able to remark this directly. But then it would be possible for a spatial polyhedron, constructed after the manner of the six-sided figure, which should close according to ordinary geometry, to remain slightly open.

Have we ever detected anything of this sort? Since olden times Euclidean geometry has always been considered to be exact. Its theorems have even been declared in the critical philosophy of Kant (1781) to be *a priori* and, as it were, eternal truths. The great mathematicians and physicists, above all Gauss, Riemann, and Helmholtz have never shared this general belief. Gauss himself even once undertook an extensively planned measurement to test a theorem of Euclidean geometry, namely, that which asserts that the sum of the angles in a triangle amounts to two right angles (180°). He measured out the triangle between the three mountains, Brocken, Hoher Hagen, and Inselberg. The result was that the sum of the angles was found to be of the right amount within the limits of error.

Gauss was attacked on many sides by philosophers on account of this undertaking. It was asserted above all that even if he had detected deviations, this would at most have proved that the light rays between the telescopes had been deflected by some perhaps unknown physical causes, but nothing about the validity or non-validity of Euclidean geometry.

Now Einstein asserts, as we have already remarked above (p. 262), that the geometry of the real world is actually not Euclidean, and he supports this statement by concrete examples. To understand the relation of his doctrine to the early discussions about the foundations of geometry we must interpolate certain reflections of principle which verge on philosophic regions.

6. Mathematics and Reality

The question is: what is the object of geometric conceptions at all? Geometry certainly has its origin in the surveyor's art of measurement, that is, a purely empirical doctrine. The ancients discovered that geometrical theorems can be proved deductively, that is, that only a small number of principles or axioms need be assumed and then the whole system of the remaining theorems can be derived from them by mere logic. This discovery had a powerful effect. For geometry became the model of every deductive science, and it was regarded as the object of rigorous thinkers to demonstrate something "*more geometrico.*" Now what are the objects with which scientific geometry occupies itself? Philosophers and mathematicians have discussed this question from all points of view and have given a great number of answers. The certainty and incontrovertible correctness of geometric theorems was generally admitted. The only problem was how to arrive at such absolutely certain theorems and what were the things to which they referred.

It is without doubt true that if a person admits the geometric axioms to be correct then he is also compelled to recognize all the other theorems in geometry. For the sequence of the proofs is convincing for whosoever can think logically at all. This reduces the question to that of the origin of the axioms. In the axioms we have a small number of theorems about points, straight lines, planes, and similar conceptions, which are to hold exactly. For this reason, unlike most statements of science and of ordinary life, they cannot have their origin in experience; for this always furnishes only approximately correct and more or less probable laws. Hence we must look for other sources of knowledge which guarantee that these theorems are absolutely certain. According to Kant (1781) Time and Space are forms of intuition, which are *a priori*, which precede all experience, and which, indeed, first make experience possible. According to this the objects of geometry must be preconstructed forms of *pure* intuition, which are at the base of the judgments which we make about real objects in *empirical* intuition (direct perception). According to this the judgment "the edge of this ruler is straight" would come about by comparing the directly perceived edge with the pure intuition of a straight line, without this process of course coming into consciousness. The object of geometric science would then be the straight line given in pure intuition, that is neither a logical conception, nor a physical thing, but some third kind of thing whose nature can be communicated only by calling attention to the experience connected with the intuition "straight."

We do not presume to pronounce a judgment on this doctrine or on similar philosophical theories. These concern, above all, the experience of space, and this lies outside the scope of our book. Here we are dealing with the space and time of physics, that is, of a science which consciously and more and more clearly turns away from intuition as a source of knowledge and which demands more precise criteria.

We must now set it down as a fact that a physicist never founds the judgment "this edge of the ruler is straight" on direct intuition. It is a matter of indifference to him whether there is any such thing as a pure form of intuition of a straight line or not, with which the edge of the ruler can be compared. Rather he would make definite *experiments* to test the straightness, just as he would test every other assertion about objects, by means of experiments. For instance, he will look along the edge of the ruler, that is, he will ascertain whether a ray of light which touches the initial and the end point of the edge also just glides over all the remaining points of the edge (Fig. 132). Or he will turn the ruler about the end points of the edge and will make the point of a pencil touch any arbitrary intermediate point of the edge. If this contact remains unaffected by the rotation, the edge is straight (Fig. 133).

Now, if we subject these processes, which are evidently far superior to intuition, to criticism, we see that they, too, go no further into the question of absolute straightness. In the first method it is evidently already assumed that the ray of light follows a straight course. How do we prove that it does? In the second method it is assumed that the points about which the ruler is turned and the point of the pencil are in rigid connexion and that the ruler is itself rigid. Suppose that we wish to test the straightness of a rod with circular cross section, which is lying in a horizontal position and is a little bent owing to its own weight; then this bending will remain unaltered by the rotation, thus the method of contact will recognize straightness where there is in reality curvature. It is useless to object that these are sources of error which occur in every physical measurement and which are avoided by every expert experimenter. What we are concerned with is to show that absolute straightness or any other geometrical property cannot be directly proved empirically, but only

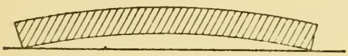


FIG. 132.

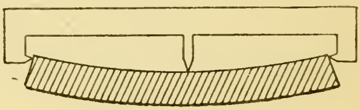


FIG. 133.

relatively to definite geometrical properties of the auxiliary means used in the measurement (straightness of the ray of light, rigidity of the part of the apparatus). If we divest the operations actually carried out of all the additional factors of thought, memory, knowledge, nothing remains except this discovery that: if two points of the ruler's edge lie on a ray of light, then so does this point or another; if two points of the ruler coincide with two points of a body, the same also holds for this or that third point. Thus, what is really ascertained is space or, rather, space-time coincidences, the meeting together of two recognizable material points at the same time and at the same place. All the rest is speculation, even such a simple assertion as that the straightness of the ruler can be determined by such experiments on coincidence.

A critical review of exact science teaches us that all our observations resolve finally into such coincidences. Each measurement states that a pointer or a mark coincides with some division on a scale at some time. No matter whether the measurement concerns lengths, times, forces, masses, electrical currents, chemical affinities, or anything else, all that is actually observable consists of space-time coincidences. In the language of Minkowski these are world-points that are marked in the space-time manifold by the intersection of material world-lines. Physics is the doctrine of the relations between such marked world-points.

Mathematical theory is the logical working out of these relations. However complicated it may be its ultimate object is always to represent the actually observed coincidences as the logical consequences of certain fundamental conceptions and principles. Some statements about coincidences occur in the form of geometrical theorems. Geometry as a doctrine that is applicable to the real world has no favoured rank above the other branches of physical science. The conceptions of forms are conditioned in the same way by the actual behaviour of the natural objects, just like the conceptions of other physical regions. We cannot allocate geometry to a special position.

The fact that Euclidean geometry hitherto reigned supreme is due to the fact that there are light rays which behave with very great accuracy like the straight lines of the conceptual scheme of Euclidean geometry, and that there are almost rigid bodies which satisfy the Euclidean axioms of congruence. The statement that geometry is absolutely exactly valid cannot be credited as having any sense from a physical point of view.

The objects of the geometry which is actually applied to the world of things are thus these things themselves regarded from a definite point of view. A straight line is by definition a ray of light, or an inertial orbit, or the totality of the points of a body regarded as rigid, which do not move when the body is turned about two fixed points, or some other physical something. Whether the straight line so defined has the properties which the geometry of Euclid asserts can be determined only from experience. Such a property of Euclidean geometry is exemplified in the theorem of the sum of the angles in a triangle which Gauss tested empirically. We must recognize that such experiments are thoroughly justified. Another characteristic property of two-dimensional geometry was given by the automatic closing of the wire hexagon (p. 261). Only experience can teach whether a definite way of realizing the straight line, the unit of length, and so forth has this property through definite physical things or not. In the former case Euclidean geometry is applicable relatively to these definitions, in the latter it is not.

Now Einstein asserts that all previous definitions of the fundamental conceptions of the space-time continuum by means of rigid measuring rods, clocks, rays of light, or inertial orbits in small limited regions certainly obey the laws of Euclidean geometry or of Minkowski's world, respectively, but not on the large scale. Only the smallness of the deviations is responsible for the lateness of their discovery. There are two ways out of the difficulty. Either we may give up defining the straight line by means of the ray of light, length by means of a rigid body, and so forth, and we may look for other realizations of the fundamental Euclidean conceptions in order to be able to retain the Euclidean system embodying their logical relationships ; or we may give Euclidean geometry itself up and endeavour to set up a more general doctrine of space.

It is clear to anyone who is not quite a stranger to science that the first way does not seriously come into consideration. Nevertheless we cannot prove that it is impossible. Here it is not logic that decides, but scientific discrimination. There is no logical path from fact to theory. Accidental ideas, intuition, fancy, are here, as everywhere, the sources of creative achievement, and the criterion of correctness is represented by the power of predicting phenomena that have not yet been investigated nor discovered. Let the reader assume for a moment that a ray of light in empty cosmic space is not the " straightest " thing there is, and let him work out the result. Then he will understand why Einstein pursued the other path.

As Euclidean geometry failed, he could have fallen back on some other definite non-Euclidean geometry. There are systems of conceptions of this sort worked out by Lobatschewski (1829), Bolyai (1832), Riemann (1854), Helmholtz (1866) and others, and these systems were evolved chiefly to test whether definite axioms of Euclid are necessary logical consequences of the others ; if they were, we should have to arrive at logical contradictions if we replaced them by other axioms. If we were to choose a special non-Euclidean geometry of this kind to represent the physical world we should simply be substituting one evil for another. Einstein went back to the physical root of phenomena, space-time coincidence, event, world-point.

7. The Measure-determination of the Space-time Continuum

The totality of marked world-points is what is actually ascertainable. In itself the four-dimensional space-time continuum is structureless. It is the actual relations of the world-points in it, which experiment discloses, that impresses a measure-determination and a geometry on it. Thus, in the real world we are confronted with exactly the same circumstances as those with which we just now became acquainted in considering surface-geometry. Hence the mathematical treatment will be the same in method.

First we shall introduce Gaussian co-ordinates into the four-dimensional world. We construct a network of marked world-points. This signifies that we consider space to be filled with matter moving arbitrarily, which may turn and be deformed in any way but is to maintain its continuous connexion; it is to be a sort of "mollusc," as Einstein expresses it. In it we draw three families of intersecting lines which we number, and we distinguish these families by the letters x, y, z. In the corners of the meshes of the resulting network we imagine clocks to be placed, which go at any arbitrary rate, but are arranged so that the difference of the data t of adjacent clocks is small. Thus the whole is a non-rigid system of reference, "a mollusc of reference." To it there corresponds in the four-dimensional world a system of Gaussian co-ordinates, consisting of a net of four numbered families of surfaces x, y, z, t.

All moving rigid systems of reference are, of course, special kinds of systems of reference which deform themselves in this way. But from our general point of view it is meaningless to introduce rigidity as something which is given *a priori*. The separation of time from space is also quite arbitrary. For, since the rate of the clocks can only be assumed to be quite arbitrary, even if continuously variable, space as the totality of all "simultaneous" world-points is not a physical reality. If different Gaussian co-ordinates are chosen other world-points become simultaneous.

But these things which do not alter when we pass from one system of Gaussian co-ordinates to another are the points of intersection of the real world-lines, the marked world-points, space-time coincidences. All really ascertainable facts of physics are qualitative relations between the positions of these world-points and thus remain unaffected by a change of Gaussian co-ordinates.

Such a transformation of the Gaussian co-ordinates of the space-time continuum denotes a transition from one system of reference to another that is arbitrarily deformed and in motion. The postulate of using in the laws of nature only what can really be ascertained or established thus brings it about that these co-ordinates are to be invariant with respect to *arbitrary transformations of the Gaussian co-ordinates* from x, y, z, t into x', y', z', t'. This postulate clearly contains the general principle of relativity, for among the transformations of x, y, z, t are also those which represent the transition from one three-dimensional system of reference to another which is moving arbitrarily. But, formally, it goes beyond this as it also includes arbitrary deformations of space and time.

In this way we have reached the foundation of the general doctrine of space which alone makes possible complete relativization. Our next step will be to link up this mathematical method with the physical reflections which we made earlier and which reached their climax in the enunciation of the principle of equivalence.

We are now in the same position with respect to the four-dimensional world as the surveyor in the woody regions, after he had marked out his co-ordinate network but had not yet begun to measure it out with his measuring chain. We must look round for a four-dimensional measuring chain.

This is furnished by the Principle of Equivalence. We know that by choosing the system of reference appropriately we can always secure that no gravitational field reigns in any part of the world if it be sufficiently small. There are an infinite number of such systems of reference, which move rectilinearly and uniformly with respect to each other, and for which the laws of the special theory of relativity hold. Measuring rods and clocks behave as expressed by Lorentz-transformations: light rays and inertial motions (see p. 254) are straight world-lines. Within this small region of the world the quantity

$$G = s^2 = x^2 + y^2 + z^2 - c^2t^2,$$

is an invariant with a direct physical meaning. For if the line connecting the origin O (which is assumed to be in the interior of the small region) with the world-point P (x, y, z, t) is a space-like world-line, then s is the distance, OP, in that system of reference in which the two points are simultaneous. But if the world-line OP is time-like, then $s = ict$, where t is the time-difference of the events O and P in the co-ordinate system in which both occur at the same point. Earlier (VI, 10, p. 245) we called s the four-dimensional distance. It is directly measurable by means of measuring rods and clocks, and so, if the imaginary co-ordinate $u = ict$ is introduced, it has, formally, the character of a Euclidean distance in the four-dimensional space:

$$s = \sqrt{G} = \sqrt{x^2 + y^2 + z^2 + u^2}.$$

The fact of the validity of the special theory of relativity in small regions corresponds exactly to the fact that Euclidean geometry can be applied to sufficiently small parts of a curved surface. But Euclidean geometry and the special theory of relativity need *not* hold in great regions. There need be no straight world-lines at all but only straightest or geodetic lines.

The further treatment of the four-dimensional world runs parallel with the theory of surfaces. First we must measure out the meshes of any arbitrary net of Gaussian co-ordinates with the help of the four-dimensional distance s. We interpret the process in a two-dimensional xt-plane (Fig. 134). Let a mesh of the co-ordinate net be bounded by the lines $x = 3$, $x = 4$, and $t = 7$, $t = 8$ (cf. Fig. 129 on p. 258). The rays of light that start out from the corner $x = 3$, $t = 7$ correspond to two world-lines which intersect and which we can draw as straight lines within a small region. The hyperbolic calibration curves $G = \pm 1$ lie between these light-lines. They correspond to the circle which, in ordinary geometry, contains the points which are at the same distance 1.

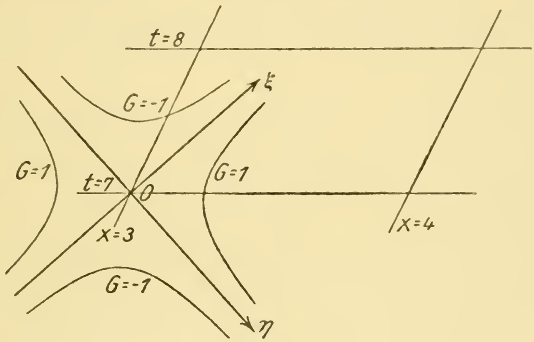


FIG. 134.

Then the application of formula (97) of the theory of surfaces leads to the expression

$$s^2 = g_{11}x^2 + 2g_{12}xu + g_{22}u^2,$$

for the invariant s, where x and $u = ict$ are the Gaussian co-ordinates of any point P of the mesh under consideration.

If we insert $u = ict$, we get

$$s^2 = g_{11}x^2 + 2icg_{12}xt - c^2g_{22}t^2,$$

or, if we call the factors differently :

$$s^2 = g_{11}x^2 + 2g_{12}xt + g_{22}t^2.$$

g_{11}, g_{12}, g_{22} are called *factors of the measure-determination* and may be interpreted directly physically. Thus, for example, for $t = 0$, $s = \sqrt{g_{11}}x$, that is, $\sqrt{g_{11}}$ denotes the true length of the spatial side of the mesh in the system of reference in which it is at rest.

In the four-dimensional world the invariant distance s

between two points whose relative Gaussian co-ordinates are x, y, z, t are represented by an expression of the form

$$\left.\begin{aligned} s^2 = g_{11}x^2 + g_{22}y^2 + g_{33}z^2 + g_{44}t^2 \\ + 2g_{12}xy + 2g_{13}xz + 2g_{14}xt \\ + 2g_{23}yz + 2g_{24}yt + 2g_{34}zt \end{aligned}\right\} \quad . \quad . \quad (98)$$

This formula may be called *the generalized Pythagorean theorem for the four-dimensional world.*

The quantities $g_{11}, \ldots g_{34}$ are the *factors of the measure-determination.* In general they will have different values from mesh to mesh of the co-ordinate net. Moreover, they will have other values for another choice of the Gaussian co-ordinates, and the new values will be connected with the original values by definite formulæ of transformation.

8. The Fundamental Laws of the New Mechanics

According to the general principle of relativity the laws of physical nature are represented by invariants for arbitrary transformations of the Gaussian co-ordinates, just as the geometric properties of a surface are invariant for arbitrary transformations of the curvilinear co-ordinates. The framework of the theory of surfaces was given by the geodetic lines. In just the same way geodetic lines are constructed in the four-dimensional world, that is, such lines as form the shortest connexion between two world-points ; and in this process the distance between two neighbouring points is to be measured by the invariant s.

Now what do the geodetic lines signify ? In such regions as are free of gravitation for an appropriate choice of the system of reference they are clearly straight lines with respect to this system. But the world-lines are either space-like ($s^2 > 0$) or time-like ($s^2 < 0$) or light-lines ($s = 0$). If we introduce a different system of Gaussian co-ordinates the same world-lines now become curved, but, of course, remain geodetic lines.

From this it follows that the geodetic lines must represent just those physical phenomena which are represented in ordinary geometry and mechanics by straight lines, namely, rays of light and motions of inertia. Thus we have found the required formulation for the *generalized law of inertia,* which comprises the phenomena of inertia and gravitation in one expression.

If the factors of the measure-determination $g_{11}, \ldots g_{34}$ relative to an arbitrary Gaussian co-ordinate system are known for every point of the net, the geodetic lines can be obtained by mere calculation. If there is no gravitational field present in a certain region relative to the co-ordinate system under consideration, then

$$\left.\begin{array}{l} g_{11} = g_{22} = g_{33} = 1, g_{44} = -c^2 \\ g_{12} = g_{13} = g_{14} = g_{23} = g_{24} = g_{34} = 0 \end{array}\right\} \quad . \quad (99)$$

for then the general expression of the distance (98) becomes reduced to $s^2 = x^2 + y^2 + z^2 - c^2t^2$. Deviations of g from this value thus denote the state which is called gravitational field in ordinary mechanics; hence inertial motions are non-uniform and curved, and ordinary mechanics gives the Newtonian force of attraction as the cause of this. The ten quantities g have thus a double function: (1) they define the measure-determination, the units of length and time; (2) they represent the gravitational field of ordinary mechanics. We say that the g's determine *the metrical field* or *the gravitational field.*

Einstein's theory is thus a wonderful amalgamation of geometry and physics, a synthesis of the laws of Pythagoras and Newton. It achieves this by thoroughly purifying the conceptions, space and time, of all the added ingredients of subjective intuition and by the utmost objectivation and relativization that is conceivable. This constitutes the significance of the new doctrine in the intellectual progress of mankind.

But the new formulation of the law of inertia is only the first step of the theory. We have introduced the g's abstractly and have found in them the means of describing mathematically the geometrical-mechanical state of the world relative to any arbitrary Gaussian co-ordinate system. Now the proper problem of the theory comes to light. It is as follows:

Laws are to be found, according to which the metrical field (the g's) can be determined for every point of the space-time continuum relatively to any Gaussian co-ordinate-system.

Concerning these laws we know the following at present:

1. They must be invariant with respect to an arbitrary change of the Gaussian co-ordinates.

2. They must be fully determined by the distribution of the material bodies.

To these there has to be added a formal condition, which Einstein has taken over from the ordinary Newtonian theory of gravitation. For if we represent the Newtonian theory as a theory of pseudo-action by contact by means of differential equations, then, like all field laws of physics, these are of the second order, and we should demand that the new laws of gravitation, which are differential equations in the g's, should also be at most of the second order.

Einstein has succeeded in deriving the equations of the metrical field or of the gravitational field from these postulates. Hilbert, Klein, Weyl, and other mathematicians have lent their efforts in investigating thoroughly and illuminating the formal structure of Einstein's formulæ. We cannot here give these laws and the arguments on which they are founded because this is impossible without the application of higher mathematics. A few indications will suffice.

We know from the theory of surfaces that curvature is an invariant with respect to any arbitrary change of the Gaussian co-ordinates, and it may be determined from measurements in the surface. The reader will remember the use of the hexagon of wires.

In an exactly analogous way invariants may be found for the four-dimensional world which are direct generalizations of the curvature invariant of the theory of surfaces. Let us consider it as arising in the following way: let all the geodetic world-lines which touch a two-dimensional surface which passes through a point P of the four-dimensional world start out from P. These geodetic lines themselves again occupy a surface which may be called a geodetic surface. Now, if we draw a hexagon within this surface, such that its sides and radii have the same four-dimensional length, this hexagon will not in general be closed; thus the geodetic surface is curved. If we make the geodetic surface through the point P take up other positions in the four-dimensional space, the curvature alters. The totality of all the curvatures of the geodetic surfaces through a point furnishes a number of independent invariants. If these are zero, the geodetic surfaces are plane, and the four-dimensional space is Euclidean. The deviations of the invariants from zero thus determine the gravitational fields and must depend on the distribution of the material bodies. But, according to the special theory of relativity (VI, 8, formula (94), p. 232), the mass of a body is equal to the energy divided by the square of the velocity of light. The distribution of matter is thus determined by certain energy-momentum-invariants. It is these to which the curvature invariants are set proportional. The factor of proportionality corresponds to the gravitational constant (III, 3, p. 53) of Newton's theory. The formulæ so obtained are the equations of the metrical field. If the space-time distribution of energy and momentum are given, the g's can be calculated, and they in their turn determine the motion of the material bodies and the distribution of their energy. The whole is a highly complicated system of differential equations. But this mathematical complexity is counterbalanced by the enormous conceptual advance which is given by its general invariance. For this is the expression of the complete relativity of all events. Absolute space has vanished finally out of the laws of physics.

We have yet to mention a terminology which usually excites aversion among non-mathematicians. We are accustomed to call the invariants of three-dimensional space which are analogous to surface-curvature or even those of four-dimensional space itself *the measures of curvature.* We say of space-time regions in which they differ from zero that they are "curved." The person of untrained mind usually becomes indignant at this. He states that he can understand something *in* space being curved, but it is sheer nonsense to imagine *space* being curved. Well, no one asks that it be imagined; can invisible light be imagined, or inaudible tones? If it be admitted that our senses fail us in these things, and that the methods of physics reach further, we must make up our minds to allow the same to the doctrine of space and time. For intuition perceives only what comes about as a mental process through the joint working of physical, physiological and psychological phenomena, and is, therefore, actually given by it. Physics does not, of course, deny that this which is actually given can be interpreted with great definiteness according to the classical laws of Euclid. The deviations which Einstein's theory predicts are so small that only the extraordinary accuracy of measurement of present-day physics and astronomy can disclose them. Nevertheless they are there, and if the sum of our experiments leads to the result that the space-time continuum is non-Euclidean or "curved," intuition must give way to the judgment pronounced by knowledge.

9. Mechanical Inferences and Confirmations

The first task of the new physics is to show that classical mechanics and physics is correct to a high degree of approximation, for otherwise it would be impossible to understand how two centuries of untiring and careful research could rest satisfied with it. The next problem is, then, to find out the deviations that are characteristic of the new theory and that allow them to be tested by experiment.

How is it that classical mechanics suffices to describe all earthly phenomena and almost all phenomena of cosmic motions? What takes the place of the conceptions of absolute space and absolute time without which, according to Newtonian principles, even the simplest facts like the behaviour of the Foucault pendulum, inertial and centrifugal forces, and so forth, cannot be explained ?

In principle we have already answered these questions at the beginning of our discussions about the general principle of relativity. We there (VII, 1, p. 249) set up as the basis of relativistic dynamics the law that distant masses as real causes have now to take the place of absolute space as a fictitious cause of physical phenomena. The cosmos as a whole, the army of stars, produces at every point and at every moment a definite metrical or gravitational field. How this is constituted on a large scale can be taught only by a speculation of a cosmological kind, such as we shall get to know later (VII, 11, p. 287). On a small scale, however, the metrical field must be "Euclidean" if the system of reference be appropriately chosen; that is, the inertial orbits and rays of light must be straight world-lines. Now, compared with the cosmos, even the dimensions of our planetary system are small, and hence the Newtonian laws hold in them with respect to an appropriate co-ordinate system so far as the sun or planetary masses do not produce local disturbances, which do not correspond to the attractions of the Newtonian theory. Astronomy teaches us that such a system of reference in which the action of the masses of the fixed stars within the region of our planetary system leads to the Euclidean measure-determination, is just at rest relatively to (or in uniform rectilinear motion with respect to) the totality of cosmic masses, and that the fixed stars execute relatively only small motions which cancel each other in the mean. An *explanation* of this astronomic fact can be given only by applying the new dynamic principles to the whole cosmos, which will engage our attention in the concluding section. Here we are for the present dealing with the mechanics and physics of the region within the planetary system. Then all doctrines of Newtonian mechanics remain almost unaltered. But we must bear in mind that the vibration plane of Foucault's pendulum remains fixed, not with respect to absolute space, but with respect to the system of distant masses, that is, that centrifugal forces do not occur in the case of absolute rotations, but in that of rotations with respect to distant masses. Furthermore, we are quite free to refer the laws of physics not to the ordinary system of co-ordinates, in which the metrical field is Euclidean and a gravitational field in the ordinary sense does not exist (except for the local fields of planetary masses), but to a system moving in any way whatsoever (or even deformed in itself); only in this case gravitational fields at once appear and geometry loses its

Euclidean character. The general form of all physical laws remains always the same, only the values of the quantities g_{11}, g_{12}, . . . g_{34}, which determine the metrical field or the gravitational field, are different in every system of reference. This invariance of the laws alone contains the difference between the new and the old dynamics ; here, too, we were able to pass over to systems of reference moving arbitrarily (or deformed), but then the physical laws did not retain their form. Rather, there were " simplest " forms of the physical laws, which were assumed in definite systems of co-ordinates at rest in absolute space. In the general theory of relativity there are no such simplest or favoured forms of the laws ; at the most, the numerical values of the quantities g_{11} . . . g_{34}, which occur in all physical laws might be particularly simple within limited spaces or be only slightly different from such simple values. Thus, geometry refers its formulæ to a system of reference which would be Euclidean within the small space of the planetary system, if there were no sun and no planets, that is, where the g_{11} . . . g_{34} would have the simple values of (99), p. 273. In reality, however, the g_{11} . . . g_{34} have not these values at all, but differ from them in the vicinity of the planetary masses, as we shall explain further later. Any other (say rotating) system of reference, in which the g_{11} . . . g_{34} have not the simple values of (99) even if there were no planetary masses, is thus in principle fully equivalent to the first. This gives us freedom to return to Ptolemy's point of view of a " motionless earth." This would mean that we use a system of reference rigidly fixed to the earth, whereby the g_{11} . . . g_{34} assume such values as correspond to the centrifugal field of the rotation with respect to distant masses. From Einstein's higher point of vantage Ptolemy and Copernicus are *equally* right. Both view-points furnish the *same* physical laws, but with different numerical values for the g_{11} . . . g_{34}. What point of view is chosen is not decided by principles but is a matter of expedience. For the mechanics of the planetary system the view of Copernicus is certainly the more convenient. But it is meaningless to call the gravitational fields that occur when a different system of reference is chosen " fictitious " in contrast with the " real " fields produced by near masses : it is just as meaningless as the question of the " real " length of a rod (VI, 5, p. 213) in the special theory of relativity. A gravitational field is neither " real " nor fictitious in itself. It has no meaning at all independently of the choice of co-ordinates, just as in the case of the length of a rod. Nor are the fields distinguished by the fact that some are produced by masses and others are not ; in the one case it is in particular the near masses that produce an effect, in the other it is the distant masses of the cosmos.

Arguments of "common sense" have been advanced against this doctrine; among them is the following: If a railway train encounters an obstacle and becomes shattered, this event can be described in two ways. Firstly, we may choose the earth (which is here regarded as at rest relatively to the cosmic masses) as a system of reference, and make the (negative) acceleration of the train responsible for the destruction. Or, secondly, we may choose a co-ordinate system rigidly attached to the train, and then, at the moment of collision, the whole world makes a jerk relatively to this system, and we get everywhere a very strong gravitational field parallel to the original motion, and this field causes the destruction of the train. Why does the church-tower in the neighbouring village not tumble down, too? Why do the consequences of the jerk and of the gravitational field associated with it make themselves remarked one-sidedly in the case of the train, whereas the following two statements are to be equivalent: the world is at rest and the train is slowing down—the train is at rest and the world is slowing down? The answer is as follows: The church-tower does not fall down because, during the retardation, its relative position to the distant cosmic masses is not changed at all. The jerk, which, as seen from the train, the whole world experiences, affects all bodies equally as far as the most distant stars and including the church-tower. All these bodies fall freely in the gravitational field which presents itself during the retardation, with the exception of the train, which is prevented by the retarding forces from falling freely. But with respect to *internal* events (such as the equilibrium of the church-tower) freely falling bodies behave just like bodies that are poised freely and are withdrawn from all influences. Thus, no disturbances of the equilibrium occur, and the church-tower does not tumble down. The train, however, is prevented from falling freely. This gives rise to forces and stresses which lead to its consequent destruction.

To appeal to " common sense " in these difficult questions is altogether a precarious proceeding. There are supporters of the theory of a substantial ether who take up arms against the theory of relativity because it is not tangible enough or cannot be pictured clearly. Some of these have finally come to recognize the special principle of relativity, now that experiments have indisputably decided in favour of it. But they still struggle against the principle of general relativity because it is contrary to common sense. To these Einstein makes the following reply: According to the special theory of relativity, the train in uniform motion is certainly a system of reference which is equivalent to the earth. Will the common sense of the engine-driver admit this? He will object that he has not to heat and oil the "surroundings" but the engine, and that it must, therefore, be the motion of the engine which shows the effect of his work. Such an application of common sense leads finally to the negation of all scientific thought. For the same common sense of the ordinary man asks why we should busy ourselves with relativity or cathode rays at all, since no material personal gain can be expected from it.

We now resume our consideration of the celestial mechanics from the point of view of Einstein, and turn our attention to the local gravitational fields which superpose themselves on the cosmic field on account of the existence of planetary masses.

We can give but a short resumé of these researches of Einstein, as they concern, in particular, mathematical consequences of the field equations.

The simplest problem is to determine the motion of a planet about the sun. In this we do best to start from the already mentioned Gaussian system in which the gravitational field is Euclidean and no gravitational field in the ordinary sense would be present in the region of the solar system, the sun and the planet being considered absent; this system is characterized by the circumstance that the $g_{11} \ldots g_{34}$ would have the values of (99), p. 273, if the sun's action be disregarded. It is merely a question of determining deviations from these values effected by the sun's mass. Einstein's field-equations serve to do this, and we find that if we assume the sun's mass to be distributed with spherical symmetry and hence the field, similarly these equations give us quite definite relatively simple expressions for the $g_{11} \ldots g_{34}$. Then we can calculate the planetary orbits as geodetic lines of this measure-determination. Its curvature which is regarded in Newton's theory as the action of the attractive force appears in Einstein's theory as a consequence of the curvature of the space-time world, of which they are the straightest lines.

Calculation now discloses that the planetary orbits determined in this way are with great approximation the same as in Newton's theory. This result is astounding if we bear in mind the totally different standpoint of the two theories. In the case of Newton we have absolute space which is unsatisfactory on logical grounds and a deflecting force which was invented *ad hoc* with the remarkable property that it is proportional to the inertial mass; in the case of Einstein a general principle, satisfying the requirements of the theory of knowledge without the addition of a special hypothesis. If Einstein's theory were to achieve no more than the subjection of Newtonian mechanics to the general principle of relativity, everyone who is seeking the simplest harmony in the laws of nature would prefer it.

But Einstein's theory does more. As already mentioned, it contains Newton's laws for planetary orbits as approximations. The exact laws are slightly different, and the difference becomes the greater the nearer the planet is to the sun. Now, in dealing with Newton's celestial mechanics (III, 4, p. 58) we have already seen that it fails in the case of just the planet which is nearest the sun, namely, Mercury. There is left an unexplained *motion of Mercury's perihelion* of 43 seconds of arc per century. But this is just the amount required by Einstein's theory. Its confirmation was, therefore, anticipated by Leverrier's calculation. This result is of extraordinary importance. For no new arbitrary constants enter into Einstein's formula, and the "anomaly" of Mercury is just as necessary a consequence of the theory as the result that Kepler's laws are valid for the planets far removed from the sun.

10. Optical Deductions and Confirmations

So far there have been found besides these astronomic deductions only a few optical phenomena which do not escape observation owing to the smallness of their effects.

One is the *displacement towards the red of the spectral lines* of the light which comes from stars of great mass. At their surfaces there is a strong gravitational field. This affects the measure-determination and causes a definite clock to go more slowly there than on the earth where the gravitational field is smaller. But we have such clocks in the atoms and molecules of luminescent gases. Their mechanism of vibration is certainly the same wherever the molecule happens to be, and thus the time of vibration is the same in those systems of reference in which the same gravitational field, say the field zero, is present.

If the time of vibration in the fieldless region of space is T, then $s = icT$ is the corresponding invariant distance of the world-points, which correspond to two successive extreme points of the vibration, relatively to the system of reference in which the atom is at rest. In a relatively accelerated system of reference in which there is a gravitational field, the same $s = ict$ is given by formula (98), in which x, y, z characterize the position of the atom and t is the time of vibration measured in this system. We may set $x = y = z = 0$ by choosing the origin of the space co-ordinates in the atom, then

$$s^2 = -c^2T^2 = g_{44}t^2.$$

Thus

$$t = T\frac{c}{\sqrt{-g_{44}}}.$$

Now it is only in a fieldless space that we have $g_{44} = -c^2$ (see formula (99), p. 273), thus $t = T$. But in the gravitational field g_{44} is different from $-c^2$, say $g_{44} = -c^2(1-\gamma)$. Thus the time of vibration is altered, and is equal to

$$t = T\frac{1}{\sqrt{1-\gamma}}$$

or, if the deviation γ is small, then the time of vibration is approximately (see note on p. 182) equal to

$$t = T\left(1 + \frac{\gamma}{2}\right) \quad . \quad . \quad . \quad (100)$$

This is the difference in the beating of two clocks which are situated at different places, for which the difference of the gravitational field given by g_{44} has the relative value γ.

Whether γ is positive or negative can be found by considering a simple case, in which the question can be answered directly with the help of the principle of equivalence. This can be done successfully for a constant gravitational field such as occurs at the immediate surface of a heavenly body. The action of such a field g may be replaced by an acceleration of the observer of the same value g and directed oppositely to the attraction. If l is the distance of the observer from the surface of the star, a light-wave which starts from it will take the time $t = \frac{l}{c}$ to reach him, and he will observe the wave as if he had during that time executed a motion of acceleration outwards of the amount g. When the light-wave reaches him he should, on this view, have the velocity $v = gt = \frac{gl}{c}$ in the direction of the light's motion; hence, by Doppler's principle (formula (40), p. 107) he observes the diminished vibration number

$$\nu' = \nu\left(1 - \frac{v}{c}\right) = \nu\left(1 - \frac{gl}{c^2}\right);$$

thus, the time of vibration $t = \frac{1}{\nu'}$ observed in the gravitational field is related to the $T = \frac{1}{\nu}$ determined in the fieldless space as follows:

$$t = T \frac{1}{1 - \frac{gl}{c^2}}$$

or, approximately,

$$t = T\left(1 + \frac{gl}{c^2}\right) \quad . \quad . \quad . \quad (101)$$

This formula gives in general the difference of rate of two clocks which are present in a constant gravitational field g but are separated by a distance l.

Accordingly, in a constant gravitational field the quantity $\gamma = \frac{2gl}{c^2}$ which occurs in (100) is positive. The time of vibration, and hence also the wave-length, becomes magnified in the case of a light-wave moving oppositely to the direction of attraction of the gravitational field. This result can be applied to the light which comes from the stars; the quantity γ will be positive. Hence all the spectral lines of the stars are displaced a little towards the red end of the spectrum. Although this effect is very small, research has just recently confirmed its existence.

At this stage we can fill in a gap which was left earlier (VI, 5, p. 216), namely, the complete explanation of the "clock paradox." In it we assumed two observers A and B, of which one, A, was at rest in an inertial system (of the special theory of relativity), whilst the other, B, set out on a journey. On B's return A's clock, by (76) page 215, is in advance of B's by the amount $\frac{\beta^2}{2}t_0$, where t_0 is the total time of the journey as measured in the system A. This formula of course holds only approximately, yet it suffices for our purpose so long as we use corresponding approximations in our other calculations.

Now we may also regard B as at rest. A then makes a journey in the reverse direction. But of course we cannot simply infer that B's clock must now be in advance of A's by exactly the same amount, for B is not at rest in an inertial system but is experiencing accelerations.

From the standpoint of the general theory of relativity we must rather take care that when the system of reference is altered definite gravitational fields must be introduced during the times of acceleration.

In the first case under consideration A is at rest in a region of space in which the measure-determination is Euclidean and gravitational fields are absent. In the second, B is at rest in a system of reference in which, during the departure, moment of turning, and arrival of A gravitational fields occur briefly, in which A falls whereas B is held fixed by external forces. Of these three gravitational fields the first and the last have no influence on the relative rates of the clocks of A and B, since they are at the same point at the moments of departure and return, and since a difference of rate occurs in a gravitational field, by (101), only when there is a distance l between the clocks. But a difference of rate occurs when A reverses his direction. If T is the time taken to reverse, during which a gravitational field arises, B being supposed at rest, then A's clock, which is at a distance l and in the gravitational field g, is in advance of B's clock, and this is given to a sufficient degree of approximation bv (101), p. 282, viz., by $\frac{gl}{c^2}\tau$. In the times, however, when A is moving uniformly and the special principle of relativity must be applied, A's clock must, conversely, be behind B's clock by the amount $\frac{\beta^2}{2}t_0$. Thus, on the whole A's clock will be in advance of B's by

$$\frac{gl}{c^2}\tau - \frac{\beta^2}{2}t_0$$

on A's return.

We next assert that this value agrees exactly with the result of the first point of view in which A was regarded at rest, namely, that it is equal to $\frac{\beta^2}{2}t_0$.

For, since the moving observer, in reversing his velocity v, assumes the velocity $-v$, his total change of velocity is $2v$. We get his acceleration by dividing this by τ, the time taken to effect this change. This gives $g = \frac{2v}{\tau}$ as his acceleration. On the other hand, at the moment of turning back half the duration t_0 of the journey is over. The distance between the two observers is then $l = v\frac{t_0}{2}$.

From this it follows that $gl = v^2\frac{t_0}{\tau}$

and

$$\frac{gl}{c^2}\tau - \frac{\beta^2}{2}t_0 = \frac{c^2}{v^2}t_0 - \frac{\beta^2}{2}t_0 = \frac{\beta^2}{2}t_0$$

which concludes the proof.

Thus the clock paradox is due to a false application of the special theory of relativity to a case in which the general theory should be applied.

A similar error lies at the root of the following objection, which is continually being brought forward, although the explanation is very simple.

According to the general theory of relativity a co-ordinate system which is rotating with respect to the fixed stars, that is, which is rigidly connected with the earth, is to be fully equivalent to a system which is at rest with respect to the fixed stars. In such a system, however, the fixed stars, themselves, acquire enormous velocities. If r is the distance of a star, its velocity becomes $v = \frac{2\pi r}{T}$, where T denotes the duration of a day. This becomes equal to the velocity of light c, if $r = \frac{cT}{2\pi}$. If r is measured in terms of the astronomic unit of length, the light-year,* we must divide this by $c \cdot 365$, T being set equal to 1 day. So soon as the distance exceeds $\frac{1}{2\pi \cdot 365}$ light-years, the velocity becomes greater than c. But even the nearest stars are several light-years distant from the sun. On the other hand the theory of relativity (VI, 6, p. 220) asserts that the velocity of material bodies must always be less than that of light. Here there seems to be a glaring contradiction.

This, however, arises only because the law $v < c$ is entirely restricted to the special theory of relativity. In the general theory it assumes the following more particular form. As we know, it is always possible to choose a system of reference such that Minkowski's world-geometry holds in the immediate neighbourhood of any arbitrary point, that is, so that the geometry is Euclidean, and that there is no gravitational field, and $g_{11}, \ldots g_{34}$ have the values of (99) on page 273. With respect to this system and in this narrow space the velocity of light $c = 3 \cdot 10^{10}$ cm./sec. is the upper limit for all velocities.

But as soon as these conditions are not fulfilled, that is, if gravitational fields are present, any velocity, either of material

* A light-year is the distance which light traverses with the velocity 300,000 kms. per sec. in one year (365 days).

bodies or of light, can assume any numerical value. For the light-lines in the world are determined by $G = s^2 = 0$, or if we restrict our attention to the xt-plane, by

$$s^2 = g_{11}x^2 + 2g_{14}xt + g_{44}t^2 = 0.$$

We can calculate $\frac{x}{t}$ from this quadratic equation, and this is the velocity of light. For example, if $g_{14} = 0$, we get from $g_{11}x^2 + g_{44}t^2 = 0$ the value $\frac{x}{t} = \sqrt{-\frac{g_{44}}{g_{11}}}$ as the velocity of light, and this depends on just how great g_{11} and g_{44} happen to be.

If we take the earth as the system of reference, we have the centrifugal field (III, 9, p. 70) $\frac{4\pi^2 r}{T^2}$, which assumes enormous values at great distances. Hence the g's have values that differ greatly from the Euclidean values of (99). Therefore the velocity of light is much greater for some directions of the light-ray than its ordinary value c, and other bodies can also attain much greater velocities.

In any arbitrary Gaussian co-ordinate system not only does the velocity of light become different, but the light-rays no longer remain straight. A second optical test of the general principle of relativity depends on this *curvature of the light-rays*. The world-lines of light are geodetic lines, just like the inertial orbits of material bodies, and hence, like the latter, will become curved. But on account of the great velocity of light the deflection of its rays is much less. We can see why this deflection should come about from the principle of equivalence without further theory. For, in an accelerated system of reference every rectilinear and uniform motion is curved and irregular, so that the same must hold for any arbitrary gravitational field.

A ray of light which, coming from a fixed star, passes close by the sun will thus be attracted to it and will describe a somewhat concave orbit with respect to the sun (Fig. 135). The observer on the earth will assign to the star a position on the extension of the ray that strikes his eye, and hence the star will appear a little displaced outwards. This deflection might be calculated from Newton's theory of attraction, in which the ray of light may be treated, say, as a comet which approaches with the velocity of light, and it is of historic interest that this idea was carried out as early as 1801 by the German mathematician and surveyor Soldner. We then get a formula similar to that of Einstein, but giving only half the value for the deflection. This is due to the circumstance that Einstein's theory assumes that the gravitational field in the neighbourhood of the sun must be more intense. It is just this apparently minute difference (which escaped Einstein's attention when he made his first provisional publication of the theory) that constitutes a particularly sharp criterion of the correctness of the general theory of relativity.

The deflection of the apparent positions of the fixed stars in the neighbourhood of the sun can be observed only during a total eclipse of the sun since otherwise the bright radiation of the sun renders invisible the stars in its vicinity.

An eclipse of the sun took place on 29th May, 1919. England sent two expeditions whose sole object was to ascertain whether the "Einstein effect" was actively present or not. One proceeded to the West Coast of Africa, the other to North Brazil, and they returned with a number of photographs of the stars surrounding the sun. The result obtained by measuring out the plates was declared on 6th November, 1919, and proclaimed the triumph of Einstein's theory. The displacement predicted by Einstein which is to amount to 1·7 seconds of arc was present to the full extent.

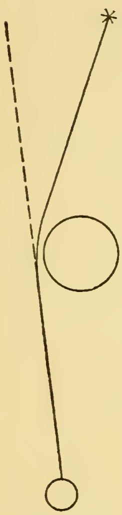


Fig. 135.

Since the remarkable achievement of the verification of this prophecy, the position of Einstein's doctrine may be regarded as assured in science.

The question whether it will be possible to find still other observable phenomena by which the theory can be tested cannot be answered with certainty. But since it is probable that the refinement of experimental observation of later decades or centuries will surpass our own by just as much as ours surpass that of Newton's time, we may expect that the new theory will be brought more and more into harmony with observation.

11. Macrocosm and Microcosm

We have seen above that the view that inertial forces are interactions necessarily leads to the logical consequence that the theory is applicable to the whole cosmos. The point is to understand why the system of reference for which Euclidean metrics hold in the neighbourhood of the solar system is relatively at rest (or in translational motion) with respect to the totality of cosmic masses. But observations of distant stellar systems, double stars, teaches us that there the same holds. This seems to indicate that the metrical field determined by the totality of all masses has everywhere the same character unless it is disturbed locally by neighbouring masses.

Since time immemorial speculations about the universe have been the favourite theme of minds given to fancy. But scientific astronomy has also occupied itself with such problems. Above all the question has been investigated whether there are a finite or an infinite number of heavenly bodies, and a decision has had to be taken in favour of a finite number. Here we can give only the merest indication of the argument. If the stars were distributed fairly regularly in space and if they were infinite in number the whole heavens would shine with a bright light, because then a star would have to be encountered in every direction somewhere or at some time, unless the light were to be weakened or absorbed on its way from the star to us. But good grounds can be adduced to support that there is no absorption of light in cosmic space. Hence we must regard the totality of stars as a gigantic accumulation which either suddenly ceases as we go outwards or at least gradually becomes diffuse.

But this view leads to a great difficulty if we start from Newtonian mechanics. Why do the stars remain together? Why do they not vanish from finite regions? We know that all stars have considerable velocities, but these are distributed irregularly in all directions and there is no indication that the parts of the whole are tending to separate.

The answer will be that gravitation holds the stars together.

But this answer is wrong. The methods of investigating such problems have long been known. They are those of the *kinetic theory of gases*. A gas consists of innumerable molecules that fly about in a disordered way, and we know the laws underlying such irregular motion.

Now it is clear that a gas which is not enclosed in rigid walls immediately expands and diffuses itself. Experiment and theory agree in teaching us that a system of bodies does not permanently remain together even if the bodies attract each other with forces which, according to Newton's law, are inversely proportional to the square of the distance.

The stellar system as a whole should behave exactly like a gas, and it seems impossible to grasp why it exhibits no tendency to lose itself in the infinite regions of space.

Einstein has given a very remarkable answer to this. He says that this is so because the world is not infinite at all. But where, then, are its limits? Is it not absurd to assume that the universe is somewhere " hedged in "?

Now, to be limited or bounded and to be finite is by no means the same. Consider the surface of a sphere. It is doubtless finite but it is without limits. Einstein asserts that three-dimensional space behaves in the same way. He finds it possible to do this because the general theory of relativity allows space to be curved. He thus arrives at the following theory of the universe. If we disregard the irregular distribution of the stars and replace it by a distribution of mass which is everywhere uniform we may ask under what conditions such a configuration of stars can remain permanently at rest according to the field equations of gravitation. The answer is that the measure of curvature of three-dimensional space must everywhere have a constant positive value just as in the case of a two-dimensional spherical surface. It is evident that a finite number of mass-points whose velocity tends to make them separate distribute themselves uniformly on a spherical surface and bring about a sort of dynamical equilibrium. An exactly corresponding result is to hold for the three-dimensional distribution of the stars. Einstein even estimates the value of this " world curvature with the help of a plausible assumption about the total mass of the stars." Unfortunately, it comes out so small * that for the present there is no hope of testing this bold idea experimentally.

It follows from the fact that the curvature of the world has the same value everywhere that the metrical field has the same character everywhere in the world, and that it is Euclidean in exactly that system of reference which is at rest with respect to all masses as a whole (or is moving uniformly and rectilinearly with respect to them). This statement contains the nucleus of the facts which Newton wished to represent by his doctrine of absolute space.

Every attempt to *picture* to oneself such a finite but unbounded " spherical " world is, of course, hopeless. It is just as impossible as trying to get an idea of the local curvatures of the world in the vicinity of gravitating masses. And yet this theory has very concrete consequences. Let us imagine a telescope in the Babelsberg Observatory directed at a definite fixed star. At the same time a telescope in the Antipodes, say at Sydney in Australia, to be directed at exactly the opposite point of the heavens. Then, according to Einstein's cosmology, it is conceivable that the observers at both telescopes see *one and the same* star, which is recognizable by, say, a characteristic spectrum. Actually, just as a person can

* According to an estimate of de Sitter the " circumference of the world," that is, the length of a geodetic world-line that returns into itself, is about 100 million light-years.

start a voyage around the earth by setting out towards the east or the west and travelling round on the same great circle in either direction, so in the spherical world of Einstein a light ray can start out from a star along a geodetic line in both directions and can meet the earth in opposite directions.

Nowadays we may still regard such reflections as the products of wanton fancy. Who knows whether they may not become, within a few centuries, empirical facts, owing to refinements in our methods of observation? It would be arrogant to deny this possibility. Even nowadays there are serious astronomers who take *Einstein's doctrine of the macrocosm* as the basis of their quantitative researches into the laws of the distribution of the fixed stars.

But Einstein's ideas also penetrate into the microcosm, the world of the atom. We have already touched earlier (V, 15, p. 187) on the question of the remarkable forces which prevent an electron or an atom from going asunder. Now, these configurations are enormous accumulations of energy in very small portions of space. Hence they will have in themselves considerable curvatures of space or, in other words, gravitational fields. It suggests itself to us that it is these fields that hold together the electric charges which tend to separate.

But this theory is only in its initial stage at present and it is quite uncertain whether it will be crowned with success. For we know from numerous experiments that new and strange laws rule in the atomic world and that they give expression to a harmony of integral numbers which is still quite unintelligible to us. It is the so-called *quantum theory* of Planck (1900). The decision rests with future investigators.

12. Conclusion

We now know, at least in coarse outline, Einstein's doctrine of space and time. We have traced its origin and development from the physical theories of his predecessors and we have seen how a clearly recognizable process of objectivation and relativization leads along the labyrinthine paths of research to the height of abstraction which characterizes the basic conceptions of the exact sciences of the present day. The power of the new doctrine is due to its immediate growth out of experience. It is an offspring of experimental science and has itself produced new experiments that bear witness to its merit. But what constitutes its importance beyond the narrow sphere of special research is the grandeur, the boldness, and the directness of the thoughts involved. Einstein's theory represents a type of thought the ideal of which is to keep a sound balance between freely creative fancy, critical logic, and a patient adaptation to fact. It is not a *world* view, if world signifies more than Minkowski's space-time manifold, but it leads whoever delves patiently into his ideas *to* a world view. For, beyond the bounds of science, too, objective and relative reflection is a gain, a release from prejudice, a liberation of the spirit from standards whose claim to absolute validity melts away before the critical judgment of the relativist.

INDEX

PRINTED IN GREAT BRITAIN AT THE UNIVERSITY PRESS, ABERDEEN

2
3
4
3