CAMBRIDGE PHYSICAL SERIES

RELATIVITY

CAMBRIDGE UNIVERSITY PRESS
C. F. CLAY, MANAGER
LONDON : FETTER LANE, E.C. 4

LONDON : H K LEWIS AND CO , LTD ,
136, Gower Street, W C. 1
NEW YORK THE MACMILLAN CO.
BOMBAY }
CALCUTTA } MACMILLAN AND CO., LTD.
MADRAS }
TORONTO : THE MACMILLAN CO. OF
CANADA, LTD.
TOKYO MARUZEN-KABUSHIKI-KAISHA

MODERN ELECTRICAL THEORY

SUPPLEMENTARY CHAPTERS

CHAPTER XVI

RELATIVITY

BY

NORMAN ROBERT CAMPBELL, Sc.D.

A MEMBER OF THE STAFF OF THE RESEARCH LABORATORIES
OF THE GENERAL ELFCTRIC COMPANY, LIMITED, LONDON

CAMBRIDGE
AT THE UNIVERSITY PRESS
1923

PRINTED IN GREAT BRITAIN

PREFACE[1]

IT is a serious responsibility to add even so slight a volume to the immense literature of relativity. A critic might well say that everything (except original developments) that can be said usefully has been said already, and that the field is sufficiently covered for all classes of English students by the translated works of Einstein himself and of Weyl and by Eddington's masterly book, the finest model of 'semi-popular' exposition our generation has produced. And in planning these chapters in the first instance I shared that view. But two considerations have changed my mind. In the first place no book of this character can be permanently complete without some description of one of the most important branches of modern physics; Chap. XIV must be replaced some day. Though, for the time being, most people are tired of relativity and desire to read no more about it, a new generation, not yet satisfied, is always arising; and it is to the needs of that new generation that the book is directed. In the second place, it remains an indubitable fact that, in spite of the attempts to enlighten him, the average physicist—the man in the laboratory, as I have ventured to call him—is still ignorant of Einstein's work and not very much interested in it. Physicists of great ability, who would be ashamed to admit that any other branch of physics is beyond their powers, will confess cheerfully to a complete inability to understand relativity. This attitude may merely prove that physicists are particularly obtuse; and, if that is so, it would be absurd for me to try to penetrate their dullness when so many others so much better qualified have failed. But there is an alternative explanation; it may be that there is some feature in the theory which is a greater difficulty to them than to others, greater to them even than to those who have no special knowledge of science at all.

I believe that the alternative explanation is correct, and it is only because I think I have discovered what that difficulty is that I venture on this task. And if I am right in my diagnosis, if my colleagues are really stumbling over the block that impeded my own progress to understanding so long, then an account of my difficulties may possibly help them to surmount their own. There is at least something new about this book. Authors of many kinds

[1] The nature and purpose of these supplementary chapters is described in the preface to Chap. XV.

have written on relativity; mathematicians, philosophers, journalists and cranks have all aired their views. But so far as I know nobody whose main occupation is the practice of experimental science has ever entered the field. It is not wholly impossible that here is the reason why so many of those who are thus occupied are still groping in darkness.

Accordingly it must be understood that I am addressing myself to a very limited class, namely to those who, either as students or in the later stages of their scientific career, are interested in relativity because it predicts something new which is experimentally true; to those whose interests in the matter are primarily mathematical or philosophical, I have nothing whatever to say. And my only object in addressing them is to help them to understand the process of reasoning which led up to the prediction of those experimentally true results. It is nothing more. A reviewer of the second edition of this book in an engineering journal was pleased to say that engineers had come to look to it to tell them what in modern physics they were to believe. He meant to be kind, but he could not have been more damning. Dogmatism is the worst fault this book could have; and though I must plead guilty to some excesses in the past, I am determined not to offend again. I am here concerned only to explain what the theory of relativity is, not to decide whether or no it is true.

All writers on relativity have found it necessary to devote some little attention to the general principles underlying physical science; the limited view that we are going to take reduces, but does not wholly remove, this necessity. I have purposely confined my remarks concerning anything that could, by any stretch of the imagination, be called philosophical to the minimum requisite for intelligibility. I have done so, not because I am uninterested in the general principles of science, nor because I think they have little bearing on relativity, but because I have said all that I have to say on the matter in my large book on *Physics*. Many of the topics discussed there were selected for no other reason than their bearing upon the difficulties of relativity.

I am indebted to my colleague, Mr J. W. Ryde, for revision of the proofs and much valuable criticism.

N. R. C.

Sept. 21, 1922.

CONTENTS

PAGE

I. THE FIRST AND SECOND PRINCIPLES 1

1. The problem of relativity. 2. The method of solution. 3. The first principle. 4. The second principle. 5. Consequences of the second principle. 6. The compounding of velocities. 7. General form of the second principle.

II. CONSEQUENCES OF THE FIRST AND SECOND PRINCIPLES 25

8. The original problem. 9. Solution of the original problem. 10. The pure relativity solution. 11. Extension of the principles. 12. The laws of electromagnetism. 13. Mass and energy.

III. THE GEOMETRICAL INTERPRETATION OF THE PRINCIPLES 50

14. The 'special' and the 'general' theory. 15. Minkowski's treatment of relativity. 16. 'Space' and 'Time'. 17. Mathematics and physics. 18. The value of the geometric interpretation.

IV. THE GENERAL THEORY 68

19. Accelerated motion. 20. The principle of equivalence. 21. The measurement of force. 22. Extension of the second principle. 23. The varieties of 'space'. 24. Spaces and fields of force. 25. The theory of tensors. 26. The law of gravitation. 27. The world-line of a particle. 28. Comparison with experiment. 29. Generalisation of the second principle. 30. The 'Einstein shift'. 31. Further developments.

V. SUMMARY 107

32. The special theory. 33. The general theory. 34. Are there alternatives?

INDEX 116

ERRATA TO CHAPTER XV

p. 24, Table I, line 1, *for* "6·17", *read* "8·25".

p. 36 footnote, *for* "Chap. XVI", *read* "Chap. XVII".

p. 37, Fig. 2, *for* "Ad" *read* "Yb".

p. 54, lines 10—11, *for* "is only one electron", *read* "are two electrons".

p. 84, lines 3—7, delete "This is.... .relativity".

p. 92, delete last sentence but one "A theory .. .currents".

RELATIVITY

I. THE FIRST AND SECOND PRINCIPLES

1. **The problem of relativity.** The principle or principles of relativity (for there are several) are analogous in a certain sense to the principles (usually called laws) of thermodynamics. The principles of thermodynamics are concerned with a certain group of physical laws, based on observations and measurements of a certain kind: these observations concern directly or indirectly the primary sensations of hot and cold, and the laws employ the conceptions of temperature and quantity of heat. The principles do not explain these laws, in the sense in which theories explain laws: they do not interpret them in terms of ideas which are not involved in the laws themselves; they rather lay down conditions which any possible laws of this group must fulfil. In just the same way the principles of relativity are concerned with a certain group of laws, concerning primary sensations of a certain kind. The laws are the laws of motion, and the primary sensations those of change of position. And just in the same way they do not profess to explain these laws (at any rate in the first instance), but rather to lay down what kind of laws are possible.

The term 'laws of motion' in this statement would be misleading without further explanation; for it is always used to denote the propositions on which Newton founded the science of dynamics, or any other propositions which may be proposed to take their place. We shall see that the later principles of relativity have a very important bearing upon the laws of motion in this narrower sense, but the earlier principles make no reference to them. The laws of motion with which Newton was concerned are those which determine under what conditions bodies are set into motion and how their motion may be changed. The laws of motion which originally gave rise to the principles of relativity are those which determine how the properties of bodies are affected by the fact of their motion, irrespective of the means by which the motion is set up or maintained. In its most general form the question from which our study starts is this:—I have examined completely some system when it is at rest relative to me, and I know all its laws. I now set the system in motion. How will the laws of that system be affected by the motion? From a knowledge of the motion and of the laws for the system at rest how can I infer the laws of the system in motion?

2. **The method of solution.** The principles of relativity attempt to answer those questions in a particular way. Suppose that after examining the system at rest and then setting it in motion, I set myself in motion also, so that I move with the system and am at rest relative to it; then, if the laws I have determined by its examination are really laws for the system at rest and hold whenever I am at rest relative to the system, when I examine the system again I shall find the same laws as before. If there is any general answer to the question that has been proposed, there must be some general relation between the laws that I determine when I am at rest and those which I determine when I am in motion. The question will be answered if I can find that relation. The principles of relativity all involve this conception of two observers, moving relatively to each other and both observing the same system; they assert relations between the observations made by such observers.

The questions, if put perfectly generally, are unanswerable. The relation between the laws for the system at rest and those for the system in motion depends generally on the nature of the system and on the character of the laws with which I am concerned. But according to the principles of relativity an answer can be given which will hold for all of an important class of laws and for all of an important class of systems. The laws and the systems must be such that if the system when it is in motion relative to the first observer O_1 is examined by another observer O_2, who is now at rest relative to the system and therefore moving relative to O_1, the laws of the system as determined by O_2 are precisely the same as those which were determined by O_1 when the system was at rest relative to him. In other words the laws and the systems must be such that the laws of the system are always the same when determined by an observer at rest relative to the system, whatever may be the motion of the system relative to any other observer. If there are such laws and such systems, then a general answer to the question we are discussing can be given, if there are any general relations between the observations made by O_1 and those made by O_2 who is moving relative to O_1. The principles of relativity assert that there are such general relations, valid for all the observations concerned in determining the laws of systems which obey the foregoing conditions. The question therefore is to be answered in this form. If the laws and system concerned fulfil the necessary conditions, an observer, having examined a system at rest relative to him, can determine how it will behave when it is moving relative to him, by imagining that the observations which he has just made on the system are being made by an observer who is moving relative to him in the same way as the system moves, and then applying to the observations made by such an observer a correction according to certain rules which are independent within wide limits of the nature of the observations made and dependent solely on the character of the motion.

That is an outline of the way in which the principles of relativity propose to answer the question; all the results which flow from those principles are determined by the mode in which it gives this answer. The conceptions, first of observations made on the same system by observers moving relative to each other, and second of the relation between the observations made by the two observers, are fundamental in our study and no progress can be made unless they are grasped. And yet—this is equally important to grasp—they are not really natural conceptions. It is hardly an exaggeration to say that never in the history of science have two observers moving relatively to each other made anything that can be called scientific observations on the same system. The vast majority of observations and all really accurate observations have been made by observers at rest relative to the earth and therefore (if the observations are made at the same time) at rest relative to each other. Comparatively rough observations are doubtless made by observers travelling in trains, or ships, or aeroplanes, but even then it is very rare, even if it has ever happened, for the systems which they observe to be observed at the same time by observers on the earth. It is difficult to think of a single instance in which the first characteristic conception of the principles of relativity has been realised at all completely, and it is quite certain that it has never been realised when observations of great accuracy are in question.

There is then no experimental evidence whatever for either of the first two principles of relativity, namely that which defines the systems the laws of which are unaltered so long as they are observed by an observer at rest relative to the system, and that which lays down the relations between the observations of two observers in relative motion. But a combination of the two principles may lead to predictions about experiments which can be made; for the same observer can observe the same system first when it is at rest and then when it is in motion relative to him. Accordingly we can examine whether the combination of the two principles is true; if it is true, it will clearly be important, for the problem it professes to solve is one of considerable interest. Moreover the resolution of the combination into its constituents and the solving of the problem in two stages rather than in one does serve some useful purpose. For though the conditions postulated by the two principles cannot be actually realised, they can be conceived; and if the principles appear to us to describe what might reasonably be expected to happen in these conceivable conditions, if they could be realised, we may be able to gather some kind of explanation of the rules given by the combination, even though the rules, regarded as mere rules and apart from the two principles which are combined in framing them, would appear perfectly arbitrary and inexplicable. This process, whereby laws (for the rules will be to some extent laws) are explained by imagining experiments made in conditions that are not actually realisable, is not peculiar to relativity; it is found in many scientific theories; for example it is essentially the same as that in which thermodynamical problems are solved by considering the behaviour of perfectly reversible engines, although no such engines actually exist.

3. **The first principle.** Let us now consider the actual principles. The first, as has been said, lays down what laws of what systems are unchanged so long as the observer who observes them is at rest relative to the system. All our physical knowledge depends on the isolation of certain parts of the world from the rest

of the world, in such a way that the behaviour of the isolated part is determined wholly by the character of that part and not by the rest from which it is isolated. If such isolation of independent systems were not possible, we should have to take into consideration the state of the entire material universe before we could lay down any law; we should have to consider that the weight of a body on the earth, for example, might depend on the eclipses of Jupiter's satellites; the task of reducing nature to order is rendered possible by its separation into parts of comprehensible size. In particular we find that the laws of the systems which we examine can usually be made entirely independent of the particular circumstances of any persons other than those actually observing them. It is plausible therefore to extend this independence to the motion of such persons, and to say that any laws of any system will be unaltered by its motion, if the whole system moves together, so that there is no motion between its parts, and if the observer observing it shares in the motion; the laws will not be affected by the motion of the system relative to any other observer.

This statement may be regarded as the first principle of relativity; it is the acceptance of that principle which has led to the name 'relativity'. For, obvious though it sounds, it has been disputed in the past, and even rejected in favour of a principle known as that of 'absolute' motion. According to the principle of absolute motion there is some system (which need not necessarily be or consist of material bodies directly observed) which forms part of every material system in the sense that the behaviour of all material systems is, or may be, affected by their relation to the absolute system, and especially by their motion relative to it. For instance the aether of nineteenth century physics was such an absolute system; the behaviour of all material systems was thought to depend in some measure on their motion relative to the aether. Now the recognition of an aether, or other absolute system, is not formally inconsistent with the first principle of relativity; for it still remains true that if the system includes, as now it must, the absolute system, its behaviour will be unaltered so long as its motion relative to the absolute system is unchanged. But since, on this view, a system, in changing its motion relative to any material system (for example, a person who is not observing it), may, and usually will, change at the same time its motion relative to the absolute system, the principle, though still true, has no applications. If there is an absolute system which is excluded from its scope, then the first principle of relativity is untrue; if it is included, the principle is useless. The first principle of relativity is in this sense a contradiction of the principle of absolute motion.

And there is another way in which the principle appears rather less obvious on further consideration. Even if we reject all absolute systems and consider only ordinary material bodies, it may happen that observations may be misleading in suggesting what bodies ought to be included in any given system. For example, if we investigated how a flag hangs round a mast in still air, we might conclude that the air was not to be regarded as part of the system, because the way the flag hangs is not actually affected by the presence of the air. But as soon as we put our flag and mast in motion (still observing it by an observer at rest relative to them) we should speedily find our error. The air ought to be included. We might also err in the other direction and expect a body to form an important part of a system in this sense, when actually it does not. Thus we shall see that, according to the first principle of relativity, a source of light is not part of the system which it illuminates; a conclusion that certainly could not have been anticipated without trial. In general we must remember that even if we admit the universal validity of the first principle as formally stated, we can only find out how to apply it and what to include in our independent systems by examining the consequences of our assumptions. The matter is not quite so simple as it appears at first sight.

Further, in applying the principle we have to be careful, not only what we mean by a system, but also what we mean by a law. In the first deductions we shall make from the principle, it can be interpreted in the strictest possible sense; everything that one observer finds to be true of the system relative to which he is at rest is found to be true by any other observer relatively at rest to the same system. But in our later deductions we shall have to interpret the principle rather differently and give a wider meaning to the term 'law'. The actual interpretation may be left until the need for it arises; but here we may notice the general meaning that is attributed to the first principle and underlies these wider interpretations. We have just seen that there is a sense in which the first principle is inconsistent with the conception of absolute motion; it is inconsistent with the belief that there is some absolute system which is a part of all other systems. But this is not exactly the belief which has been the subject of so much disputation in philosophy, to which the term 'absolute' properly belongs. The question which is there discussed is whether there is any meaning in saying that a body moves, if it is not said at the same time relative to what other body it moves; is there anything that can be termed absolute motion as distinguished from relative motion? Now if by motion we mean change of position, determined experimentally, absolute motion is, of course, meaningless; for in virtue of the meaning of the word, position, and therefore motion, is relative to something; we cannot determine experimentally the position of a body except by observing its relations to other bodies. When we say, as a result of experiment, that a body moves we mean that it moves relative to something else (usually ourselves), just as when we say a body is large we mean that it is larger than something else; in both cases we merely use an abbreviated form of words. Philosophy, which studies words and not their meaning, seizes on this abbreviated form of expression; believing apparently that any set of words which obey the rules of grammar must mean something, it holds that the statement that a body moves must mean something—even when no other body is mentioned. And in order to give the phrase a meaning, the philosophers who support the conception of absolute motion conclude that there must be some super-human observer or some super-material system relative to which the motion of any body can be determined uniquely. The first principle of relativity, in its wider applications, rejects this idea, as it rejects the other and less general idea. It denies that there is any super-material system or super-human observer, and holds that all systems are material and all observers human, and that the observations of one observer are as good and as valid as those of another who is related in exactly the same way to the system which he observes. Any proposition which would lead to a preference for the judgements of one observer over those of another it rejects, unless the preference can be justified by a difference in relation to the system observed. When we are comparing the observations of two observers, any relation by means of which we compare them must be symmetrical, so that a rule for determining the observations of A from those of B must also hold for the determination of the observations of B from those of A.

4. **The second principle.** We now pass to the second principle, which concerns the relation between the observations made on the same system by two observers who are moving relatively to each other and do not form part (in the sense described) of the system they observe. In particular it concerns the measurements which such observers will make on the distances between the places at which events happen and the times at which they happen; but we shall see that the relations between such measurements lead indirectly to relations between other measurements.

It is here that the first real difficulty in understanding the principles of relativity enters. For, before they were propounded we all had very definite ideas as to what these relations must be, and these ideas are so ingrained in us that it has become difficult to distinguish how far they are necessary to our way of thinking, so that any other relations are inconceivable, and how far they are mere generalisations from experience, so that other relations, though strange, would not be inconceivable. The second principle of relativity asks us to abandon these ideas for others; and in order that it may be understood exactly how far the acceptance of it involves a complete reconstruction of our manner of thinking, it will be well to inquire what solution we should have been disposed to offer to the problem it professes to solve, and what evidence we could bring in favour of our solution.

Let us start by taking a very simple and perfectly definite case. There are two thin rods, AB and CD, sliding along each other in the direction of their length with a constant relative velocity v (Fig. 1). There are two observers, one O_1 at rest relative to AB, and another O_2 at rest relative to CD; both observers, who are of course moving relative to each other, are observing both rods. Then we should be disposed to argue thus. In the course of the motion there will be instants at which A and B coincide with C and D; let us call the coincidence of A and C the event AC. The other events are BC, AD, BD. Now suppose that O_1 and O_2 are both measuring distances along the rods, O_1 from a point P on AB, O_2 from a point Q on CD; suppose that O_1 finds the distances PA, PB to be a_1 and a_2, while O_2 finds the distances QC, QD to be b_1, b_2. Let the times at which the events AC, BC, AD, BD occur be t_1, t_2, t_3, t_4. These times (accord-

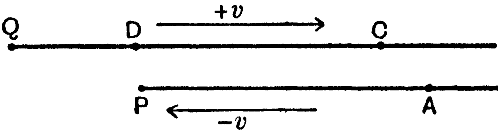

Fig. 1.

(The arrows next to each rod show the direction in which it moves relative to the other rod.)

ing to the view we are taking) are the same for both O_1 and O_2, but the distances at which the events happen are different for O_1 and O_2. In fact we have the following table, where x and X are the distances at which events happen from O_1 and O_2 respectively.

TABLE I.

Event	x	X	t
AC	a_1	b_1	t_1
BC	a_2	b_1	t_2
AD	a_1	b_2	t_3
BD	a_2	b_2	t_4

In virtue of the fact that the relative velocity of the rods is constant and equal to v (as observed by O_1) or $-v$ (as observed by O_2), there will be some relation between the distances and the times. The passage of C from A to B (observed by O_1) must occupy the time $(a_2 - a_1)/v$, and that of A from C to D (observed by O_2) the time $(b_2 - b_1)/-v$. Hence we have

$$v = \frac{a_2 - a_1}{t_2 - t_1} = \frac{a_2 - a_1}{t_4 - t_3} = \frac{b_2 - b_1}{t_1 - t_3} = \frac{b_2 - b_1}{t_2 - t_4} \quad \text{.........(4·1).}$$

Now consider the differences between x and t for two of the events, e.g., AC, BC. O_1 finds that the difference in x is $\delta x = a_1 - a_2$, while O_2 finds the difference in X is $\delta X = 0$; both find the difference in t is $\delta t = t_2 - t_1$. Comparison with (4·1) shows that

$$\delta x - \delta X = v \,.\, \delta t \quad \text{....................(4·2).}$$

If any other pair of events is examined, we get (4·2) again. Thus for the pair BC, AD,

$$\delta x = a_1 - a_2, \quad \delta X = b_2 - b_1, \quad \delta t = t_3 - t_2 = (t_1 - t_2) - (t_1 - t_3).$$

The equation (4·2) is perfectly general and states, in one form at least, the relation we should expect between the measurements of O_1 and O_2.

If O_1 and O_2 choose (as obviously they may) to measure distances from A and C, instead of from P and Q, and to measure times from the instant when A and C coincide, so that $a_1 = b_1 = t_1 = 0$, then the relation can be put into a somewhat different form. If we examine the table we shall find that for all four events

$$X = x - vt \quad \text{.........................(4·3)}$$

or

$$x = X - (-v)\,t \quad \text{.................(4·3·1).}$$

(4·3) expresses in a very obvious way our expectation, for we shall argue that we can arrive at the distance of an event as measured by O_2 (or O_1) by taking the distance as measured by O_1 (or O_2) and allowing for the distance travelled by O_2 (or O_1), as estimated by O_1 (or O_2), during the time that has elapsed since they were at the same point; the allowance is vt (if made by O_1) or $-vt$ (if made by O_2). Accordingly (4·3) expresses more directly the relation which we should expect, but it must be remembered that (4·1), to which we shall subsequently refer, is really identical with it and differs only by a mere algebraical alteration.

One further point should be raised before we leave this matter. Suppose that O_1 wanted to measure the distance between C and D, or the length of this part of O_2's rod, when it is moving relatively to him; how must he proceed? He must note with what two points on his own rod, AB, C and D are coincident at the same moment and then measure the distance between these two points on his own rod. There is no other way in which he can make the measurement; this method defines what he means by the length of CD. Now D is at A when $t = t_3$; C is at A when $t = t_1$, and at B when $t = t_2$; consequently when $t = t_3$, C is at a point for which x_1 is given by

$$\frac{x_1 - a_1}{a_2 - a_1} = \frac{t_3 - t_1}{t_2 - t_1} \quad \text{......................(4·4)}$$

or

$$x_1 - a_1 = v\,(t_3 - t_1) \quad \text{.................(4·4·1).}$$

But l, the length of CD, is defined by O_1 to be $x_1 - a_1$. Consequently

$$l = v\,(t_3 - t_1) \quad \text{......................(4·5).}$$

For future reference it may be well to note that this is also the length which would be estimated by observing the time interval between the passage of the two ends of CD past the same point on AB and multiplying the time interval by the velocity of CD relative to AB. For $(t_3 - t_1)$ is the time interval between AC, AD or (in virtue of (4·1)) between BC and BD. In fact these two methods of measuring the length are physically identical and differ merely in the algebraical form in which they are expressed. Lastly this length CD, estimated by O_1, agrees with that estimated by O_2 in virtue of (4·1), which gives $v(t_3 - t_1) = b_1 - b_2$.

So much for what we expect. The basis of our expectations is our belief that O_1 and O_2 will agree perfectly and always as to the time at which any given event happens. Now the second principle of relativity asks us to imagine that this belief is not universally true; that two observers may differ in the time they assign to an event; and, consequently, that a second event, which happens just before the time of this event to one observer and just after it to the other, may precede this event in the estimate of one and follow it in the estimate of the other. As has been said before, the principle does not ask us to believe that such things do happen, for the experiment cannot be tried in the conditions and with the accuracy it postulates; it only asks us to imagine that they might happen. But perhaps this request is the more serious. For the object of asking us to imagine anything is to make the ultimate predictions of what will happen more intelligible; and it may well be urged that an argument which involves the imagining of the inconceivable can hardly lead to intelligibility.

But is it really inconceivable that two observers might differ as to the time of an event and even as to the time sequence of two events? I think the difficulty that we find in conceiving such a possibility can probably be traced to two sources. First, there is the fact that all the events in the mental history of a single observer have a definite time sequence, the alteration of which is strictly inconceivable, at least to most of us. The whole of our conscious life is nothing but a perception of this time sequence, and, if it is suggested that some perception which we experienced after some other perception really happened before it, we shall be reduced to that state of mental confusion which provokes ecstasy in the philosopher and the fiercest anger in all normal men.

But the second principle of relativity has nothing to do with this personal time sequence; according to it every observer has his own time sequence which is perfectly normal. What it does assert is that the time sequences of different observers, who can only compare their time sequences by conversation, may be slightly different. Is it really so incomprehensible that observers may differ even in so fundamental a matter; is not the wonder rather that they agree as often as they do? The second source of the idea of a fixed sequence of events is the Law of Causation. We have had it so drilled into us by nineteenth century philosophers that causality is the mainspring of nature and that the cause always precedes the effect, that we have not only come to believe it but find it difficult even to imagine that it is untrue. And if every event has a cause, how can we imagine that, to some other observer, that event may precede the cause. But, as I have suggested, causation is a much more complicated (and scientifically a much less important) thing than these philosophers would have us believe. The matter cannot be discussed fully here and a partial discussion would do more harm than good. It must suffice to point out that, so far as the relation of cause and effect holds at all, each observer will be able to arrange his events unambiguously in a causal sequence, but different observers may possibly differ in their causal, as in their time sequences, and one may regard as the effect what another regards as the cause; but even this would not happen in any of the circumstances that have so far been imagined in the successful applications of the principles of relativity. Once more, is it so hard to believe that observers might differ?

But further, even if we admit that all events have a definite and unalterable time sequence, the question remains whether all observers can determine it correctly; there is the possibility that one of the differing observers may be judged to be wrong. This solution of the difficulty would be a very bad one, for it cuts at the root of the fundamental ideas of relativity which, implied in the first principle and elaborated in the second, involve the acceptance of the observations of one observer as neither more nor less valid than those of another. But there is the alternative, not so objectionable, that both observers may be wrong; and for this reason, and for others, it will be well to examine rather more nearly what is involved in the determination of the time of an event by the two observers.

The time of an event is determined by its simultaneity with some one of a standard series of events, which we may call generally a clock; when we say that an event happened at 9.30, we mean that it happened simultaneously with the event in the clock which we call 9.30. (Of course we choose our standard series in a particular way; but the discussion of that matter is irrelevant to our present purpose.) The two observers in relative motion can determine the time of events in one of two ways. First, they may both observe the same clock, which may be at rest relative to either or neither of them. But then, since all signals are propagated with a finite velocity, they have to apply a correction for the difference between the time that the event really happened in the clock and the time when it reached them at a finite distance. Since one or both of the observers must be changing their distance, this correction will be very complicated. Are we so sure that we know to the last conceivable accuracy how to make it? Second, they may observe different clocks which have been compared together when they are both at rest and both in precisely the same condition relative to the observer making the comparison, and which have been found to agree precisely. But they must take their clocks with them when they set out on their journey and they must thereby set them in relative motion. Is it really inconceivable that the clocks might be upset slightly by being so set in relative motion?

Those suggestions may possibly enable some readers to overcome their initial repugnance from imagining that observers could differ as to the time of the same event. Those who have not overcome it had better read no further; for unless the effort of imagination is successful, relativity must always remain unintelligible. I would repeat once more that only imagination, not belief, is required there is no question of actual fact. Nobody has ever denied that all pairs of observers always have agreed, and in all realisable conditions always will agree, in their estimates of times to the utmost accuracy with which such estimates can be made.

After this warning, we may proceed to state the second principle of relativity, and again we shall state it first for the simplest possible case. The circumstances considered are again those of Fig. 1

in which two thin rods slide along each other in the direction of their length with constant velocity. (From this point until the end of Section III, velocity will always mean constant velocity; accelerated motions are definitely excluded for the present from our consideration.) Some of the propositions inherent in the older view are accepted; thus, it is supposed that both observers will observe at some places and at some times the four events which consist of the coincidences AC, BC, AD, BD, and that the relative velocity of the two rods is arithmetically the same by whichever observer it is determined, though it will differ in sign if directions are measured in the sense indicated in the figure. These propositions were not expressed formally before; but they are really quite as dubitable and therefore need expression as much as those which form the rest of the principle. We can again draw up a table similar to Table I, but we shall now have to allow for the possibility that T, the time of an event observed by O_2, is not the same as t, the time observed by O_1; accordingly we have

TABLE II.

	Observer O_1		Observer O_2	
Event	x	t	X	T
AC	a_1	t_1	b_1	T_1
BC	a_2	t_2	b_1	T_2
AD	a_1	t_3	b_2	T_3
BD	a_2	t_4	b_2	T_4

In place of (4·1) we have

$$v = \frac{a_2 - a_1}{t_2 - t_1} = \frac{a_2 - a_1}{t_4 - t_3} \quad \text{..................(4·6)}$$

$$= \frac{b_1 - b_2}{T_3 - T_1} = \frac{b_1 - b_2}{T_4 - T_2} \quad \text{..................(4·7).}$$

The second principle of relativity states that, for the relation (4·2) characteristic of the older view, we must substitute the relation

$$c^2 (\delta t)^2 - (\delta x)^2 = c^2 (\delta T)^2 - (\delta X)^2 \quad \text{...........(4·8),}$$

where c is the value assigned to the velocity of light, i.e. with the centimetre and second as units, about 3×10^{10}. This relation appears at first sight so different from (4·2) that it may be difficult to believe that one can be a substitute for the other. For the moment let us leave this question on one side, together with others, such as how the relation came to be proposed; let us examine its consequences and return later to consider its meaning. It should be mentioned however that it is not usual to lay down (4·8) as the formal statement of the principle; (4·8) is usually deduced from other propositions which are regarded as more fundamental. Again certain conditions, which will appear presently, have to be imposed in order that (4·8) may express the principle at all; but I think our mode of procedure is the best.

5. Consequences of the second principle. In order to examine the consequences it will be well to introduce the following quantities: (1) l_1, the length of AB as measured by O_1, (2) L_2, the length of CD measured by O_2. By the meaning of length we have

$$l_1 = a_2 - a_1 \quad\text{.......................(5·1)},$$

$$L_2 = b_1 - b_2 \quad\text{.......................(5·2)}.$$

(3) l_2, the length of CD measured by O_1, (4) L_1, the length of AB measured by O_2. Since the method of p. 10 is again the only method for measuring these lengths, we have again from (4·5)

$$l_2 = v(t_3 - t_1) = v(t_4 - t_2) \quad\text{.................(5·3)},$$

$$L_1 = v(T_2 - T_1) = v(T_4 - T_3) \quad\text{...........(5·4)}.$$

What follows is mere algebra. Applying (4·8) to the pair of events AC, BC, we have

$$c^2(t_2 - t_1)^2 - (a_2 - a_1)^2 = c^2(T_2 - T_1)^2 \text{...........(5·5)}.$$

Applying (4·8) to the pair AC, AD, we have

$$c^2(t_3 - t_1)^2 = c^2(T_3 - T_1)^2 - (b_2 - b_1)^2 \text{...........(5·6)}.$$

Substituting for the t's from (4·6), for the T's from (4·7), for $a_2 - a$ from (5·1) and for $b_2 - b_1$ from (5·2), we have

$$l_1^2\left(\frac{c^2}{v^2} - 1\right) = L_1^2 \cdot \frac{c^2}{v^2} \quad\text{....................(5·7)}$$

$$l_2^2 \cdot \frac{c^2}{v^2} = L_2^2\left(\frac{c^2}{v^2} - 1\right) \text{....................(5·8)}.$$

Therefore

$$\frac{L_1}{l_1} = \frac{l_2}{L_2} = \frac{1}{\beta} \quad\text{.......................(5·9)},$$

where

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

β is a real quantity greater than 1, if $v < c$.[1]

[1] Note that German writers now use β for $\frac{v}{c}$.

Again, since from (4·6), (5·1), (5·4)

$$(T_2 - T_1)/(t_2 - t_1) = L_1/l_1 \quad \text{.............}(5\cdot10)$$

and from (4·7), (5·2), (5·3)

$$(t_3 - t_1)/(T_3 - T_1) = l_2/L_2 \quad \text{.............}(5\cdot11),$$

we have

$$\frac{T_2 - T_1}{t_2 - t_1} = \frac{t_3 - t_1}{T_3 - T_1} = \frac{1}{\beta} \quad \text{.................}(5\cdot12).$$

So far only two pairs of events out of the possible six have been used. If we take the pair BC, AD we shall find that, using (5·9) and (5·12), (4·8) is an identity. If we take BC, BD and AD, BD we find (5·9) again and in place of (5·12)

$$\frac{T_4 - T_3}{t_4 - t_3} = \frac{t_4 - t_2}{T_4 - T_2} = \frac{1}{\beta} \quad \text{..............}(5\cdot12\cdot1).$$

And if we take lastly AC, BD we find again (4·8) is an identity.

The occurrence of these identities means that we have assumed more than was necessary, and that we may prove some of our assumptions instead of assuming them. Thus we may regard ourselves as having proved that, if (4·8) is true in the circumstances postulated, O_1 and O_2 must agree in their estimates of the relative velocity. The fact that we can prove (5·9) from two pairs of events is easily seen to be due to our assumption that v is the same whichever pair we take, that is to say, that the relative velocity of the bodies is really uniform and does not change during the motion.

Consider now our results. If v were zero, we should have $l_1 = L_1$, $l_2 = L_2$; the observers would agree in their estimates of the lengths of the rods AB and CD. We have said nothing so far (either here or on p. 10) about the units of measurement employed by O_1 and O_2. This result shows that (4·8) can be true, only if these units are the same when they are at relative rest, so that O_1 and O_2 then agree concerning the length of any rod. This is one of the conditions which was omitted in stating the second principle (and in stating the older view). But when they are in motion, each observer finds the rod which is moving relative to him shorter in the ratio $1/\beta$ than he did when it was at rest, and shorter than the observer at rest relative to the rod finds it. Each rod appears to contract when set in motion, when observed by the observer relative to whom it moves. This is one of the best known consequences of the principles of relativity and, in its early days, excited a vast amount of discussion. But it must be noted that any difficulty raised by the apparent contraction can be referred to the discrepancy with regard to times. The contraction appears to occur because the observers have no way of estimating the length of a rod moving relative to them except by noting the positions of its ends at the same time. If they agreed as to the times when events happen, they would agree as to the length of rods. The paradox of length is therefore only another, and probably less startling, aspect of the paradox of time. It is less startling because here it seems fairly reasonable to retain the conception of the real length of the rod and to assert that one of the discrepant observers is right and the other wrong in estimating it; for we naturally prefer the judgement of the observer relative to whom the rod is at rest and who can measure its length by familiar laboratory methods. If we adopt that point of view (and I can find nothing in the fundamental ideas of relativity to render it untenable) we may simply say that neither rod really changes in length, but that each appears to contract because of the discrepancy in time judgements. Put in that form, the conclusion will probably present no difficulty whatever.

Now turn to the time intervals. We again see that if $v=0$, the time intervals agree, so that it is one of the conditions of (4·8) that the observers should use clocks adjusted to go at the same rate when at relative rest, or, if they are all using the same clock and making corrections for their distances, that the corrections should be such as to produce agreement when the observers are all at relative rest, whatever their distance from the clock. But when v is not zero, we find, as has been indicated, a discrepancy between the two observers. The relation between the time intervals when the observers are in relative motion, given by (5·12), may then be put in words, necessarily somewhat complicated, thus: The time interval between two events which consist in the passage of the same point of one system past two different points in the other is shorter when measured by an observer at rest relative to the system of which the same point is concerned in the two events than when measured by an observer at rest relative to the system of which different points are concerned in the two events. It is now perhaps rather less plausible to assert that one observer is right and the other wrong; for the events we are considering are coincidences between points in relative motion; none of these events is *in* one system or *in* the other; they are all necessarily relations between the two systems, and it is more difficult to see why the observer on one system should have a better right to pronounce about them than the observer on the other.

If we had to choose between the pronouncements of the two observers we should probably choose that of him to whom the events appeared to happen at the same place, especially if times were being estimated by the observation of a single clock corrected for distance; for the corrections of this observer would be simpler and therefore less liable to error. If we had to choose, we should say that this observer determined the real time and that the other differed from him because the time as it appears to him is not the real time but affected by error. But if the reader will think the matter out very carefully, he will find that he cannot maintain that attitude against objections without giving to one observer a preference which is not based solely on his motion relative to the material system observed. Such a preference is inconsistent with the fundamental idea of the first principle. We must at some point in our study of relativity agree to imagine a state of affairs which our ingrained conceptions tell us (though without adequate reasons) is impossible.

As a result of (5·9) and (5·12) we may now put (4·8) in a form similar to (4·3), and so complete the analogy of the two views. We again have to choose the instant from which we measure times and the points from which we measure distances so that

$$a_1 = b_1 = t_1 = T_1 = 0.$$

We shall then find by algebraical substitution that the x, X, t, T for the same event in Table II are related by the equations

$$X = \beta(x - vt) \quad \text{......................(5·13)},$$

$$T = \beta\left(t - \frac{vx}{c^2}\right) \quad \text{..................(5·13·1)},$$

or

$$x = \beta(X + vT) \quad \text{..................(5·14)},$$

$$t = \beta\left(T + \frac{vX}{c^2}\right) \quad \text{..............(5·14·1)}.$$

The resemblance to (4·3) is much more obvious than it was in comparing (4·8) with (4·2). We have merely introduced the apparent contraction into the relation between x and X, and have added a correcting term, depending on the distance, into the relation between t and T. (But, once more, if we try to regard this correcting term as on the same basis as the correction for finite velocity of signals, which is introduced in estimating times all from the same clock, and as a mere correction to this correction, we shall find ourselves in the difficulties just mentioned in face of an acute objector.) And we can now see why it is that (4·8) appears to differ so completely from (4·2), although it is a substitute for it and only adds small correcting terms. The reason is that (4·8) is really not the analogue of (4·2), but of the statement that the observers agree about times, or that $t = T$. In applications of the principle to any velocities realisable in material systems, δx^2 is very small indeed compared with $c^2 \delta t^2$; if numerical values are put in, (4·8) is almost indistinguishable from $\delta t = \delta T$.[1] In stating (4·2) our attention was concentrated on the relation between x and X, and we forgot about t and T; (4·8) states both relations at the same time and, when they are stated in that form, the relation of t and T appears predominant.

6. **The compounding of velocities.** We have now discussed how the two observers will differ in their measurements of times and lengths. We may now ask how they differ in their measurements of velocity. We suppose that there is a point E on some

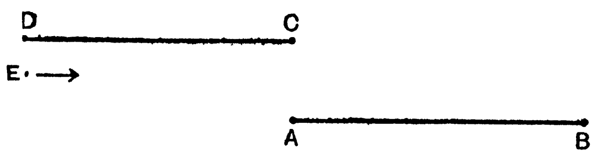

Fig. 2.

other rod, sliding along the common surface of AB and CD (Fig. 2). The relative velocity of AB and CD is again v, that of

[1] I have preferred, unlike most writers, to insist very little on the smallness of the correction. It does not seem to me any harder to *imagine* a large correction than a small one; if it is a question of *believing* that the correction is actually necessary, there is simply no evidence for either; the very small ones cannot be detected; the large ones cannot be realised.

E and AB u, that of E and CD U; we ask what is the relation between v, u, U.

According to the older view, if the directions are as in the figure,

$$u = U + v \quad \text{.......................(6·1).}$$

The problem can be solved according to the principles of relativity by the direct method we have used already; we might draw up a table of the events concerned in the measurement of u, v, U and apply to them (4·8). In so doing we should have to make the assumption that, if E is judged by O_2 to be coincident with a point X on CD, and at the same instant X is judged by O_2 to be coincident with a point x on AB, then both E and X are judged by O_1 to be coincident with x at the same instant. This is a new assumption and should be noted; but it is plausible and need not be further discussed.

However this method is cumbrous and has been used sufficiently; we are now in a position to use our previous results in order to abbreviate; accordingly we shall base our solution on (5·13), (5·14), leaving readers who think fit to work out the more direct method for themselves. Suppose (as is always necessary when using these relations) that O_1 and O_2 agree to measure times from the instant at which the points from which they measure distances are coincident. Let these points be A and C and let D be the point judged by O_2 to be coincident with E when $T = 0$. Let the distance CD measured by O_2 be L_2. Then for the event ED $X = -L_2$, $T = 0$. Consequently from (5·14) the event ED happens for O_1 when $x = -\beta L_2$ and when $t = -\beta \frac{vL_2}{c^2}$; that is to say E is at $-\beta L_2$ when $t = -\beta \frac{vL_2}{c^2}$. Let the event EC $(X = 0)$ happen for O_2 at $T = T_1$; then it happens for O_1 at $x = \beta v T_1$ where $t = \beta T_1$. For O_2 E has moved from D to C in time T_1 and its velocity is accordingly $U = L_2/T_1$. For O_1 it has moved from $-\beta L_2$ to $+\beta v T_1$ in the time from $-\beta \frac{vL_2}{c^2}$ to βT_1; consequently its velocity is

$$u = \frac{\beta (L_2 + vT_1)}{\beta \left(\frac{vL_2}{c^2} + T_1\right)}.$$

If we substitute UT_1 for L_2, we obtain

$$u = \frac{U+v}{1+\frac{Uv}{c^2}} \quad \text{......................(6·2)},$$

which by mere algebraical transformation gives

$$U = \frac{u-v}{1-\frac{uv}{c^2}} \quad \text{........................(6·3)}.$$

The transformation shows, as it should, that the relation holds for both observers equally if we take into account the different signs they give to v.

Comparing (6·2) with (6·1) we see that, according to the principle of relativity, the velocity obtained by 'compounding' two velocities in the manner contemplated is always less than that predicted by the older view. The result of such composition is especially remarkable when one of these compounded is equal to c; for if either U or v is c, u also is c, whatever is the other velocity compounded with it; no two velocities which are both less than c can, if compounded, give a velocity as great as c. These results are, of course, at variance with older ideas; once more, any difficulty there may be in imagining they could occur can be reduced immediately to the fundamental difficulty of imagining that two observers could differ concerning the simultaneity of events; and once more, there is no question of observing such anomalous results, for the necessary conditions cannot be realised. This last remark is of especial importance here; for some writers have expressed the result by saying that no *addition* of one velocity less than c to another can produce a velocity as great as c. Now usually when we speak of adding to a velocity or increasing it, we are thinking of rather different circumstances. We are thinking only of two systems, that relative to which the observer is at rest and the moving body; and we increase the velocity of that body by acting on it with some force inherent in the observer's system. Such conditions can be realised easily enough. But there is nothing in the conclusions we have reached so far which predicts anything about what will happen in such conditions; it would not be inconsistent with the principle, as stated at present, to suppose that a body might by such means be given a velocity greater than that of light relative to some other system.

The special case when one of the compounded velocities is c is of particular importance and will be the subject of the next section. Here we may note that the conclusion reached has given a name to the second principle of relativity. We have concluded that a body moving relatively to one observer with the velocity of light in vacuo is moving with the same velocity relative to any other observer who is moving with any constant velocity relative to the first. The velocity of such a body is constant for all observers (in uniform relative motion); the principle which leads to this proposition is called the principle of constant light velocity. (The velocity of light will always mean the velocity in vacuo, unless the contrary is stated; as we shall see on p. 27 the principle does not apply to the velocity in other media.)

The constancy of the velocity of light follows immediately from (4·8). For if the events considered are the passage of light between pairs of points in the two systems, $\delta x = c\delta t$ and $\delta X = c\delta T$, both sides of (4·8) are zero. Many writers indeed regard (4·8) as nothing but the algebraic expression of the constancy of the velocity of light. I think however that the passage from the latter to the former is something more than mere expression in symbols, and that it is the something more which enables many of the most important deductions to be made. But we need not stop to consider this question; everyone agrees that (4·8), together with a few minor assumptions, which are here introduced incidentally, is an adequate expression of the second principle of relativity.

7. **General form of the second principle.** A slight extension of the principle must be made before this section is ended. We have considered only the relation between measurements of lengths made by the two observers when these lengths are parallel to the relative motion and slide along each other. We have to consider how lengths in other directions, and those which, while parallel to the motion, lie a finite distance apart, are related in the measurements of the two observers. The pronouncements of the second principle on these matters are (1) that a line which appears to one observer perpendicular to the relative motion appears so also to the other observer—which means, in terms of the happening of events, that if ab and AB are lines in their own systems which

appear to O_1 and O_2 to be perpendicular to the motion, then both observers will agree that the instant at which any one point on *ab* coincides with a point on AB is also the instant at which all other points on *ab* coincide with points on AB; (2) that the distances between events both of which happen on such lines appear to be the same to both O_1 and O_2. (1) and (2) imply, among other things, that *all* dimensions parallel to the motion appear to change in accordance with (5·9), while dimensions perpendicular to the motion are the same for both observers and the same as when the systems are at relative rest.

It follows that, if rectangular axes are adopted by both observers for the description of the positions of points in their own systems, and if both place the x (or X) axis along the direction of motion, we have, in place of (5·13),

$$X = \beta (x - vt) \qquad \text{(7·1·1)},$$
$$Y = y \qquad \text{(7·1·2)},$$
$$Z = z \qquad \text{(7·1·3)},$$
$$T = \beta \left(t - \frac{vx}{c^2}\right) \qquad \text{(7·1·4)}.$$

It follows again by mere algebra that in place of (4·8) we have

$$c^2 (\delta t)^2 - (\delta x)^2 - (\delta y)^2 - (\delta z)^2 = c^2 (\delta T)^2 - (\delta X)^2 - (\delta Y)^2 - (\delta Z)^2 \qquad \text{(7·2)}.$$

There is no need to discuss here whether these additions to the principle are totally independent of the part first stated, or what is the logical connection between them. But it may be well to note that they are in complete accordance with our view that the discrepancies in respect of length are a manifestation of those in respect of time. For suppose that KL, MN are two lines in one system parallel to the relative motion, and x, y two points in the other. If x lies on KL and y on MN at some instant during the motion, they will lie on those lines at all instants; for that is what is meant by saying that KL, MN are parallel to the motion, and that is what is assumed in assuming that there can be such a thing as the motion of one body in a single direction relative to another. The distance between xy perpendicular to the motion will be the distance apart of KL, MN; the estimate of it is quite independent of any judgement where, at any instant, x is on KL

and y on MN. The principle asserts, as we expect, that in such conditions, when measurement of length is independent of measurement of time, the lengths measured by the two observers agree.

Lastly we may inquire about velocities perpendicular to the direction of relative motion. The solution of any problems that arise is easily obtained from (7·1). Thus suppose E in Fig. 2 appears to O_2 to have the velocities (U_x, U_y, U_z) parallel to perpendicular axes (X, Y, Z), while to O_1 it appears to have velocities (u_x, u_y, u_z) parallel to (x, y, z). Then, as seen by O_2, E moves from (X_0, Y_0, Z_0) to $(X_0 + U_xT, Y_0 + U_yT, Z_0 + U_zT)$ during the interval from 0 to T; consequently, as seen by O_1, it moves from

$$\{\beta X_0, Y_0, Z_0\} \text{ to } \{\beta (X_0 + U_xT + vT), Y_0 + U_yT, Z_0 + U_zT\}$$

during the interval from 0 to $\beta \left(T + \frac{vU_xT}{c^2}\right)$. Consequently

$$u_x = \frac{U_x + v}{1 + \frac{vU_x}{c^2}}, \quad u_y = \frac{U_y}{\beta \left(1 + \frac{vU_x}{c^2}\right)}, \quad u_z = \frac{U_z}{\beta \left(1 + \frac{vU_x}{c^2}\right)} \quad (7\cdot3).$$

It should be observed that the constancy of the velocity of light still obtains; for it will be found that if $U_x^2 + U_y^2 + U_z^2 = c^2$, then $u_x^2 + u_y^2 + u_z^2 = c^2$.

II. CONSEQUENCES OF THE FIRST AND SECOND PRINCIPLES

8. **The original problem.** We must now inquire what the principles we have stated will explain. We must seek to combine them in such a way that the fantastic element drops out and we are left with a statement concerning solid fact which we can test by experiment.

Everyone knows that the principles of relativity were introduced to explain certain difficulties in the theory of light propagation, and especially that raised by the Michelson-Morley experiment. But let me imagine for the moment that my reader does not know that, and let me tell him what the experiment was. A beam of light from a suitable source is split by a semi-transparent mirror into two beams which travel along two paths OA, OB at right angles. At a distance l along each of these paths the two beams fall perpendicularly on mirrors which reflect them to (or very near) their starting point, where they produce a set of interference fringes. The whole apparatus is mounted on a frame which can be rotated about a vertical axis. As the frame rotates the interference fringes do not move. My reader will at once exclaim, How obvious! Why should anyone expect them to move! If they had moved, there might have been something to explain; as it is there is nothing.

That is a point which it is important to make at the outset. The principles of relativity may ask us to imagine strange things, but they ask us to believe nothing that is in the least strange. Most of the predictions based on them are so obvious that we can hardly believe that they are not true.

The Michelson-Morley experiment is difficult to explain only when considered in combination with other experiments. There are three others which are especially important; I shall state their results very briefly in a form suited for our purpose, leaving the reader who wants to know more to turn to any treatise on Light.

The Doppler effect. If the velocity of an observer towards or away from a source of light is changed by an amount v, the apparent frequency of the light is increased or decreased in the ratio $\frac{c \pm v}{c}$, where c is the velocity of light.

Aberration. If an observer, moving along a line making an angle θ with the line along which the light from a source reaches him, changes his velocity by an amount q, then the apparent direction of the source changes through an angle $\frac{q \sin \theta}{c}$.

Fizeau experiment. A medium which, when at rest relative to the source and to the observer, appears to have the refractive index n appears to have a refractive index $n' = n\left\{1 + \frac{v}{c}\left(n - \frac{1}{n}\right)\right\}$, when it is moving with a velocity v towards the source from the observer.

There is again no difficulty in understanding any of these experiments considered alone. The difficulty arises when we try to explain them all at the same time, and especially when we try to explain them on the aether theory of light propagation as it was held in the middle of the nineteenth century. According to that theory light consists of vibrations propagated through an all-pervading medium, having some at least of the properties of an elastic solid, just as sound is propagated through air. The velocity of light through this medium is always c. If an observer is moving relatively to the medium with the velocity v, the velocity of light relative to him should be that compounded (according to the 'older view') of c and of v. The Doppler effect follows immediately from this theory. The source is emitting the same number of waves per second whatever the motion of the source and of the observer. If both are at rest in the aether, the observer receives each second as many as are sent out. If he is moving towards the source with a velocity v relative to the aether, or if the source is moving towards him with the same velocity he receives some of the waves before they would have reached him if there had been no motion; he receives more per second and the apparent frequency is increased. Again, the direction from which he appears to receive the light changes with his motion relative to the aether just as the direction from which the wind appears to come varies with our motion through the air. The Fizeau experiment is less easy to explain completely; for if the moving medium carries the aether with it, we should expect the apparent velocity of the light to change in the ratio $\dfrac{\frac{c}{n} - v}{c}$, and the refractive index in the inverse ratio which would make $n' = n\Big/\left(1 - \frac{v}{c} . n\right)$; if it does not carry the aether with it, we should expect no change. The actual change lies between these extremes, and we have to adopt the not very satisfactory idea that the medium confers on the aether only part of its velocity, this part varying with the refractive index. Lastly the Michelson-Morley experiment is a great stumbling block. For the fact that the earth's motion relative to the sun produces aberration in the light received from stars seems to show that the earth must be moving through the aether. In some position of the frame carrying the mirrors, OA must lie in the direction of this motion, OB across it; when the frame is turned through a right angle these positions must be interchanged In the first position, the light traverses the distance l with velocity $c + v$ in one direction and with velocity $c - v$ in the other; the total time it takes is

$$\frac{l}{c+v} + \frac{l}{c-v} = \frac{2lc}{c^2 - v^2};$$

on the other hand it traverses the distance l along OB with velocity c, there and back, and its time is $\frac{2l}{c}$. The first time is the longer; the light travelling along OA will be delayed. When OA and OB are interchanged, the light along OB will be delayed; consequently the times of arrival of the two beams at the place of interference will be changed by the rotation of the apparatus, and the interference fringes correspondingly displaced.

9. **Solution of the original problem.** If we are to bring any of these experiments within the range of the principles of relativity as they have been expounded so far, we have to assume, or at least to imagine, that the velocity of light can be measured in the same way as the velocity of a material body, namely by noting the positions of the 'light' at different times. For if we do not make that assumption, there is no point of contact between the imaginary

experiments, on which we have based the conclusions drawn from the principles, and the observations that we have to explain. Probably most readers will feel no difficulty about this assumption, but the necessity for it should be noted carefully. In all the experiments we have to explain, the velocity of the light is actually measured by interference or spectroscopic methods, which, so far from being measures of the distance something moves in a given time, do not involve any moving parts at all. There are ways of measuring the velocity of light which approximate more nearly to those applicable to a material body, e.g. the rotating mirror and toothed wheel methods; but even in these the time measured is not that in which light moves from one place to another, but that in which it returns to the same place after reflection. The nearest attempt to measuring the velocity of light by a 'direct' method is that which depends on the occultation of the satellites of Jupiter; but even here there are many assumptions involved which might possibly be doubted. Of course we have very good reason to believe that all these methods do truly measure the velocity of light; but it must not be forgotten that the belief is not absolutely certain, and that it would be possible to remove all the discrepancies we are discussing by simply denying that our present theory of spectroscopy and interference is correct. Nobody has suggested such a solution, which would doubtless introduce more difficulties than it removes[1]. It is only because that theory is unchallenged, and because we have not the smallest doubt that interference does measure the velocity and does give results that would agree with the 'direct' method if only it could be carried out, that we can calmly proceed to interpret the Michelson-Morley experiment, in which all relevant material bodies are at relative rest, in terms of

[1] It is worth noting that attempts are made from time to time to show that the classical wave theory or Maxwell's theory do not really predict the shift of the interference fringes which Michelson and Morley expected and failed to find; and that, therefore, there is nothing for relativity to explain. Such attempts make use of the doubts suggested in the text, namely whether in these circumstances interference fringes are an adequate measure of velocity. But such inquiries really lead nowhere. Theories very seldom predict quite definitely any law unknown at the time of their formulation; and neither Huyghens nor Maxwell carried his theory into the particular detail necessary for the unambiguous treatment of the famous experiment. The only question that can be reasonably asked is whether these theories, developed in the obvious and natural manner, do predict a positive result; the fact that *before* the experiment was tried, nobody doubted that they did so predict is a sufficient answer.

principles which necessarily assume that the bodies considered are in relative motion.

Now let us take the experiments in turn, retaining for the present the conceptions and terms of the aether theory. The Michelson-Morley result, of course, follows at once. Even if OA and OB are moving differently relative to the aether, an observer fixed relative to OA will still always find for the velocity of light along OA the same value as an observer fixed relative to OB finds for the velocity along OB; if OA and OB (which are relatively at rest) are equally long, the light will always take the same time to travel along each of them and back again. No displacement of the fringes is therefore to be expected.

The Fizeau experiment appears as a direct consequence of (6·3). Let n be the refractive index of the medium determined by an observer O_1 at rest relative to it. Then, on the Huyghens' theory, the velocity of light relative to him is c/n. Consequently the velocity of light between two points in the medium, determined by an observer O_2 moving relatively to the medium with a velocity v (in the direction opposite to that of the light), is by (6·3)

$$U = \frac{c/n - v}{1 - \dfrac{v}{nc}} \quad \text{......................(9·1).}$$

But $c/U = n'$, the refractive index observed by O_2. Hence

$$n' = n\left(1 - \frac{v}{nc}\right)\Big/\left(1 - \frac{vn}{c}\right) \quad \text{..............(9·2),}$$

which, if $\frac{v}{c}$ is so small that we may neglect terms in $\frac{v^2}{c^2}$, becomes

$$n' = n\left\{1 + \frac{v}{c}\left(n - \frac{1}{n}\right)\right\} \text{.................(9·3).}$$

This simple explanation of the Fizeau experiment is a very notable achievement of the doctrine of relativity, although the result was known, and had been explained in some measure in terms of electromagnetic theory, before the advent of those principles. It should be pointed out, however, that in applying the first principle here, we are assuming that the motion of the observer relative to the source of light is immaterial. For the supposed observer at rest relative to the medium is moving relative to the source; if his velocity relative to the source had been important, we could not have assumed that he would find the refractive index to be n, the normal value. It can be tried directly whether the motion of the source affects n for velocities of the magnitude of those used in the Fizeau experiment; in fact it is tried when the Doppler effect is measured first with a prism and then with a grating spectroscope. The agreement of those two instruments shows that the refractive index of the prism for light of given frequency is the same whether or no the source is in motion relative to the prism. Of course the result is to be expected on the aether theory; but apart from that theory, the assumption of it in the application of the first principle was precarious. In dealing with the other experiments, this question does not arise, for the only thing which we assume to be unchanged during a change of the motion of the source relative to the observer is the velocity of light in vacuo; and the constancy of that velocity follows from the second principle and not from the first.

In explaining aberration we have to consider the effect of compounding the velocity of light with the velocity of the observer in a different direction. Suppose that an observer O_2 at rest in the aether, taking his x axis along the direction of relative motion of O_1, sees a source of light in a direction making an angle Θ with that direction. Then for O_2,

$$U_x = c\cos\Theta, \quad U_y = c\sin\Theta, \quad U_z = 0.$$

Consequently, O_1 sees the light in a direction θ with x, where

$$u_x = c\cos\theta, \quad \sqrt{u_y^2 + u_z^2} = c\sin\theta,$$

and u_x, u_y, u_z are given by (7·3). Hence we find easily

$$\sin(\Theta - \theta) = \frac{\frac{q}{c}\sin\Theta - \left\{\left(1 - \frac{q^2}{c^2}\right)^{\frac{1}{2}} - 1\right\}\cos\Theta\sin\Theta}{1 + \frac{q\cos\Theta}{c}} \quad (9·4).$$

If $\frac{q}{c}$ is so small, that terms in $\frac{q^2}{c^2}$ can be neglected, (9·4) becomes

$$\Theta - \theta = \frac{q}{c}\sin\Theta \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots (9·5).$$

It might seem at first sight as if the Doppler effect, which is ordinarily interpreted as a manifestation of a difference in the velocity of the observer or the source relative to the disturbance, could not be reconciled with the conclusion that all observers have the same velocity relative to the light and that the velocity of the source is immaterial. But, in relativity as in some other subjects, obvious arguments are usually unsound and 'it stands to reason' is a sure sign of fallacy. Let us consider the matter more closely. In Fig. 3 let E_1, E_2 be two successive crests of the wave; AB, CD

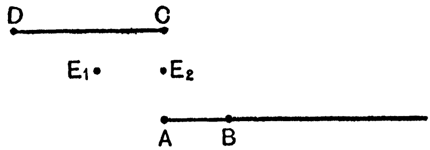

Fig. 3.

are as usual the two observers' systems of which CD is the aether. Choose as zero of time the instant of the event E_2AC. Let B be the point on AB which coincides with E_1 and C at the same time. Let the event E_1BC happen at time T_1 to O_2 and at time t_1 to O_1. Then, by definition, T_1 is the period and $1/T_1$ the frequency measured by O_2. Since C is moving with velocity v relative to O_1, $AB = vt_1$. But E_2 is moving relative to AB with velocity c; therefore E_2 arrived at B when $t = \frac{vt_1}{c}$. Consequently the period measured by O_1, being the difference between the times at which E_2 and E_1 arrive at B, is $t_1 - \frac{vt_1}{c} = t_1\left(1 - \frac{v}{c}\right)$. Hence the frequency appears greater to O_1 than to O_2 in the ratio $\frac{c}{c-v} \cdot \frac{T_1}{t_1}$. But T_1 is the interval between two events both of which happen at C; therefore, by (5·12), $T/t_1 = \frac{1}{\beta}$. Our final result is therefore that the frequency is changed in the ratio $\frac{1}{\beta} \cdot \frac{c}{c-v}$. It differs from that deduced on the 'older view' only by terms which involve $\frac{v^2}{c^2}$; and since these terms in any actual case are so small that no experiment can tell the difference, relativity agrees with the older view and with observation[1].

[1] For the benefit of those who know a little about the subject it may be worth while to point out at once that this factor $1/\beta$ is *not* that concerned in the disputed 'Einstein shift' in the sun's spectrum. The correction introduced by it is even smaller than that shift. See p. 103.

10. The pure relativity solution. But the explanation of the experiments does not depend upon the aether theory. Let us now forget for the moment all about that theory and discuss the matter again, remembering only that, if we are to forget it at all, we must forget it completely and make no use of conceptions. There is now no question of explaining the Michelson-Morley experiment; there is nothing to explain; there is no moving system; all the portions of the apparatus which are parts of the system investigated are at relative rest. Again nothing need be changed in our explanation of the Fizeau experiment; for it will be seen that no reference was made to the aether in explaining it. The aether enters only as an explanation of the experimentally established fact that the refractive index is independent of the velocity of the source; we now base our application of the second principle directly on the fact.

But with the Doppler effect and aberration a curious difficulty seems to enter which requires a moment's consideration. It is known experimentally that the Doppler effect is determined by the motion of the source as much as by that of the observer; indeed the magnitude of the effect is proportional to the relative velocity of source and observer. On the other hand, aberration does not depend on the velocity of the source. For there are stars which we have every reason to believe are rotating round axes nearly perpendicular to the line of vision. The central parts of these stars are moving across the line of vision, while the poles are not (or rather the poles and the centre are moving with different velocities across the line of vision). Consequently, if aberration depended on relative velocity of observer and source, the centre should be displaced by a very appreciable amount relative to the poles, and the star should appear, not as a point, but as a line so drawn out to be invisible. The aether theory, at least on the older view, by dividing the relative velocity of source and observer into two parts, makes a clear distinction between the two cases. The argument on p. 26 shows that the Doppler effect must depend on both parts, the velocity of the source relative to the aether and that of the observer relative to the aether. But from the same argument it appears also that aberration should not depend on the first part, because the direction of motion of the light, once it has left the source, is independent of the motion of the source; an observer at rest in the aether should perceive the real direction of the light, whatever the motion of the source; it is only the second part that is effective. But if we leave the aether out of account, we cannot divide the relative velocity of source and observer into two parts. If this velocity is effective at all, the whole of it must be effective; and it would seem therefore that if the Doppler effect depends on this velocity, so must aberration—a conclusion contradicted, as we have seen, by experiment.

But relativity does not tell us 'on what the Doppler effect or aberration depends'; it merely tells us what will be the difference of experience of two observers. And the only relative velocity it introduces is the relative velocity of these observers, together with the velocity of light relative to either of them. The velocity of the observers relative to the source does not enter into the explanation of aberration at all; and it enters into the explanation of the Doppler effect only in the proposition that it is irrelevant, because the refractive index is independent of this relative velocity. The apparent difficulty arises because we have interpreted the answer given by relativity in terms of ideas which it does not consider at all. We could leave out all reference to the aether on p. 30, and the explanation would stand unaltered.

But since it may seem that this explanation of the matter is unsatisfactory and that relativity ought to give some answer to the question 'on what velocity the Doppler effect and aberration depend', we will look into the matter a little more closely. The question implies that there is some 'real' frequency and some 'real' direction of the light, and that the difference from this real frequency and real direction is the real Doppler effect or aberration which must depend on some velocity. The appearance of the word 'real' is always a danger signal of confusion of thought; but without any deep inquiry into the fallacies which the word may here conceal, we may see very easily that there is a great difference between the 'real frequency' and the 'real direction' of the light which is quite sufficient to account for the difference between them that we are discussing. It is that we can determine the real frequency while we cannot determine the real direction. The real frequency is the frequency as observed by an observer at rest relative to the source, whether or no he is at rest relative to the aether; it is something that can be observed directly. But the real direction is not that observed by an observer who is (and always has been) at rest relative to the star; for the direction in which he would see the star, if he were now occupying the position of our observer, would be that in which our observer would have seen it if, since the light which he now receives was emitted, he had been moving relative to his present position with the negative of the velocity of the star relative to him during the interval. Or, to put the matter in a less complicated manner, the real direction of the light seen by the observer is not the real direction of the star from him at the moment of observation; it is the real direction from him of the star at the moment it emitted the light which he now receives. Since we do not in general know what has been the relative motion of star and observer during the passage of the light from one to the other, we do not know and cannot know the real direction. All we know is the direction in which the light is seen. If we mean by the 'real aberration' the difference between the direction in which the light is seen and the 'real direction', and ask whether it depends on the relative velocity of source and observer, the only possible answer to the question is that we do not and cannot know on what it depends, because we do not know what it is. On the other hand, if we mean by the real Doppler effect the difference between the real and observed frequency, we do know what that is and can show experimentally that it is proportional to the relative velocity of observer and source.

And here it may be convenient to notice parenthetically a set of objections to the doctrines of relativity which have received much attention. They are based on the questions concerning 'absolute rotation' which were first asked by Newton about the water rotating in the bucket. The matters have really nothing to do with relativity; they fall wholly outside the special theory, which considers only uniform translation; and so far as they fall within the general theory, it gives to the questions raised answers indistinguishable experimentally from those of the Newtonian theory. Such difficulties as there are concern theories of cosmogony rather than theories of motion or forces. The main question, put in the widest form, is why a plane which does not rotate relative to the fixed stars is of such great importance in purely terrestrial experiments; it is not answered by saying that this axis is at absolute rest, for we may still ask why then the fixed stars are at absolute rest; it can only be answered by tracing back the history of the earth and the stars to the time when they were in dynamical connection. That problem is not one for relativity to solve.

The matter is mentioned here because the supposed inconsistency with relativity arises in the same way as that just considered. The conception is introduced of a super-human observer who can measure directly magnitudes (real velocity of rotation or real direction of light) which actual observers can only determine indirectly and by the aid of some theory. This conception is inconsistent with relativity, which cannot therefore answer questions in which it is involved. If the reader, in considering such objections which he will find discussed elsewhere, will resolutely refuse to ask questions involving such a conception, he will find that the objections cannot be raised. But some resolution is needed; for the conception is very deeply ingrained in most minds; it is the conception on which are based many of the classical problems of metaphysical reality. Of course it may be urged that doctrines inconsistent with such a conception must be false; but that is a question beyond our province; all that we must insist is that, in this matter, relativity, though possibly false, is self-consistent.

11. **Extension of the principles.** Most of the chief theories of science were invented primarily to explain some definite group of known laws; most of them have gone beyond their primary purpose and explained also some laws which, though known at the time of their invention, were not taken into account in the first formulation of the theory; some of them have also explained in advance, or predicted, laws which were not already known. These advances of the theory have always and necessarily involved some modification of it, but the change has been so slight or so obvious that there has never been any hesitation in regarding the original and the modified theory as substantially the same. In these respects the principles of relativity resemble the chief theories of science. The laws of the propagation of light were their original concern; and those laws alone are all that they will explain in the form in which they have been stated

hitherto[1]. But a very slight modification of them will bring within their scope an entirely new, and highly important branch of physics, namely that which appears in our text-books as 'Electricity and Magnetism.'

Indeed it was necessary that they should be extended to that branch. For we believe that the propagation of light is fundamentally an electromagnetic process; if the principles of relativity had been capable of resolving the difficulties that we have discussed only when light was regarded from the standpoint of the corpuscular theory (for that has essentially been our standpoint so far), and not when regarded from the standpoint of the electromagnetic theory, then they would have been quite unacceptable. Further, the laws of electromagnetism include some propositions which clearly ought to fall within the scope of relativity. They do predict in some measure what will be the effect of setting in motion relative to the observer the system that he observes; they deal with the properties of moving charges and moving magnets. Indeed some progress had already been made in combining these laws of moving systems with the electromagnetic laws in such a way as to explain the optical phenomena we have just been discussing. The Fizeau effect had been explained completely in this manner and, rather less satisfactorily, the Michelson-Morley experiment. Again these are purely electrical experiments, such as that of Trouton, designed, like the Michelson-Morley experiment, in the hope of detecting the earth's motion through the aether and, like it, yielding a negative result. All these features of electromagnetism impel us to endeavour to bring it within the range of our doctrines.

But with our principles in this present form we can make no progress. Our principles are concerned with systems in which the only phenomenon is motion and the only thing that happens is that something moves; but electromagnetism, though concerned with motion, is concerned also with phenomena that are not motion; indeed one important branch of it, electrostatics, is concerned wholly with bodies all of which are at relative rest. Again the first principle in its present form would seem only to tell

[1] I do not mean to imply that the form in which they are stated here is precisely their original form; for that is not true. But I have tried to pick out from the original form those elements which were strictly necessary in the first instance.

us that, if we set an observer with all his apparatus in motion relative to some other observer or to something which is not part of his system, the phenomena he observes will be unaltered. But this proposition, though doubtless true, does not actually help us; the information which is actually required in order that the electromagnetic properties of moving systems may be brought within the scope of relativity concerns the effect of setting the observer with some of his observing instruments in motion relative to the charges, currents and magnets that he observes. There can be no doubt that these latter form part of his system, in the sense of p. 5, and that his observations will be affected by his motion relative to them, for it is precisely the effect of that motion that we want to determine.

It is necessary therefore to extend somehow both the first principle and the second. The first principle is extended by announcing what remains unchanged when an observer changes his motion relative to part of the system which he is observing. His observations are bound to change with his motion; so much is implied by the mere statement that he is moving relative to that part. The principle now asserts that, though his observations may change, the laws which he deduces from his observations, or—it would be better to say—the form of those laws, is unchanged. This distinction between observations and the laws based on them needs careful consideration; a full discussion of the matter, though very interesting, is beyond our present scope; a brief and incomplete statement must suffice.

The fundamental idea of the first principle is that, of observers in relative motion, one is as good as another; there is no question of the observations of one being true and real and those of the other misleading and merely apparent. Good or true or real here means fulfilling the aims of scientific inquiry—it may mean more, but at least it means that. Part again at least of those aims is the reduction of complex experience to laws; the fewer in number and the simpler in character those laws are, the better the aims are fulfilled. Now there are involved in most physical laws the values of some magnitudes or other; the typical law is one which asserts a numerical relation between the values of two or more magnitudes; the law is identified by that relation and its simplicity is determined by the character of that relation. (We shall have to modify this statement later; but it will serve for the present. See p. 59.) A law is not materially altered in character or simplicity if the values of the magnitudes change, so long as the nature of the relation between them is unaltered. Thus, if a re-examination of the law relating the volume and mass of silver led to the belief that the density of silver is 10·9 and not, as we now believe, 10·5, we should not regard the law as altered materially for the primary purposes of scientific inquiry; but we should regard it as altered, if we found that the mass was not proportional to the volume and therefore that there was no such thing as density. The numerical value of the density is really immaterial to pure science (though it may be important in its applications); what is important is that there is such a thing as density and a law defining it.

Accordingly whatever values an observer assigns to the magnitudes he measures, his observations, for the purpose of science, will be as good and as true and as real as those of another, who assigns different values to the same magnitudes, so long as both find the same relations, characteristic of a law, between those values. So long as one finds proportionality where the other finds it, a square law where the other finds it, and so on, there will be no reason to prefer the observations of one observer to another, however much they may differ in the values they assign to magnitudes. It may be well to note (although the matter will not actually arise) that this indifference extends even to the values assigned to 'universal constants'. We have no particular affection for $4{\cdot}77 \times 10^{-10}$ as the charge on an electron; we have no reason to prefer an observer who finds that value to one who finds $5{\cdot}77 \times 10^{10}$, so long as both find there is such a thing as the charge on an electron, and that it is a universal constant throughout their own observations. The remark is important, because various writers (including Einstein himself) have expressed themselves occasionally as if the second principle were a logical consequence of the first, and as if the constancy of the velocity of light (being a universal constant to any one observer) must have the same numerical value for all observers, in virtue of the fact that all find the same laws. I am not sure that any serious writer has meant to maintain that position, but it is well to object formally to it. In order to be consistent with its other applications, all the first principle has to assert regarding this velocity is that each observer will find that there is a velocity of light which is constant for him; it need not assert that they will all find the same numerical value for that velocity. The velocity of light should be regarded as a special case which is dealt with by the second principle.

To sum up then, the first principle now asserts that two observers in uniform relative motion, even if they have different motions relative to the system they are observing, will always be able to interpret their observations in laws of the same form, although they will differ concerning the numerical values of the magnitudes involved in those laws.

We now turn to the second principle. The extension that we make here is that the propositions which have been stated about 'events', carefully defined as meaning the coincidence of points in the relatively moving systems, are true of all events in the ordinary and wider sense; whatever the nature of the event may be, if O_1 observes that any event happens at (x, y, z, t), then O_2 observes that it happens at (X, Y, Z, T), where the relation between (x, y, z, t) and (X, Y, Z, T) is (7·1), the convention being retained that the origin of time is the instant at which the origins of distance coincide. This new part of the principle follows from the old if it is assumed that if O_1 and O_2 observe respectively that the event P (not being a coincidence of points) happens at the same distance from them and at the same time as the two events Q_1 and Q_2 (which are coincidences of points), then they will agree that Q_1 and Q_2 happen at the same time and place. Of course this assumption is very plausible, but with relativity plausibility is no test of truth; we have denied assumptions just as plausible. It would not have been incomprehensible if principles so essentially concerned with motion had recognised a great difference between events according as motion is or is not concerned in them.

One consequence of the assumption may be mentioned. If O_1 and O_2 are determining times by local clocks (see p. 13), each of O_1's clocks must be events which, for O_1, happen at the same place. Although those events may not be coincidences of moving points, we may now extend to them the result expressed in (5·12).

Accordingly we conclude that O_2 judges that in virtue of the motion all O_1's clocks are going slow in the ratio $1/\beta$, and O_1 judges the same about O_2's clocks. This proposition is introduced, because it is included in almost all expositions of relativity; but it does not seem to me really a very accurate or illuminating way of expressing the meaning of (5·12).

Again, some writers seem to regard this new assumption as a necessary consequence of the second principle and not as an extension of it, suggesting that all the phenomena studied in physics can actually be reduced to coincidences of moving points. They draw attention, for example, to the use of a coincidence between a spot of light and the divisions of a graduated scale to measure many magnitudes, e.g. electric currents. But of course such suggestions are absolutely false and would not appeal to anyone actually familiar with experiment. The coincidence of the galvanometer spot with the division on the scale measures a current only if many conditions are fulfilled, if the instrument is constructed in a certain manner and if it is connected to the rest of the circuit in a certain way. And these conditions cannot be described in terms of coincidences of points; no coincidence would be altered if we replaced iron by wood and copper by glass; but the deflection of the galvanometer would be very greatly affected by the change. We need a definitely separate assumption if we are to extend the second principle to events that are not coincidences.

12. **The laws of electromagnetism.** Let us now consider how we might apply these principles to electromagnetic laws, and let us start from Coulomb's inverse square law. O_1, with whom we will identify ourselves for the moment, is at rest relative to the point charge A; he has another charge B with which he explores the electric field of A. He measures the force which A exerts on B, and records in his note book the position of B and the time corresponding to this measurement. Of course he finds the inverse square law. O_2, moving relatively to A and to O_1, carries out exactly the same experiments and records the result of them in exactly the same way. According to our extended first principle, we know that O_2 must also find the inverse square law; but we do not know what value he will find for the constant in it, i.e. for the charge on A; that is one of

the things we seek to deduce from our principles. For this purpose we must imagine O_2 conducting other experiments. We know that a moving charge is equivalent to a current and produces a magnetic field. We must imagine O_2 making magnetic as well as electric measurements. O_2 will find a magnetic field, and this field must be related to his measurements of the charge on A and its velocity relative to him by Ampère's law. Again, the field he finds will be changing because, though the velocity of the charge relative to him is constant, its distance from him is changing. Now in virtue of Faraday's law of induction a changing magnetic field is associated with an electric field such that the E.M.F. round any circuit is proportional to the change of the magnetic flux through it. O_2 has already measured the electric field and hence, if we assume we know the results of his measurements, a condition is imposed upon the magnetic field which he will find. We have therefore a relation which, in virtue of our first principle, O_2 must find between the electric and magnetic fields, the distance and the time; by suitable calculation we can find out what that relation must be without knowing what numerical values he attributes to the fields at any given time and place.

Now we make use of the second principle. The occurrence of a given value for the electric and magnetic field is an event. According to the principle the events which for O_2 happen at (X, Y, Z, T) are the same events as those which happen for O_1 at (x, y, z, t). The events are not the same, we suspect, in the sense that O_1 and O_2 find the same values for the electric and magnetic field; if that were so, all this elaborate argument would be unnecessary. They must be the same in the sense that one is determined wholly by the other; that there is some relation between the different values, which is the same for all events, and enables us to correlate an event for O_1 with an event for O_2. For that is just what we mean when we say that two observers observe the same event; we do not mean that they both observe exactly the same thing, for they may view it from different distances or under different conditions; what we do mean is that everything in one observer's observations can be correlated in a simple and unique way with something in the other's. Accordingly, if our principles are true, O_1 by comparing with his measurements those of O_2 which represent the same event, must be able to find a simple unique relation between his measurements of the electric field (for he perceives no magnetic field) and O_2's measurements of the electric field and magnetic field. The necessity for such a relation, together with that between O_2's electric and magnetic fields, may enable us to determine exactly what the relation is, and thus to find (for the first time in this argument) a numerical relation between O_2's measurements and our own, which will solve the problem of what change will be made by setting in motion relative to the observer the point charge we have been investigating.

We *may* be able to attain this result; we cannot be sure until we work out O_2's relations and try to fit them to our own. As a matter of fact we shall in this instance succeed. We should succeed similarly if we had been observing, not merely a point charge, but any distribution of electric and magnetic fields which obey (as all fields do) Ampère's, Faraday's and Coulomb's laws. We should always succeed, by following out the argument sketched, in showing that there must be one and only one relation between O_2's measurements of those fields and our own, if the conditions are to be satisfied (1) that the laws are valid for O_2 as well as for ourselves, and (2) that the same events, consisting of simply and uniquely related values for the fields, are those which occur at times and distances given by (7·1).

But, it might be thought from this account of the matter, the argument will be extraordinarily complex. On the contrary, as it is usually expressed, it is wonderfully simple. The way it is usually put is this. Coulomb's law is expressed by Poisson's equation

$$\frac{de_x}{dx} + \frac{de_y}{dy} + \frac{de_z}{dz} = 4\pi\rho \quad \text{..............(12·1)};$$

Ampère's and Faraday's laws by Maxwell's equations

$$\left(4\pi\rho u_x + \frac{de_x}{dt}\right) = \frac{dh_z}{dy} - \frac{dh_y}{dz}, \text{ and so on ......(12·2)};$$

$$\frac{1}{c^2}\frac{dh_x}{dt} = \frac{de_y}{dz} - \frac{de_z}{dy}, \text{ and so on ......(12·3)};$$

the absence of real magnetic poles by

$$\frac{dh_x}{dx} + \frac{dh_y}{dy} + \frac{dh_z}{dz} = 0 \quad \text{.................(12·4)}.$$

Here (e_x, e_y, e_z) are the components of the electric field, (h_x, h_y, h_z) those of the magnetic field, ρ the density of electric charge, and (u_x, u_y, u_z) the components of its velocity.

In these equations we now replace (x, y, z, t) by (X, Y, Z, T) according to (7·1) and inquire whether we can find any quantities (E_x, E_y, E_z), (H_x, H_y, H_z), P, (U_x, U_y, U_z) such that if they are now substituted for the e's, h's, ρ, u's, the equations are identically true. That is a relatively simple algebraical question to answer. The reader will find by mere substitution[1] that the following quantities solve the problem:

$$E_x = e_x \quad \ldots\ldots\ldots\ldots\ldots(12{\cdot}5{\cdot}1), \qquad H_x = h_x \ldots\ldots\ldots\ldots\ldots(12{\cdot}6{\cdot}1),$$

$$E_y = \beta\left(e_y - \frac{v}{c^2}h_z\right)\ldots(12{\cdot}5{\cdot}2), \qquad H_y = \beta\,(h_y + ve_z)\ldots(12{\cdot}6{\cdot}2),$$

$$E_z = \beta\left(e_z + \frac{v}{c^2}h_y\right)\ldots(12{\cdot}5{\cdot}3), \qquad H_z = \beta\,(h_z - ve_y)\ldots(12{\cdot}6{\cdot}3).$$

$$\mathsf{P} = \beta\left(1 - \frac{vu_z}{c^2}\right)\rho \quad \ldots\ldots\ldots\ldots\ldots\ldots(12{\cdot}7),$$

while the U's are related to the u's by (7·3).

We say then that the values denoted by capitals are those measured by O_2 while the values for the same magnitudes denoted by small letters are those measured by O_1. The motion will convert the fields, density of charge and velocities measured denoted by the small letters into those denoted by the large.

We can easily obtain the corresponding relation for the total charge of any body, q or Q. For $q = \int \rho\, d\omega$ and $Q = \int \mathsf{P}\, d\Omega$, where ω,

[1] It should be observed that all the differential coefficients are 'partial' and refer to the change of one variable at a time; so that

$$\frac{d}{dX} = \left(\frac{d}{dx}\cdot\frac{dx}{dX} + \frac{d}{dt}\cdot\frac{dt}{dX}\right), \text{ when } T \text{ is constant,}$$
$$= \beta\left(\frac{d}{dx} + \frac{v}{c^2}\cdot\frac{d}{dt}\right);$$
$$\frac{d}{dT} = \left(\frac{d}{dx}\cdot\frac{dx}{dT} + \frac{d}{dt}\cdot\frac{dt}{dT}\right), \text{ when } X \text{ is constant,}$$
$$= \beta\left(v\frac{d}{dx} + \frac{d}{dt}\right).$$

Note also that the equations (12·5, 12·6, 12·7) are self-consistent; for if we change our standpoint, identify ourselves with O_2, and ask how to obtain O_1's values from O_2's, we may solve the equations and obtain another set in which small and capital letters are interchanged and the sign of v reversed.

Ω are the volumes of the body. Since O_2 observes the body contracted in the ratio $1/\beta$ along x and otherwise unaltered $d\Omega = d\omega/\beta$. Further, if the charge is at rest relative to O_1 (as we are supposing), $u_x = 0$. Hence

$$Q = \int \beta\rho \cdot \frac{d\omega}{\beta} = \int \rho\, d\omega = q \quad \text{..............(12·8)}.$$

The two observers agree numerically concerning the charge on any body.

The process seems so magically simple compared with that previously described that we shall be inclined to inquire whether it is really the same and if no additional assumptions have been introduced. That is a difficult question to answer certainly. Whenever we express laws by differential equations, we alter their meaning somewhat, for we can only compare the equations with actual measurements when they have been integrated. Thus, if we proposed to compare Poisson's equation with experiment by measuring e_x, e_y, e_z, ρ, we should find that we should have to integrate the equation over a finite volume before the comparison could be made. Differential equations always really express a theory, from which laws can be deduced by integration. But if we overlook this difference, which is inherent in all mathematical physics, I do not think any other remains; the mathematical argument could be related step by step to the physical argument as it was stated originally.

Let us consider then what we have proved. First that, in their application to electromagnetic laws as expressed by the familiar equations, the first and second principles of relativity are consistent with each other. The translation of O_2's observations into O_1's by means of the second principle (as extended) is consistent with the assumption made by the first (as extended) that those laws are valid for both observers. If and only if the principles of relativity are true, Maxwell's equations are valid for all observers and not only for an observer fixed in the aether. We have found again from those equations that the velocity of light is the same for all observers (since c appears unaltered in the transformed equations); we have also found again (7·3) for the composition of velocities. It follows that if light can be explained as an electromagnetic phenomenon from the standpoint of O_1, it can always be so explained from the standpoint of O_2, that the Michelson-Morley result (which depends on the constancy of c) must be obtained, together with the formulae for aberration and the Doppler effect (which depend on (7·3)).

Secondly, we have found a numerical relation between O_1's values and O_2's. The older theory, which regarded the equations as true for an observer fixed in the aether, arrived at similar formulae for the effect of setting charges and magnets in motion relative to the aether, by assuming (for example) that a charge moving through the aether was equivalent to a certain electric current obeying Ampère's law. But these formulae were not quite the same as ours; they differ from them by the omission of the factor β and the term in $\frac{vu_x}{c^2}$; the difference made by this omission is always too small to be detected by experiment. Hence the experiments, such as that of Rowland or Roentgen, which were held to establish the equivalence of a moving charge and a current, support equally the formulae (12·5) and (12·6). But our view of the matter is rather different and introduces a new and satisfactory unity. According to the older view, electrostatics and electrodynamics were two separate and independent studies; a charge had certain properties at rest, and certain others in motion; at rest it produced an electric field, in motion a magnetic field, which were two utterly different things. But our view is that they are merely aspects of the same thing. When the observer, previously at rest with regard to the charge, is set in motion relatively to it (or when the charge is set in motion relatively to him) he loses a part of its electric field, which is the exact equivalent to the magnetic field which he gains. We cannot distinguish between electrostatics, of which the laws are asserted by Poisson's equation, and electrodynamics, of which the laws are asserted by Maxwell's equations; they are essentially parts of a whole and lose much of their meaning when separated.

The form in which we have discussed Maxwell's equations is that applicable to a vacuum in which there are no material dielectrics or magnetisable bodies. And this, according to electronic theory, is the only form in which they are really valid. Their modification to suit material media by the introduction of K and μ is, according to that theory, merely a summary and not wholly a satisfactory way of taking into account the presence of the charged bodies of which material media consist; the more scientific way is to keep the equations unaltered and to introduce the charges in the form of ρ's and u's. It is in this manner that we must apply relativity to any experiments, such as that of Fizeau, which involve the presence of material media; we must first explain them for some observer on the basis of the electronic theory and then consider how the observations of any other observer will be affected by his motion relative to the first. The explanation of the Fizeau experiment depends, on this view, on the electronic theory of refraction; we consider how the fields due to the electrons which vibrate under the incident light will be changed by the motion, and how, in virtue of that change, their effect on the incident light will be changed. This mode of looking at the matter leads to the right result, given by (9·3). Indeed it led to the right result (apart from inappreciable terms in v^2/c^2) without relativity, and on the older theory which applied Maxwell's equations only to observers fixed in the aether; according to that theory the Fizeau result is an expression of the magnetic field set up by the motion through the aether of the electrons in the refracting medium. (See Chap. XIV of the *first* edition of this book.) All the crucial optical experiments can be explained on the electromagnetic theory of light as well as on the corpuscular; and the purely electrical but closely related experiments, to which brief reference has been made, can also be explained.

But it is to be noted that we can only bring within the range of our theory electromagnetic experiments which can be completely explained in terms of the electronic theory. Thus, we cannot at present bring Ohm's law within its range; we cannot say exactly how the resistance of a conductor would differ when measured by two observers in relative motion; we cannot even be sure that an observer moving relative to a conductor would find Ohm's law strictly true; we do not know whether the first principle applies to that law. In order to have that knowledge we must be able to express conduction in terms of the motion of electrons, and we cannot yet do that quite satisfactorily. But it is easy to see that if Ohm's law does remain true, the change in resistance can only depend on terms in v^2/c^2, which are quite inappreciable experimentally.

13. Mass and energy. It remains to draw some conclusions concerning the mass and energy of a moving charged particle which, though throwing no additional light on the fundamentals of relativity, are very important for it, since they can be confirmed by experiment and thus used to support the principles.

O_1 and O_2 will determine the mass (m, M) of a particle with charge (q, Q) by finding what values will fit the equations

$$\frac{d}{dt}(mu_x) = qe_x,\ \frac{d}{dt}(mu_y) = qe_y,\ \frac{d}{dt}(mu_z) = qe_z \quad (13 \cdot 1 \cdot 1)$$

$$\frac{d}{dT}(MU_x) = QE_x,\ \frac{d}{dT}(MU_y) = QE_y,\ \frac{d}{dT}(MU_z) = QE_z \quad \ldots\ldots(13 \cdot 1 \cdot 2).$$

Suppose O_1 is at rest relative to the particle; let m_0 be the mass that he finds. Then $(u_x, u_y, u_z) = 0$, though $\frac{du_x}{dt}, \frac{du_y}{dt}, \frac{du_z}{dt}$ are finite. Also, if (as we suppose) there is no magnetic field due to other charges moving relatively to O_1, $(h_x, h_y, h_z) = 0$. Taking these conditions into effect, and using (12·5), (12·6), (12·7) we find that (13·1·1) is algebraically equivalent to (13·1·2) if we put[1]

$$M = \beta m_0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(13 \cdot 2).$$

In obtaining this result it has been assumed that the first principle applies to Newton's second law of motion as well as to the electromagnetic laws, if that law is stated in the form (13·1) that the rate of change of momentum is equal to the force. The result is confirmed by experiments on the deflection of rapidly moving

[1] In (13·1), $\frac{d}{dT}, \frac{d}{dt}$ are the 'full' coefficients, determined by the rate of change with T, t irrespective of change of phase. Since the u's are zero, the events happen in the same place for O_1, and $dT = \beta\, dt$. Further the β under the differential sign in $\frac{d}{dT}(\beta m_0 U_x)$, etc., must be differentiated; for if O_1 is always at rest relative to the particle v, his velocity relative to O_2 changes during the interval. Since $U_x = v$, $U_y = U_z = 0$, it is found that

$$\frac{d}{dT}(\beta U_x) = \left(1 + \frac{v^2}{c^2} \cdot \beta^2\right) \frac{dU_x}{dt} = \frac{du_x}{dt}, \text{ while } \frac{d}{dT}(\beta U_y) = \beta \frac{du_y}{dt}.$$

Of course there is something suspicious about this argument which first assumes that the relative velocity of O_1 and O_2 is uniform and then that it is variable. But if the assumptions are carefully stated in terms of limits, and it is considered what happens as $\frac{dv}{dt}$ tends to the limit zero, it will be found that they are self-consistent.

electrons in magnetic fields and (in conjunction with the Bohr-Sommerfeld theory) by observations on X-ray spectra (see Chap. XV, p. 84 etc.[1]).

Accordingly the principle receives notable support. It may be noted that if the law of motion has been stated in the more familiar form (which is that of d'Alembert rather than of Newton), that the force is equal to the product of the mass and the acceleration, a rather different result would have been obtained. It would have been found that the mass varied with the velocity, but that the mass was slightly different according as the body was moving already in the direction of the force (longitudinal mass) or at right angles to it (transverse mass). Experiment can scarcely decide between the two alternatives, but the greater simplicity of the first result makes it much more attractive. In dynamical problems treated in the light of the principles of relativity we must therefore be careful to introduce the laws of motion in the right form.

The simplest of these problems requires somewhat complicated analysis without introducing anything that is important for our present purpose, which is to make clear the meaning and foundation of the principles. Accordingly we may leave such problems as that of an electron moving in electric and magnetic fields or revolving round a charged nucleus to other chapters, where the results obtained will be simply stated without proof. But two more interesting conclusions, which can be more easily deduced, may be stated, although again they require nothing but straightforward calculation.

Let us calculate the kinetic energy W of the particle. The gain in energy must be equal to the work done by the forces $Q(E_x, E_y, E_z)$. Hence

$$\frac{dW}{dt} = Q(E_x U_x + E_y U_y + E_z U_z) \quad \text{.........(13·3)}.$$

[1] There is a very bad mistake on pp. 85, 92 of that chapter. It is stated or implied that the explanation of the motion of the perihelion of Mercury follows from the relation between the mass of a body and its velocity which has just been stated and which is deduced from the first two principles of relativity. Really that explanation requires another principle, which has not yet been introduced; in the usual terminology, it is a consequence of the 'general', not the 'special' theory of relativity. Equation (13·2) does predict *some* motion of the perihelion on Mercury, if we assume that gravitational forces behave in the same way as electrical, but its amount is much smaller than that observed.

We substitute for (E_x, E_y, E_z) from (13·1·2). These equations are symmetrical with regard to (X, Y, Z) and there is no reason to maintain the convention that U_x is equal and parallel to v while $U_y = U_z = 0$. We may write instead $U_x^2 + U_y^2 + U_z^2 = v^2$. There results

$$\frac{dW}{dt} = v^2 \frac{d}{dt}\left(\frac{m_0}{\left(1 - \frac{v^2}{c^2}\right)^{\frac{1}{2}}}\right) + \frac{m_0}{\left(1 - \frac{v^2}{c^2}\right)^{\frac{1}{2}}} v \frac{dv}{dt} \quad \text{......(13·4)}$$

$$= \frac{d}{dt}\left(\frac{m_0 c^2}{\left(1 - \frac{v^2}{c^2}\right)^{\frac{1}{2}}}\right) = \frac{d}{dt}(\beta m_0 c^2) \quad \text{...........(13·5).}$$

Choosing the constant of integration so that $W = 0$ when $v = 0$,

$$W = m_0 c^2 (\beta - 1) \quad \text{.................(13·6),}$$

which, when $\frac{v}{c}$ is very small, reduces to the familiar expression

$$W = \tfrac{1}{2} m v^2 \quad \text{......................(13·7).}$$

Again, since $M = \beta m_0$, we have

$$M - m_0 = \frac{W}{c^2} \quad \text{....................(13·8).}$$

That is to say, the 'extra' mass which a body possesses in virtue of its motion, over and above that which it possesses at rest, is proportional to its kinetic energy. In acquiring kinetic energy, the body also acquires mass; energy W appears to have a mass $\frac{W}{c^2}$. We have *proved* this result only when the energy is kinetic and derived from the motion of a charged body in an electric field; we may suspect that it is true whatever the form of the energy and however acquired. Some interesting speculations in atomic theory (see Chap. XVII) are founded on this suspicion; but at present it is unconfirmed. In any case the general proposition that all energy has mass is not a direct consequence of our principles.

III. THE GEOMETRICAL INTERPRETATION OF THE PRINCIPLES

14. The 'special' and the 'general' theory. We have now examined sufficiently the first and second principles of relativity, which form what Einstein calls 'the special theory'. In the next section, we shall introduce the other principles (which can be regarded as further extensions of the first and second) and so reach the 'general theory'. The general theory is necessarily much more difficult for a physicist to understand than is the special theory. For the special theory (as I hope the reader is convinced) can be stated intelligibly in terms of the experimental concepts which we employ in the laboratory; it can be stated in the form of assertions that, if certain experiments were tried, certain results would be found. The experiments cannot actually be tried, but nobody seems to find any difficulty in imagining them; such difficulty as they have is in imagining that the supposed results would follow. On the other hand the general theory cannot be stated in terms of experimental concepts—or, if it can be so stated, the statement would be so appallingly complex that nobody would understand what it really meant or how anyone could lay it down as a general and fundamental scientific proposition. The general theory can only be stated in terms of the purely mathematical concepts of generalised space. Now the special theory can also be stated in terms of generalised space; and it is probably only through the statement of this special theory, first in terms of one set of concepts and then in terms of the other, that the gulf can be bridged that separates these mathematical concepts from those which any physicist can understand. Accordingly in this section no new material will be introduced; we shall simply devote ourselves to stating in a different form what has been stated in the two preceding sections. If the reader can only grasp that the two statements are only two ways of saying the same thing, he may begin to hope to understand the general theory of relativity and all its consequences.

The course that we are adopting is historically correct. The form of statement which we shall reach at the end of the section (though not the manner of arriving at it) is due to Minkowski. It was only after and (as Einstein is always the first to acknowledge) because he showed how attractive and simple this form is, that the later developments, blossoming into the full general theory, began. The ideas that we shall reach in this section are fundamental in all modern work on the subject.

15. **Minkowski's treatment of relativity.** Let us start at the beginning again, and consider the two observers moving relatively to each other along a single straight line and measuring times and distances along that line; further once more we will start by supposing that the 'older view' is correct and that the relation between the measurements of O_1 and O_2 is given by (4·3). In Section I we supposed that their measurements were entered in note books which were compared after the experiment was over. But in place of (or besides) recording their observations by writing numerals in columns, the observers might record them by plotting. Each of them might take a sheet of squared paper, plot distances as abscissae and times as ordinates, and represent each event they observe as a point on the diagram. If they are observing the events of p. 9, their diagrams will be like Fig. 4. (Note that O_2's distances are all negative.) When the diagrams are compared, they will not be the same; even if the same units of distance and time have been taken by both observers and plotted on the same scale, the points, representing the events, will not fit if the diagrams are placed one over the other; they will not fit however the diagrams are turned round; they form different patterns.

Nevertheless, there will be a certain resemblance between them. In each diagram (remember we are adopting the older view) all the events which happen at the same time, but at different distances from the observer, will lie on a horizontal straight line parallel to the x axis, the position of this line depending on the time at which they happen. Again, all the events which happen at the same place in O_1's system (e.g. AC, AD or BC, BD) will lie on a vertical straight line in O_1's diagram; they will also lie on a straight line in O_2's diagram, but this line, instead of being vertical, will be inclined to the t axis at an angle $-\alpha = \tan^{-1}(-v)$, where $-v$ is the relative velocity of the two systems measured by O_2. Similarly all the events which happen at the same place in O_2's system (e.g. AC, BC or AD, BD) will lie on a straight line, which will be vertical in O_2's diagram, but in O_1's inclined to the t axis at the angle $+\alpha = \tan^{-1} v$. And if we introduce a point E moving relatively to both systems with uniform velocity, the events in which

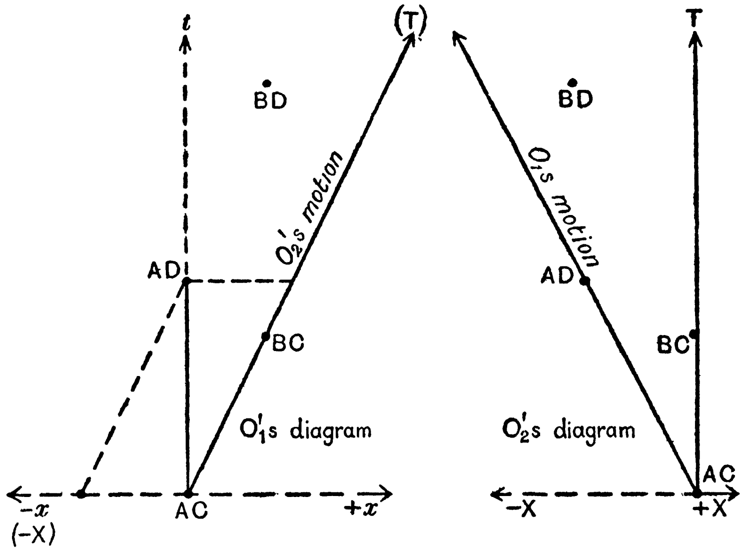

Fig. 4.

E is concerned will again lie on a straight line in both diagrams, but once more the lines will be differently inclined to the axes.

Now suppose for a moment that O_1 believes in absolute motion, and is convinced that he is at absolute rest, so that he considers his diagram right and O_2's wrong. With the resemblances just described in mind he may say to O_2: 'Your diagram is wrong, because you have plotted it with the wrong axis of t. My axis of t (being the line representing events which happen at $x = 0$ for all values of t) represents events which happen at my origin of distance; and all lines parallel to my axis of t represent events which happen at other fixed points in my system, points which I maintain are absolutely at rest. But your axis of time (T) and lines parallel to it, represent events which happen at the same place in your system, which is in motion. If you insist on using as your axis of T a line representing moving points, you must remember that the points are in motion and not adopt a procedure legitimate only if the points are at rest. My diagram shows that the line representing events which happen at your origin of distance (the line which you propose to use as axis of T) is not at right angles to the axis of x; if you use this line as your axis of T you are using oblique and not rectangular coordinates and must adapt your procedure accordingly'. O_2 may, of course, reply that O_1 has no special reason to suppose that he is at rest and that his diagram, rather than O_2's, is right. However, if they finally agree that neither of them is absolutely right, but that it is convenient in order to obtain agreement to assume that one of them is right, then they can reduce their diagrams to agreement. If the luck of the toss decides that O_1 is to be taken as at rest and that O_2 is to change his diagram to suit that of O_1, then O_2 will proceed to plot his observations on O_1's diagram—or, if he prefers, on a copy of it—in accordance with O_2's suggestion. He will take O_1's axis of x as his axis of X and use the same scale as O_1 in plotting distances. He will take the line AC, BC as his axis of T. Since his estimates of time agree with those of O_2, he must represent by the distance AC, BC measured along this line the same time as O_1 represents by the distance AC, BC measured vertically parallel to his axis of t; that is to say, his scale of plotting times must be greater than that of O_1 in the ratio $\sec\alpha$. The coordinates of AD will then be those shown by the dotted lines in Fig. 4 and the point for each event plotted from O_2's observations with O_2's axes will fall exactly on that plotted for the same event from O_1's observations with O_1's axes. The diagrams will have been brought into agreement.

It may be well to show that this re-arrangement of axes will produce the desired result. A change from rectangular to oblique axes implies that the coordinates of a point with the former (x, t) will be related to those of the same point with the latter (X, T) by linear equations of the form

$$X = ax + bt, \quad T = cx + dt \quad \text{............(15·1).}$$

The conditions of the transformation which we have imposed are (1) $T = t$, (2) $X = x$, when $t = 0$ and the origins of coordinates coincide, (3) $X = 0$, when $x = vt$, i.e. O_2's axis of T is O_1's line $x = vt$.

From these three conditions, it follows at once that $a = 1$, $b = -v$, $c = 0$, $d = 1$. Consequently (15·1) reduces to (4·3).

Now let us again abandon the older view and turn to that asserted by the second principle of relativity; we shall substitute (5·14) for (4·3) as giving the relation between O_1's and O_2's measurements. And here, for mere algebraic convenience, it will be well to alter (5·14) slightly. We are now going to take as unit of time the period in which light travels the unit distance. As a consequence $c = 1$ and, if we now make v mean the velocity expressed as a fraction of c, c can be struck out wherever it occurs in our equations. It will be seen that the relations between (x, X), (t, T) become quite symmetrical:

$$x = \beta (X + vT); \; t = \beta (T + vX) \quad \text{.........(15·2)},$$

where β now means $(1 - v^2)^{-\frac{1}{2}}$.

If now O_2 plots his observations on O_1's diagram using for X O_1's axis of x, but taking as axis of T the straight line by which O_2 represents events happening at the point $X = 0$ in O_2's system, his points (as might be expected) will not coincide exactly with those of O_1 for the same events, whatever scales O_2 uses for distance and time. By no choice of his axis of T or of the scales of his plotting can he make his points fit O_1's in all parts of the diagram. On the other hand—this is the important fact to which Minkowski drew attention—he can make them fit if he alters his axis of X. The rule which he must adopt in choosing his new axis and scales is best stated in the geometrical form given by Minkowski.

Let O_1 draw in his diagram the rectangular hyperbolas (Fig. 5) of which the equations are $t^2 - x^2 = 1$ and $t^2 - x^2 = -1$. Let OT, the line representing to O_1 events at O_2's origin of coordinates, cut the curve at P, so that $TOt = \tan^{-1} v = \alpha$. The line OX, cutting the other hyperbola in Q and making an angle α with Ox, is the line which O_2 is to take as his axis of X. In place of Op and Oq by which O_1 represents his unit time and unit distance, O_2 is to take to represent his unit time and distance the lengths OP, OQ. It is quite a simple matter to prove that, if O_2 chooses his axes and his scales of plotting in this way, the values of X, T for any point in the diagram will be related to those of (x, t) by (15·1) and, consequently in virtue of the second principle, that the same point will represent the same event to both O_1 and O_2. For the conditions which we must now impose on (15·1) are (1) at Q $X = 1$, $T = 0$, (2) at P $X = 0$, $T = 1$. But the coordinates of Q are

$$x = (1 - v^2)^{-\frac{1}{2}}, \quad t = v(1 - v^2)^{-\frac{1}{2}},$$

and of P

$$x = v(1 - v^2)^{-\frac{1}{2}}, \; t = (1 - v^2)^{-\frac{1}{2}}.$$

It follows at once that (15·1) reduces to (15·2).

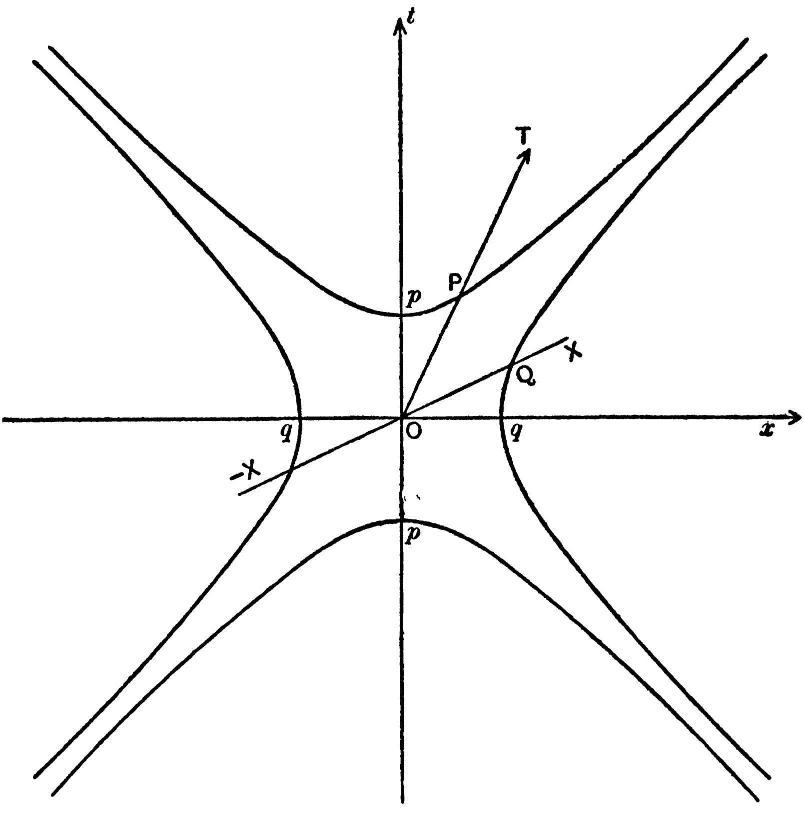

Fig. 5.

We may then express the difference between (4·3) and (5·14), or between the older view and the principles of relativity, by saying that, in the circumstances supposed, the former permits us to change our axis of t without any accompanying change in the axis of x, while the latter demands that if we change one axis we should also change the other in a suitable manner. Minkowski now argues that, when the difference is expressed in this manner, it appears at once that the principle of relativity is intrinsically more reasonable than that which it replaces, and becomes still more reasonable when this mode of expression is developed somewhat further. Accordingly, if we had no experiments concerned with velocities large enough to make the difference between the two principles appreciable, it would still be preferable to accept that of relativity; the older view is essentially paradoxical. It is in this step of the argument that the real difficulty seems to me to occur. In the next paragraph the argument will be stated; thereafter we shall return and consider it in some detail.

16. **'Space' and 'Time'.** The conception of changing the axes to which we refer measurements is not new; it is quite familiar in dynamics, where analysis is often made easier by referring the motion first to axes fixed in a moving body and then translating our conclusions into terms of some other axes to which measurements are actually referred. This method of solving certain problems is rendered possible and its limitations are assigned by a certain feature of the Newtonian laws of motion, namely that they retain their form (in the sense discussed on p. 37) (1) if the origin of coordinates is changed and the axes translated parallel to themselves, or (2) if the axes are rotated in any way round the origin, or (3) if a uniform translation is imposed on the axes. In order that they may retain their form for changes (1) and (2), it is necessary that the axes be shifted or rotated as a rigid body; if the position of one axis were changed, while the others were left unaltered, the axes would become oblique (if they were originally rectangular) and the form of the law would be changed. Accordingly the condition for no change in form of the laws for changes (1) or (2) of the axes is, broadly, that if we take a new axis of x (say), we must also take new axes of y and z. Now we have just seen that changes (3) can also be regarded as a change of axis, a change of the t axis. The older view, as we have seen, held that this change of axis need not be accompanied by any change of the other axes; the principle of relativity on the other hand says that this change

of axis is on a par with the others and, like them, must be accompanied by a suitable change of the related axes. Regarded in this way, surely the new view is the simpler and the more reasonable if all the changes (1), (2), (3) are nothing but changes of axes, surely it is more reasonable to suppose that the same rule applies to all. The time axis in Fig. 5 is perfectly symmetrical with the space axis; surely it is preferable to regard space and time as substantially the same thing, both subject to the same rules, than to regard them as two things to which it happens that the same rule applies.

If we once accept this idea, a further development, leading to even greater simplicity and reasonableness, follows. We started on all this argument from the question whether O_1's and O_2's representations of events by points could be made to fit. Now the fitting of one set of points with another is not really a question of axes at all; axes are convenient, or perhaps necessary, in setting out the points from the measurements; but once the points are set out, the fit is determined wholly by the distances between the points. If two groups of points are such that the distance between two points in one group is always equal to that between the corresponding two points in another, then, and then only, the two groups will fit. Now distances between points can be determined from their coordinates referred to any axes; but they can also be determined without any use of axes or any knowledge of the axes from which they were originally set out; they can be determined with a graduated measuring rod. Accordingly our conclusion that the observations of the two observers can be made to fit, even though they are made with different axes, inevitably suggests that what is really being measured is something of the nature of distance, and that the agreement between the two sets of measurements really means that they are measurements of the same distance.

Now the distance l between two points in space, of which the rectangular coordinates are (x_1, y_1, z_1), (x_2, y_2, z_2), is given by

$$l^2 = \delta x^2 + \delta y^2 + \delta z^2 \quad \text{.................}(16{\cdot}1),$$

where $\delta x = x_1 - x_2$, etc.

Two observers determining the positions of such points will always arrive at the same value for l, even if they use different axes and find different values for δx, δy, δz. On the other hand, the second principle of relativity as expressed by (7·2) (we now remove again the restriction that the motion studied is in one dimension only) asserts that both observers find the same value for s, where

$$s^2 = \delta t^2 - \delta x^2 - \delta y^2 - \delta z^2 \quad \text{..............(16·2)},$$

the c dropping out with our new convention about units. The similarity between (16·1) and (16·2) is striking; (16·2) merely contains one more term, of the same form as the rest, on the right-hand side and shows a change of sign. s may be regarded as a distance, not in our ordinary three-dimensional space, but in a four-dimensional 'space-time,' in which the fourth dimension is that which is ordinarily called the time, but is essentially similar, according to our present view, to the remaining three 'space' dimensions. It is true that the space terms in δx, δy, δz are negative, and that the term in δt differs in sign from the remainder, but these differences are trivial. Time, we may say, differs from space in being represented appropriately by an imaginary rather than a real quantity; for, if we suppose time to be essentially imaginary and write $t = iu$, we can write

$$-s^2 = \delta u^2 + \delta x^2 + \delta y^2 + \delta z^2 \quad \text{..............(16·3)}.$$

The only difference now is that the distance s is also imaginary, while l is real.

We propose then to restate the second principles of relativity (and the first in so far as it is involved) in the following form. Events are points in a four-dimensional space-time, of which three dimensions are ordinary space, the fourth mathematically imaginary time. The measurements which we have supposed the two observers to make are simply determinations of the positions of these event-points in space-time. The measurements differ because the two observers make them from different axes; each chooses as his axis of t the path in space-time of the point which he takes as his origin of spatial measurements. The positions of any point relative to his axes is different for the two observers, but the distance between any two event-points is the same for all observers; it is something absolute and independent of observation or at least of the observer who makes it. This statement is equivalent to (16·3) which is again equivalent to (7·2), which is merely the algebraical expression of the second principle.

17. Mathematics and physics. Such is the argument in outline. Doubtless others would want to alter it a little so as to make it appear to them still more convincing, but nobody is likely to maintain that it has been seriously distorted. If the reader finds it quite easy and intelligible he may pass on directly to Section IV. But if he is one of those to whom this book is addressed, he will not find it so; he will find it getting steadily less tractable as it proceeds and finally dashing over a precipice of imaginary quantities into a whirlpool of incomprehensibility. For such a reader I am going to try, not to make the final conclusion really comprehensible —for I believe that to be impossible—, but to show him more exactly why it is incomprehensible.

The reason lies in a fundamental difference of attitude between the mathematician, who puts forward the argument, and the physicist, who fails to understand it. The figures which we enter in our note books as the result of any experiment represent relations of two perfectly different kinds. First, they represent numerical relations; the numbers that they denote are connected in the way in which numbers, and numbers only, can be related. Thus the numbers in one column may be proportional to those in the other, or proportional to their square, or to their cosine, or so on. Secondly, they represent physical relations, the kind of relations with which experimental science is concerned. Thus the numbers in one column may represent the pressure of the gas of which those in the other represent the temperature, or the mass of some portion of a substance and its volume, or the distance of a material point from some origin and the time which it has taken to reach that point. The difference of attitude between the mathematician and the physicist is that the former is primarily interested in relations of the first kind, while the latter is primarily interested in relations of the second kind. Relations between numbers, and the much more general relations between 'variables' of all kinds which have been developed from them, provide the subject matter of a great part of mathematics, including all of it which has any immediate connection with physics; mathematical physics simply is the study of these relations and of the ideas directly arising from them. On the other hand the art of experiment deals with the second kind of relation; the difficulties of experiment and the precautions that have to be taken in order that experiments may be of value are all concerned with these. The troubles which we should have with the 'dead space' in securing right figures to represent the relation between pressure and temperature, or with the purity of materials and variation of temperature in securing the right relation between mass and volume are nothing but troubles in making the physical relations what we intend them to be. For the things between which we assert the numerical relation of proportionality are not (e.g.) any mass and volume, but those between which the physical relation holds which is expressed by saying that they are the mass and volume of the same pure substance at the same temperature.

The difficulty that we are studying arises from the fact that the same numerical relation may be associated with very different physical relations. In the three examples given, the numerical relation is the same in each case, and the three experiments are therefore practically indistinguishable to the mathematician; for his purpose it does not matter whether the numbers which he is given to study were obtained by one kind of experiment or by another; if they are the same numbers, or numbers with the same numerical relation, he has no need to inquire further. But to the physicist the difference between the three experiments is far more important than the resemblance; it makes all the difference in the world to him whether he has got to struggle with leaks in glass apparatus and mercury spilt on the floor, or with a balance that shifts his zero and draughts which set up temperature inequalities.

If mathematicians were always pure mathematicians and experimenters pure experimenters, no trouble between them would arise; they would each go their way, talking, it is true, about different things in a different language, but with no possibility of conflict. Conflict arises only when the two attitudes are confused. In this particular instance, the first lapse from single-minded virtue must be set against the experimenters. We ought never to have 'plotted' our results. For 'plotting' is simply the substitution of one pair of relations for another pair, the numerical relations being the same in both pairs, but the physical relation different. The numerical relation between (e.g.) the pressure and temperature of our gas is the same as that between the ordinates and abscissae of the point in the diagram by which we represent that pressure and temperature in our diagram; but the physical relation is utterly different; in one case it is a relation between two properties of a perfect gas, in the other between two perfectly different properties of a sheet of paper; chalk and cheese are not more different. This neglect of the physical relation and insistence on the numerical relation serves of course very useful purposes; it so happens that we can recognise and appreciate certain numerical relations directly, and literally at a glance, when they are associated with 'spatial' relations, and cannot appreciate them so readily when they are associated with other physical relations or when, in our note book, they are associated with no physical relation at all. In practice it will never mislead us; for we are not likely to forget the physical relations. But it may mislead others, or at least lead them into paths where we find it difficult to follow. It appears to us paradoxical when the mathematician announces that space and time are merely different aspects of the same thing; for we know that measuring times is perfectly different from measuring spaces. But his assertion is an immediate and almost inevitable outcome of our practice of representing every kind of magnitude, pressure or volume, mass or temperature, as a kind of space for the special purposes served by plotting.

But the mathematician adds his quota to the confusion. If he would use terms and ideas that were clearly analytical and appropriate only to numerical relations, we might find his argument hard, but we should not be so puzzled. He would then state the conclusion at which he arrives in some such terms as these: The variables (x, y, z, u) form a continuous four-dimensional series such that the expression $ds^2 = dx_1^2 + dx_2^2 + dx_3^2 + dx_4^2$ is invariant for linear transformations of the variables of the type

$$x_m = \sum_n a_{mn} X_n \quad \text{.....................(17·1)}$$

subject to the conditions

$$\sum_n a_{mn}^2 = 1, \quad \sum_m a_{mn} \cdot a_{mo} = 0 \quad \text{...........(17·2),}$$

and subject to some other conditions if one of the variables is imaginary. Such a statement would appear to us wholly unobjectionable—and also wholly uninteresting. Our interest and our objections are aroused because he uses terms which are intimately associated with physical relations, but not with the particular physical relations that are concerned in the matter under discussion. Thus he speaks of 'change of axes' and 'distances'; those terms are associated in our minds with magnitudes which are related by the special type of physical relations which we call spatial; that is to say, with magnitudes that are lengths. If one or both of the magnitudes are not lengths, these terms are not appropriate. If we are considering pressures and volumes, even if we are plotting them, we do not attribute any physical meaning to these terms; we should not expect our diagram of the isotherms of a gas to remain the same if we changed the axes from which it was plotted; and we should not expect that, in comparing two states of a gas, or the same gas observed in two different conditions, the expression $\delta p^2 + \delta v^2$ will represent anything that is experimentally important. The mathematician confuses us because he uses the terms 'change of axes' and 'distance', when one of the magnitudes concerned is not a length but a time interval. And he not only uses the term, but he uses the idea; he uses the physical relation which would hold if the magnitudes were lengths, but have no application if they are not lengths, to suggest purely numerical relations. The conception of the fitting of the observations is expressible in purely analytical terms; but nobody would have thought of introducing that analytical criterion and calling it 'agreement' or 'observing the same events', if it were not for the analogy with spatial fitting—an analogy which from our point of view is wholly false. The reader whom I am addressing doubtless felt that the procedure discussed on p. 52 was wholly unnatural and artificial. Why should we expect the two sets of observations to 'fit' when plotted on a diagram? It is not a criterion of agreement which a physicist would think of employing; it is plausible only if we allow ourselves to confuse spatial and non-spatial magnitudes.

The misunderstanding arising from the misuse of the word 'distance' is so important that it will be well to state rather more formally one essential difference between the true spatial distance concerned in (16·1), and the mathematician's generalised 'distance' concerned in (16·2). Spatial distance is a magnitude directly measurable as such and without regard to any other magnitude. (16·1) is a true numerical law, stating a relation between one fundamental magnitude l and other fundamental magnitudes δx, δy, δz. (It happens that all the magnitudes are of the same kind and directly comparable with each other; but this feature is not immediately relevant.) On the other hand, s is not a fundamental magnitude, directly measurable as such; it is a mere name for a particular function of the magnitudes δx, δy, δz, δt which are directly measurable; moreover, if it were a measurable magnitude at all, and not a mere name for the function, it would be a derived magnitude, a magnitude that can be measured only by measuring something else. And (16·2) is not a numerical law; it does not state anything that is the result of experiment; what experiment states (or is imagined to state) is that the function of δx, δy, δz, δt which we propose to call s is the same for all observers; it does not tell us that this function is s or is equal to s—that is a mere definition; for once more, we cannot measure s except by measuring x, y, z, t.

We pass to another difference between mathematician and physicist closely connected with the first. To us no distinction is more important than that between what does actually happen and what might happen or can be imagined; we spend our whole working lives in making that distinction. But to the mathematician it is utterly trivial. He wants numbers to argue about; it does not matter to him whether the numbers arose from some observations which were actually made or whether somebody wrote them down from pure imagination. It is this difference between us that makes our attitude toward 'four-dimensional space' and imaginary quantities so different. The mathematician says: I see that you sometimes fix the varying position of a point by a single series of numbers entered in a single column, sometimes by two columns, sometimes by three. Why this limitation; why not four? I am going to imagine that there are four columns fixing the position. It is vain to protest to him that we can actually find conditions in which one, two, or three columns will fix what we call position, and that we cannot actually find conditions in which four are required; he brushes the objection aside; it does not interest him. We know to our cost that, while it is easy enough to plot the relation between two magnitudes, it is difficult, tedious, and not very illuminating to plot a relation between three, and wholly impossible to plot one between four. All that does not matter to the mathematician; he passes gaily from the possible two-dimensional plot to the impossible four-dimensional plot without giving the matter a thought; he can argue about one as well as about the other. And yet he transfers to his imaginary four-dimensional plot, terms and ideas, such as distance, which originally arose only from the study of actual two- and three-dimensional plots. In the same way with imaginary quantities. We enter in our note books numbers of which the squares are positive. The mathematician can argue equally well about numbers of which the squares are negative. I will imagine, he says, that you have entered such numbers; and wanders off in his airy way, leaving us to gasp and protest that the very meaning of measurement, as we understand it, implies that we never could write down such numbers.

18. **The value of the geometric interpretation.** Such are the differences between us and the mathematician which make his argument so hard to understand. Are we then to conclude that the argument stated on pp. 56—58 is mere nonsense, and that we can merely disregard it in the same spirit as the framers of it disregard the realities of physics? If words have the meaning we attribute to them, then, judged by any possible standard of formal accuracy, the argument is nonsense, nonsense as complete as Lear's rhymes; it is based on an inextricable confusion between numerical and physical relations. And if words such as 'distance' are to be used in some sense wholly different from that to which we are accustomed, then the argument ceases to be an argument at all—even a bad one. However, formal accuracy is a fetish to which much more valuable things may be, and have been, sacrificed; when we consider 'rejection', we must consider also the purpose to which the train of thought is to be put. An anthropologist anxious to study the geographical distribution of English eccentricities would be wise to reject Lear, but not a harassed uncle with young children to amuse. To what purpose is the Minkowski theory to be put?

In this section we have regarded the geometrical interpretation as merely an alternative method of expressing the experimental facts about the relations between the observations of two observers (or what we have agreed to regard temporarily as experimental facts) that were discussed in the last section. And if it is such an alternative method of expression it cannot be definitely wrong or false, unless the facts are wrong. On the other hand, if they are right (and this is the assumption on which the whole of this section is founded), it has great advantages; it enables the necessary mathematical deduction from the facts to be made with ease and elegance. This is the aspect of the principles of relativity which is the main theme of most other serious expositions, very little attention is paid to it here, because purely mathematical developments are expressly excluded from the scope of the book. But it is of such immense importance that some slight reference to it must be made.

The fundamental idea of these developments is that the history of any particle is to be represented by a 'world-line' drawn in space-time, which defines the relation between the position of the particle at any moment and the time of that moment. The history of any system is represented by the world-lines of all its particles. Thus, in Fig. 4, the history of A is represented by the line (straight in this instance) through AC, AD, that of B through BC, BD and so on. An event happening to a particle is the intersection of its world-line with some other world-line; thus the event BC is the intersection of the world-lines of B and C. Strictly, these conceptions are applicable only when the whole history of a particle is described by its motion and when there are no events other than the coincidence of particles. But if there are other events, e.g. the passage of the particle through a particular electric field, we can represent such an event by imagining some second particle of which the world-line crosses that of the first particle, while its presence at a point of space-time implies the presence of the particular electric field at that position and time.

If, then, we know the world-lines of all particles, both real material particles and those which are invented to carry with them magnitudes such as electric or magnetic fields, we should know the entire history of the material universe. And, if we know some of the world-lines, we may be able to deduce the course of others, in somewhat the same way as we can deduce information about the third side of a triangle from knowledge about the other two. A physical calculation, involving any laws which can be brought within the scope of the principles of relativity, is reduced to a calculation in four-dimensional geometry, one dimension being imaginary. Now four-dimensional and other 'non-Euclidean' geometries have been a favourite study of pure mathematicians; a great deal is known about them and very complete methods of treating them have been developed. One consequence of the Minkowski interpretation is to make this large and healthy branch of pure mathematics available for the purposes of mathematical physics. Its methods are not easy to those who have not the mathematical instinct; they are harder to them than straightforward algebraic calculation from (7·1) or (7·2), just as 'vector' methods (which they resemble) are harder than Cartesian. But as they are harder, so also they are more elegant and even more powerful in the right hands. Most of us, who are not mathematicians, seeing how results which we can only derive by pages of symbols follow in a few lines, wish that we too could handle vectors with familiarity. Just so, once the general properties of world-lines are worked out, all the problems of the electromagnetics of moving systems, or those of the propagation of light from which we started in Section II, follow with ridiculous ease.

But the statement just made that the methods are those of a certain kind of geometry must not be misinterpreted. They are purely analytical; they deal throughout with symbols and not, like the methods of Euclid or of Newton's *Principia*, with pictures. Their utility is no evidence against the view urged here that the geometrical analogy, while good in some respects, is in others quite misleading. They bear to physical geometry, the experimental study of the laws of the geometrical magnitudes, length, area, angle and so on, much the same relation as the theory of theta-functions (which were developed from Fourier series, which again were introduced to treat problems in thermal conduction) bears to the physical science of heat.

A second use of the Minkowski interpretation is to explain the first and second principles, to tell us *why* they are true. For this purpose the geometrical analogy is developed in great detail and resemblances traced between the laws implied by the first and second principles and statements of the properties of four-dimensional space. This aspect of it interests primarily philosophers and that inveterate but exceptionally muddle-headed philosopher, the newspaper reader; it is of no service to us. To explain is to interpret in terms of more acceptable ideas and the ideas of Minkowski are essentially unacceptable. It does not help us to understand why all observers at rest relative to a system make the same observations of it or why observers in relative motion may differ as to the time sequence of events, if we are told that all our observations and all our measurements are nothing but the placing of points in a four-dimensional and partly imaginary time-space. For, in so far as we can understand such a statement, we know it to be false; time and space are not the same thing, neither is imaginary, and there are observations which are concerned with things that are neither time nor space. No facts are incomprehensible, though they may be surprising; an 'explanation' which is incomprehensible is no explanation at all. The more the analogy is developed, the more prominent its defects are likely to be; and for this reason it has been treated here as lightly as possible.

But we cannot maintain this position much longer. For lastly, Minkowski's interpretation of the first and second principles suggests, especially or perhaps only to mathematicians, extensions of them. This is, by universal admission, its most important function. In the next section we shall endeavour to grasp these suggestions and their consequences. The task will be difficult because it is unlikely that the suggestions will appeal to us. But any success that may be achieved will come only if, for the time being, we try to rid ourselves of an antagonism to Minkowski's ideas and to make them part of our normal mental equipment. In this attitude of mind, let us address ourselves to the task.

IV. THE GENERAL THEORY

19. **Accelerated motion.** So far we have supposed that all the motions we are studying are made with uniform velocity. We now remove this condition and repeat the questions from which we started. We ask how the observations of any observer will be affected by setting him in motion, not necessarily uniform, relative to the system that he observes. We propose to answer the question in the same kind of way, namely by inquiring first how, if at all, the observations of an observer O_2 at rest relative to the system will be changed by its motion relative to the first observer O_1, and, second, what will be the relation between the observations of the two observers; and we shall try again to express our conclusions mathematically in some form similar to that in which our previous answer was expressed.

The answer that we give to the first part of the question must be unchanged, for that answer is inherent in the fundamental ideas of the whole doctrine of relativity. If O_2 is at rest relative to the *whole* system under consideration, then his observations must be the same as those of any other observer similarly at rest and, in particular, the same as those of O_1 when it was at rest relative to him. But if, when the system is set in motion, some part of it remains at rest relative to O_1, then we cannot expect that the observations of O_2 will be precisely similar to the original observations of O_1. What we do expect is that the form of the laws satisfied by the observations of O_2 and of O_1 shall be the same, to such an extent that there can be no reason for O_1 rather than for O_2 to assert that he is at absolute rest. O_1 and O_2 may differ as to the value of any magnitudes which they measure, but they must agree as to what magnitudes are concerned and what are the numerical laws between these magnitudes.

The laws that are of particular importance when we are studying non-uniform, or accelerated (which of course includes retarded), motion are the laws of dynamics, and the associated magnitudes are forces; for, according to Newtonian dynamics, accelerated motion can exist only when the moving body is under the action of forces. And part of the answer that has just been given is implied by Newtonian dynamics—at least as it is interpreted nowadays. If two observers at relative rest observe the motions of the same system, they will agree perfectly in the laws that they find and in the numerical values they attribute to forces acting in the system; and this agreement will persist if the observers are in *uniform* relative motion. But if they are in accelerated relative motion, then, though they will both find that the system obeys the same laws and that its parts move under forces, they will give different values to the forces acting on its parts. The familiar example always quoted in this connection is that of two observers; one on the earth and sharing in its rotation; another close at hand, but separated from the earth, at rest relative to the 'fixed' stars and not sharing in the rotation. Both will find that a free body near the earth moves towards it with uniformly accelerated motion and is therefore under the action of a constant force, but they will attribute different values to this force. The observer at rest relative to the stars will find the value g, independent of the position on the earth at which the experiment is made (assuming the earth to be a perfect sphere); the observer on the earth will find a value $g - \omega^2 r \cos\lambda$ where $\omega^2 r \cos\lambda$ is the term arising from the centrifugal force due to rotation and dependent on the latitude λ. A third observer, falling freely towards the earth, would find that other unsupported bodies moved with uniform velocity and would therefore assert that there was no force at all.

20. **The principle of equivalence.** That is a clear deduction from Newtonian dynamics, but an obvious implication of this deduction, when expressed in that form, would not have been accepted by Newton. He would have admitted that the three observers would attribute three different values to the force near the earth's surface, but he would not have admitted that all three were equally right. He would probably have maintained that the only observer who was right was the one at rest relative to the stars; he alone determined the true force. If that statement is to have any meaning, there must clearly be some way of determining what is the 'true force' apart from the observations which have

been mentioned; though the three different forces are equally satisfactory for describing these experiments, there must be some other experiments for the description of which they are not equally satisfactory; there must be some which can be described satisfactorily only in terms of the true force. Whether there actually are such experiments is a matter we shall have to discuss; but even if he could not have invented any at the time, Newton would have maintained that such experiments were possible and might be found at any moment. It is here that the principles of relativity diverge from those of Newton. The general theory of relativity, which we are about to discuss, starts from the statement that no such experiments are possible, that by no possible means can two observers, relatively accelerated, determine which of them is measuring the true force and which measuring a fictitious force including a term arising out of his own acceleration; or, as Einstein expressed it, calling the doctrine the Principle of Equivalence, that a gravitational field of force is exactly equivalent to an acceleration. This is the extension of the first principle that is proposed to cover non-uniform motion.

It is apparent at once that this principle is plausible only if the fields of force to which it is applied are gravitational. For in other fields of force, the acceleration depends upon the nature of the body on which the force acts. Acceleration of the observer would increase by the same amount the acceleration of all bodies and increase the forces on them by the product of this acceleration by the mass. But some forces (e.g. electrical forces) do not depend at all on the mass according to our present view; the introduction of electrical forces depending on the mass would make havoc of electrostatics. On the other hand, since gravitational forces are proportional to the mass, the forces added by the acceleration of the observer are of the same nature as, and indeed proportional to, those due to 'pure gravitation'. However there seems no objection to confining a principle of this nature to gravitation, for gravitation is exceptional. Indeed its peculiar features, its universality and the strict proportionality of mass to weight, which fall in so well with the principle of equivalence, are precisely those which have proved the stumbling block to those who have tried to make experiments on it or devise theories to explain it. Anything which will assimilate gravitation to 'centrifugal force', which shows the same features, rather than to electrical or other 'physical' forces of more normal type, is clearly to be welcomed. Here there is no difficulty, but only an attractive simplicity.

But there are difficulties, and it will be well to consider them before we proceed. The first needs only bare mention. It is that while we can readily enough conceive an observer moving with uniform velocity relative to the system which he observes, and yet wholly independent of that system in the sense of p. 5, it is not so easy to conceive of an independent observer who is accelerated relatively to the system he observes. For the mere fact of his acceleration proves that, in ordinary parlance, there is a force between him and the system; and of course the existence of a force between two things directly implies that they are not independent. And yet if the observers are not independent of the systems observed, the new principle is utterly different from those we have considered before; these principles assumed necessarily the independence of the observer and the system observed, as the reader may find for himself if he attempts to repeat the arguments of previous sections denying such independence.

21. **The measurement of force.** But the second difficulty needs closer examination, if only because it is so often passed over too lightly. It arises from an ambiguity in the term 'force', to which we must devote some little attention.

'Force' may mean, and in the history of science has meant, three things, which may be termed muscular force, static force and dynamic force. Muscular force is something appreciated by direct sensation. When we set a heavy body in motion by the action of our limbs, we experience certain sensations which everybody knows and nobody can describe. Such muscular force is not a strictly scientific conception; we cannot compare directly the muscular force which we exert with that which others exert, and we cannot measure it directly; but it provides the foundation for static force, which is a scientific conception and a measurable magnitude. Let us sketch very briefly the kind of experiments which enable us to pass from muscular to static force and to measure the latter.

We find that bodies can be set in motion, not only by the action of our limbs, but also by placing them in certain relations to other bodies. Thus if we attach a body to the end of a stretched spring and release the catch, the spring will contract and move the body. The analogy with muscular force leads us to say that the spring exerts force on the body. We try now to find laws which will enable us to represent this force exerted by the spring as a measurable property of the spring in any particular state of stretch. Of the laws which enable us to attain this end, the following are the most important. (1) We can find two springs A and B, such that when each is in a defined state of stretch, they 'balance'. That is to say, if we attach the two springs, both at the same time, to the same body (any body will do) we find that when their directions are properly related (i.e. when they are pulling in 'opposite' directions) and when they are both stretched to the defined state, they will remain in that state and produce no motion in the body. If we now find a third spring C which in its defined state balances A in this sense; then we find, as an experimental fact, that it will also balance B. This law, which is sometimes stated as the law of the equality of action and reaction, enables us to define equal forces, forces being—it must be remembered—properties of the springs or other similar bodies on which the experiment is made. (2) We take a pair of springs A and C the forces of which are equal in this sense, and another pair B and D the forces of which are also equal; we attach all four to the same body, so that A and B pull in one direction, C and D in the opposite direction. We find that the forces when so combined still balance. This law, sometimes stated as the law of the independence of forces, enables us to define the addition of forces.

We can now measure the static force characteristic of any body or system in terms of any arbitrary unit. We find, for example, that the earth exerts a static force which will stretch a spring when a body is hung on the end of it. There is a force characteristic of the earth and of the distance from it just as there is a force characteristic of a spring. The force is somewhat different from the forces due to springs because it depends on the properties of the body on which it is exerted, but that is immaterial for our present purpose; we can measure without any ambiguity the force exerted by the earth on a particular body. But there is one point we must carefully note. All measurement of static force is made by means of the balancing of forces, such balance implying that the parts of the system consisting of the body exerting the force and that on which it exerts force are at relative rest. If they are in relative motion, there is, by the very meaning of the term, no possibility of measuring static force. In a certain very definite sense, there is no such thing as the static force between two bodies in relative motion or, more accurately, between two bodies of which the distance is changing. (One body revolving in a circle round another is usually said to move relative to it, but there is no change of distance and no motion in the particular sense employed here.)

We now pass to dynamic force. This conception arises from the inquiry, which is Galileo's great contribution to science, how a body moves when the balance of the static forces on it is upset. We conceive a body balanced between two equal measurable static forces; the system producing one of them is removed; the body starts into motion in a manner which proves to be determined by the system producing the other force. We ask exactly how this motion is determined by the force due to that system, meaning by the force that measured in the same condition of the system when the body was at rest and the force balanced. The experiment can be made easily only with gravitational force, because it is the only static force which remains constant while the body moves over an appreciable distance; but there are very good reasons which will at once occur to the reader, why the result which he obtained is to be extended to all static forces. Galileo's result was that a body moving under constant static force has a constant acceleration in the direction of the force and that the acceleration is proportional to the force.

But though static force always thus produces acceleration, the converse is not true, and there are notable examples in which a body is accelerated when there is no static force. Thus a copper disc set spinning between the poles of a magnet is accelerated (i.e. retarded), but when it is at rest relative to the magnet no static force can be discovered between them. We say nowadays that the force which causes the acceleration of the disc is caused by the motion and does not exist in its absence; but if in that statement we meant static force, we should be talking nonsense, for static force is something that can be investigated only when the system is at rest. We have changed the meaning of force by recognising dynamic force, a conception which has a meaning only in virtue of Galileo's discovery as generalised by Newton. The dynamic force on a body A due to a system B simply means the acceleration of A relative to B multiplied by a factor characteristic of A, called its mass, with the exact means of determining which we need not concern ourselves here. Static force and dynamic force are not the same thing; on the contrary they are as different as Box and Cox. For if, when we say there *is* a force, we mean that we can measure it, then if there is a static force on A due to B, there is no dynamic force, and vice versa. But there is a close connection; for the dynamic force which a spring exerts on a body when the catch is released is proportional to—equal to, if we choose units suitably—the static force which it exerted just before the catch was released.

With this explanation let us return to the earth's gravitational and centrifugal forces on a body at the equator. The only observer who is in a position to determine without any ambiguity the static force exerted by the earth is the one on the earth, and he can determine it only when the body is at rest relative to the earth. In that case he will find a static force $g - \omega^2 r \cos \lambda$.

The observer at rest relative to the stars might measure statically the force exerted by the earth on a body at rest relative to him, if the earth's motion were one of rotation only; for then, since there would be no change of distance between that body and the earth he might hold that he had secured a balance. But since the earth revolves round the sun and is therefore changing its distance relative to him, he cannot secure a balance—unless he assumes that uniform motion does not invalidate a balance; but in making that assumption he would be practically substituting dynamic for static force. The observer falling freely cannot measure the static force exerted by the earth at all. He can show that the force exerted on a body at rest relative to him by a spring at rest relative to him is zero, but that tells him nothing about the static force exerted on the body by the earth. He cannot secure at once a balance between forces exerted by the earth and forces exerted by the measuring springs at rest relative to him.

Accordingly, if by the gravitational field of force of the earth we mean the field of static force, the observations of the three observers are by no means equal in value. The only one whose observations are indubitably valuable is the one who is on the earth, and the only indubitably true force is the one which, in the ordinary mode of expression, is said to be compounded of the true gravitational force and the centrifugal force[1]. In respect of statical force, acceleration is *not* equivalent to a gravitational field; and if it is applied to the effects of statical force the principle of equivalence presents great difficulties. Consider, for example, the barometric pressure at the earth's surface. The principle of equivalence in its natural interpretation would seem to predict that an observer falling freely towards the earth would conclude that the barometric pressure was zero; for it would be zero if there were no gravitational field. But that conclusion would be almost certainly false. Accordingly, in drawing conclusions from any results we may obtain from the principle of equivalence, we must always carefully inquire whether we have applied it to dynamical or statical forces.

22. **Extension of the second principle.** With this warning let us accept the principle of equivalence as an extension of the first principle and see whither it leads us.

In Section II, p. 37, we introduced the assumption, which we regarded as a natural development of the first principle, that the form of certain laws was the same for two observers moving relatively to each other with uniform velocity. This assumption, combined with the second principle, which described how the observations made by such observers are related, enabled us to deduce a relation between the values of certain magnitudes, involved in those laws, as determined by the two observers. We found how their estimates of magnetic and electric forces and of the masses of moving electrified particles must be related. We have now, in accepting the principle of equivalence, made a similar assumption

[1] The reason why we regard g as the true gravitational force is that if we separate this component from the measured statical force $g - \omega^2 r \cos\lambda$ we get a much simpler relation between the force and the properties of the earth than we do if we take the measured force unseparated. That is to say, we call it the true gravitational force because it satisfies the Newtonian theory of gravitation. But this theoretical separation of the measured force has nothing to do with the questions under discussion in the text.

for observers moving relative to each other with uniform relative acceleration; we assume that both of such observers will find that the law of gravitation has the same form. If we had a principle, corresponding to the second, relating the space and time measurements of the observers, we might follow a similar argument and hope to arrive at a conclusion concerning the values which they will attribute to gravitational magnitudes.

But we have as yet no such principle; and even if we had, the problem which we might expect to solve by means of it is not exactly that which we propose to solve. For we are *not* going to assume that we know exactly the form of the law of gravitation, as we assumed that we knew the form of the electromagnetic laws: We are going to try to find, not merely the numerical values attributed to magnitudes involved in a law of known form, but the form of the law. These two differences, namely that we have nothing corresponding to the second principle and that we do not even know the form of the law, will make our argument follow a somewhat different course from that which we considered before; but the line of thought may be clearer if the resemblance between the two arguments is borne in mind.

It might seem at first sight that the two additional sources of ignorance in this new problem must make a solution impossible; but by bold measures the difficulties can be overcome. We avoid the difficulty arising from our ignorance of the relation between the observations of the observers by assuming that the law we are seeking will retain its form unchanged whatever that relation may be; and we avoid that arising from our ignorance of the form of the law by assuming that there is only one law which will retain its form, so that if we can find any law that retains its form, we know that it must be the law of gravitation. The problem then which faces us is this: To find a law which will retain its form when the change is made from one observer to another whatever may be the relation between the observations of the two observers. That law, if we can find it, must be the law of gravitation which is to take the place of the familiar Newtonian law that the attraction is proportional to the product of the masses and the reciprocal of the square of their distance. If we can find the law, then we may be able to work backwards and deduce an extension of the second principle.

The method of attack is suggested by our previous conclusions. We saw that all those conclusions could be expressed, rather cryptically, but to the mathematician satisfactorily, by the single statement that the quantity s^2, given by (16·3), is unchanged if we make certain changes of axes or certain algebraic transformations consisting in the substitution for (x, y, z, t) of (X, Y, Z, T) which are functions of (x, y, z, t). Such transformations are associated with the change from one observer to another; the only transformations which were relevant before were those corresponding to the change between two observers with uniform relative velocity. These transformations are expressed by (7·1)—if the c is struck out in view of the change of units adopted on p. 54; but in order that there may be continuity between our previous conclusions and those to which we are about to proceed it will be well to write them in a rather different notation better adapted to our more general treatment. If we write (x_1, x_2, x_3, x_4) for (x, y, z, it), and now replace $-s^2$ by s^2, equation (7·2), stating that the left-hand side is unchanged by (or invariant for, to use the mathematical term) the transformation (7·1), is equivalent to the statement that

$$s^2 = \delta x_1^2 + \delta x_2^2 + \delta x_3^2 + \delta x_4^2 \quad \text{..............(22·1)}$$

is invariant for the transformation

$$\left.\begin{array}{l} x_1 = \lambda_{11} X_1 + \lambda_{12} X_2 + \lambda_{13} X_3 + \lambda_{14} X_4 \\ x_2 = \lambda_{21} X_1 + \lambda_{22} X_2 + \lambda_{23} X_3 + \lambda_{24} X_4 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{array}\right\} \quad \text{......(22·2)},$$

if the λ's, being independent of the X's, are such that

$$\left.\begin{array}{l} \lambda_{11}^2 + \lambda_{12}^2 + \lambda_{13}^2 + \lambda_{14}^2 = 1 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{array}\right\} \quad \text{...........(22·3)},$$

$$\left.\begin{array}{l} \lambda_{11} . \lambda_{21} + \lambda_{12} . \lambda_{22} + \lambda_{13} . \lambda_{23} + \lambda_{14} . \lambda_{24} = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{array}\right\} \quad \text{......(22·4)},$$

where all λ's which have not the suffix 4 are real, those which have it once, imaginary, while λ_{44} and the determinant formed of all the λ's are real and positive. All that is merely a very complicated way of saying that what is said in (7·1) and (7·2); and the transformation (22·2), subject to the conditions stated, is that corresponding to a change between observers in uniform relative motion. The change between two observers relatively accelerated will correspond to some other transformation. We do not know what this transformation is; but we do not need to know, because the bold assumption that we have made is equivalent to the statement that any conclusions we are going to draw must be valid for all possible transformations. (We may note in passing that if we regard all transformations as relevant, we are practically regarding all motions as equivalent. We may withdraw any limitation, expressed or implied, to observers moving with *uniform* relative acceleration, and assume that the form of the law we are seeking will retain its form for all observers however they are moving relatively to each other.)

Let us consider then the effect upon s^2 given by (22·1) of making any transformation whatsoever, that is to say, or substituting for (x_1, x_2, x_3, x_4) quantities (X_1, X_2, X_3, X_4) related to them by the equations

$$\left.\begin{array}{l} x_1 = f_1(X_1, X_2, X_3, X_4) \\ x_2 = f_2(X_1, X_2, X_3, X_4) \\ \cdots\cdots\cdots\cdots\cdots\cdots \end{array}\right\} \quad \text{.............. (22·5)},$$

where the f's may be any functions whatever—or at least any analytic functions, possessing differential coefficients. For this purpose, let us cease to deal with finite differences between the coordinates δx_1, etc., and a finite s, and deal only with infinitesimal differences dx_1, etc., and an infinitesimal ds.

We then write for (22·1)

$$ds^2 = dx_1^2 + dx_2^2 + dx_3^2 + dx_4^2 \quad \text{........... (22·6)}.$$

For brevity let us substitute the suffixes m, n, o, p for the suffixes 1, 2, 3, 4, and agree that any one of the letters may represent any one of the numerals. Thus X_m will mean the collection of (X_1, X_2, X_3, X_4), ΣX_m will mean $X_1 + X_2 + X_3 + X_4$. We shall shortly have to use multiple suffixes. Then X_{mnop} will mean the collection of $(X_{1234}, X_{2341}$, etc., etc.) and $\sum\limits_{m,n,o} X_{mnop}$ will mean the sum of those members of the collection which all have the same last suffix, e.g. X_{1234}, X_{2314}, etc., etc. Generally letters appearing under the Σ are different for different terms of the sum; the remainder are the same for all members. It is unnecessary to set out all the forms that will be used; the reader will be able to follow the natural development of the notation. According to this notation we write (22·2)

$$x_m = \sum_n \lambda_{mn} X_n \quad \text{................. (22·2·1)}$$

and write (22·4)

$$\sum_n \lambda_{mn} . \lambda_{on} = 0 \quad \text{................. (22·4·1)}.$$

We have then from (22·5), which we now write

$$x_m = f_m(X_n) \quad \text{.....................(22·5·1)},$$

$$dx_m = \sum_n \frac{\partial x_m}{\partial X_n} dX_n = \sum_n \frac{\partial f_m(X_n)}{\partial X_n} dX_n \quad \text{.........(22·7)}.$$

$\frac{\partial x_m}{\partial X_n}$ is, of course, a function of x_m or X_m, changing in general as they change. But when we are considering infinitesimal changes only, they may be regarded as constants throughout those changes. Substituting from (22·7) in (22·6) we obtain

$$ds^2 = \sum_{m,n} g_{mn} dX_m \,.\, dX_n \quad \text{..............(22·8)},$$

where the coefficients g_{mn} (the g's as we shall call them) are squares and products of the partial differential coefficients $\frac{\partial x_m}{\partial X_n}$. The g's thus depend (1) on the nature of the transformation (22·5), and (2) on the values of the coordinates x_m or X_m in the neighbourhood of which the infinitesimal differences lie; but for any particular transformation and in any particular neighbourhood they are constant. (22·8) therefore differs from (22·6) in the fact that in the latter the g's are the same for all transformations considered and are constant over the whole range of the coordinates; in fact (22·8) reduces to (22·6) if we put

$$g_{mm} = 1, \quad g_{mn} = 0 \quad \text{.................(22·9)}.$$

Accordingly the result of the most general transformation is simply to change the values of the g's in (22·8); the particular group of transformations (22·2) associated with uniform relative velocities are a special set for which the g's happen to retain throughout the transformations the special constant values given by (22·9). But if we are not bound to particular values of the g's, but allow that the g's may change their numerical values during a transformation involving infinitesimal differences, then we can say that, even in the most general transformation, ds^2 given by (22·8) is invariant. By that we mean that the expression retains its form and the numerical value of the whole during the transformation, although the numerical values of the coefficients change. This kind of invariance is quite familiar to the mathematician, but the reader untrained in analysis should note carefully that it is not quite the same or so simple a kind of invariance as that which we considered on p. 77.

There s^2 was unchanged in numerical value when we replaced x_m by X_m and made no other change whatever; now when we make the substitution we have at the same time to change the g's.

23. The varieties of 'space'. But there is a question here which needs careful attention. On p. 58 we saw that there was an analogy, not very complete, between s and a distance; this analogy indeed gave s its importance. In the argument that is about to be set forth, this analogy is even more important than it has been hitherto; all the developments of the general theory of relativity arise from the analogy between s and a distance. We must ask therefore whether the analogy is preserved when we substitute (22·8) for (22·6); are there any expressions, relating a distance to coordinates, similar to (22·8), when the g's have values other than those given by (22·9)?

There certainly are such expressions when there are only three coordinates. For suppose that we start with the Cartesian orthogonal coordinates and a distance given by (22·6), and then substitute the Eulerian polar coordinates (r, θ, ϕ). The distance between two neighbouring points is now given by

$$ds^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta\, d\phi^2 \quad\text{............(23·1)},$$

which is identical with (22·8) if we write (r, θ, ϕ) for (X_1, X_2, X_3) and put

$$g_{11} = 1,\ g_{22} = X_1^2,\ g_{33} = x_1^2 \sin^2 X_2 \quad\text{........(23·2)},$$
$$g_{12} = g_{23} = g_{31} = g_{13} = g_{32} = g_{21} = 0.$$

Since these values of the g's are constants or functions of the coordinates, they are values which might arise from such transformations as we are considering. In fact they arise from the transformation

$$\left.\begin{aligned} x_1 &= r\sin\theta\sin\phi \\ x_2 &= r\sin\theta\cos\phi \\ x_3 &= r\cos\theta \end{aligned}\right\} \quad\text{.................(23·3)}.$$

Accordingly ds may retain its character as a distance if the transformations we are considering are analogous to changes from one set of coordinates to another. But this result is not very helpful. In the first place, it is not possible to represent all transformations as changes of coordinates, even if we take into account all conceivable kinds of coordinates; certain conditions have to be fulfilled in order that such representation should be possible[1]. But more serious is the objection that the physical meaning of this kind of transformation is unsatisfactory. It is intolerable to suppose that in translating the observations of one observer into those of another, we have to make a change analogous to that from orthogonal to polar coordinates; we may expect that the observers can use any kind of coordinates in making their measurements without invalidating the translation, but we must expect also that, if there is to be coherence between the observers, both observers must use the same kind of coordinates. Accordingly we pass over this method of preserving the analogy with distance and consider another.

The second method arises from the fact that we habitually use the term distance in more than one sense. The primary meaning of the distance between two points is the length of the straight rod lying between them, or (if the points are separated by a solid body) the length of the straight rod which will fit between calipers set so as to touch the two points. But when we speak of the distance between two points on some surface, e.g. the surface of a sphere, we do not mean this distance; we mean the distance measured along the surface by means of a flexible tape pulled taut. This distance is the same as the distance measured by calipers only if the surface is plane. Now the relation between this 'surface-distance', as we shall call it, and the coordinates is not (16·1) even if we use orthogonal coordinates, corresponding very closely to the Cartesian (x, y, z). Thus orthogonal coordinates on the surface of a sphere are the arcs, measured along the equator and a meridian, by which it is possible to pass from the origin to the point in question; these arcs (which may be termed longitude λ and latitude β) are surface-distances, just as (x, y) are distances. These coordinates are then precisely analogous to Cartesian coordinates, and yet the surface-distance ds' between two neighbouring points expressed in terms of them is given, not by $ds'^2 = dx^2 + dy^2$, but by

$$ds'^2 = d\beta^2 + \cos^2\beta\, d\lambda^2 \quad \text{..............(23·4)}.$$

[1] One obvious condition is that all the functions f should be single-valued, so that to a single point defined by one set of coordinates corresponds a single point defined by the other. Another condition is that to every point defined by one set must correspond a point defined by the other. The complete mathematical expression of these conditions is complicated; but it will be readily realised that there are conditions.

Here again we have a distance in which the g's are not constants, 0 or 1, but are functions of the coordinates. If we applied the same principles of any other surface, e.g. an ellipsoid or a cone, always taking as coordinates (x_1, x_2) surface-distances at right angles, we should always again find that the surface-distance between two neighbouring points was given by

$$ds'^2 = g_{11}\,dx_1^2 + 2g_{12}\,dx_1 \,.\, dx_2 + g_{22}\,dx_2^2 \qquad \ldots\ldots\ldots(23{\cdot}5),$$

where the g's were functions of the coordinates, the nature of these functions varying with the kind of surface. They would in general also vary with the position of the points on it; for the universality of the relation (23·4) for the sphere is a consequence of its characteristic property, constant curvature; the form of the g's is in fact largely determined by the curvature of the surface in the neighbourhood of the points in question.

Here we have a much more satisfactory analogy and one that leaves us much more latitude than that based on change of coordinates; for there is no reason to believe that any set of g's, resulting from any transformation whatever, might not be characteristic of such relations between surface-distances. Of course the difficulty recurs that we know nothing about such relations as (23·5) when there are four coordinates; we do not know anything about them even when there are three; surface-distances are a characteristic of surfaces, which are necessarily two-dimensional. But again this does not trouble a mathematician; he can readily conceive of similar mathematical relations holding when there are three or four coordinates, and he therefore says that he can conceive of different kinds of three- or four-dimensional 'spaces' corresponding to different kinds of surfaces. Here we cannot hope to follow him; for, once more, there is all the difference in the world for us between a surface-distance in two dimensions which we can measure with a flexible tape (or a closely fitting polygon of straight rods) and a 'space-distance' (as I suppose we should call it) in four-dimensions which is simply not a measurable quantity at all. But enough has already been said of this difference; let us try to adapt ourselves to the mathematician's way of thinking.

Our new idea, then, is that the transformations we are considering are analogous to a change from one kind of space to another, the same kind of coordinates (orthogonal, polar, or so on) being used in each. The different observers, relatively accelerated, are making observations in spaces of different kinds, or, perhaps it would be better to say, trying to fit the observations that they make to different kinds of spaces; the translation of the observations of one observer into terms of those of another is very roughly analogous to the making of a plane map from measurements made on the curved surface of the earth. The g's appearing in the expression for ds^2, determine in what kind of space the observer is making measurements; ds^2, involving the g's, describes the kind of space in which the corresponding observer is working.

24. **Spaces and fields of force.** But the g's have another significance. Different accelerations of the observer O_2 relative to O_1 correspond to different transformations for the translation of O_2's observations into terms of those of O_1. We do not know what difference in acceleration corresponds to what difference in the transformation; but we do know that a difference in one corresponds to a difference in the other. Different transformations again mean different g's. But the acceleration of O_2 relative to O_1 (who, we may suppose, experiences no gravitational field) determines, in accordance with the principle of equivalence, the gravitational field that O_2 will experience. Accordingly the g's must be measures in some sense of the gravitational field. If we start with an observer who experiences no field and transform to relatively accelerated observers who experience a field, we must be able to express the field that they experience in terms of the g's. And if there is any general law of gravitation, valid for all these observers, it must be expressible in terms of the g's, together with the coordinates; for the field will vary with the coordinates. Expressed in terms of the geometrical analogy, this conclusion is that the kind of space in which an observer works is determined by the gravitational field that he experiences. In fact the gravitational field is nothing but a kind of space; the difference between two gravitational fields is analogous to that between a plane and a curved surface. In the absence of a gravitational field, all bodies move with uniform velocity, and the appropriate transformation is (22·2—22·4), which corresponds to what mathematicians call a 'flat' space. (We need

not inquire more precisely why they call it flat.) If there is a gravitational field, the space is no longer flat, but becomes curved or wrinkled; what we experience as a gravitational field is, according to this view, simply a curvature or wrinkling of the space in which we are working. We may assume that the space in which we are working is flat (just as the observer falling freely to the earth assumes that there is no gravitational field); but then the space which to some other observer (for example one at rest on the earth) appears flat will appear to us wrinkled. There is no absolute flatness or curvature of space; it is all a question of point of view. But the view that an observer takes, or the particular portion of space that he assumes to be flat, will determine the gravitational field which he finds at any other part of the space.

That is a general description of the startling ideas which have attracted so much popular attention to the general theory of relativity. They are certainly helpful in understanding the process by which the conclusions of that theory are derived; indeed the process can hardly be followed without them. But, as has been insisted, the analogy is not complete, and if it is pressed too far, we either have to pretend to believe propositions which, as physicists, we know to be false, or we have to suppress our objections at the grave risk of landing ourselves in hopeless confusion of thought. In what has been said already there are several grave difficulties; but these the reader had better discover for himself. What we are concerned with here is to understand what the argument is; its value can be estimated when we know its conclusions.

25. **The theory of tensors.** An account of the geometrical interpretation was introduced at this point in order to show that, in assuming the invariance of (22·8), we were not abandoning the fundamental notions on which our argument is based. But the only result of it that we need immediately (and this is a result which could have been reached without any use of it) is that the law of gravitation that we are seeking must be expressible in terms of the g's and the coordinates. In addition we know two other things about the law; first that it must retain its form for all observers; second (this assumption has not been explicitly stated before, although it has often been implied) that it must be such

that it is possible that, for some observers and for some values of the coordinates, there is no gravitational field. For we know or guess from experience that observers infinitely distant from any of the material systems they observe and at rest relative to them are aware of no such field.

The process of reasoning whereby from this scanty knowledge we deduce the form of the law is purely analytical and consists of the application of a branch of pure mathematics, due originally to Riemann, known as the theory of tensors. The main argument must here be interrupted to give an outline of this theory, very brief, but sufficient for our purpose.

We may start by defining a tensor as a set of four quantities (a_1, a_2, a_3, a_4) which are functions of the coordinates (x_1, x_2, x_3, x_4) and are such that when (X_1, X_2, X_3, X_4), given by (22·5), are substituted for (x_1, x_2, x_3, x_4), they become (A_1, A_2, A_3, A_4), where

$$\left.\begin{aligned} a_1 &= \lambda_{11}A_1 + \lambda_{12}A_2 + \lambda_{13}A_3 + \lambda_{14}A_4 \\ a_2 &= \lambda_{21}A_1 + \lambda_{22}A_2 + \lambda_{23}A_3 + \lambda_{24}A_4 \\ &\dots\dots\dots\dots\dots\dots\dots\dots \end{aligned}\right\} \dots\dots\dots(25\cdot1),$$

or generally
$$a_m = \sum_n \lambda_{mn} A_n \dots\dots\dots\dots\dots\dots(25\cdot1\cdot1),$$

the λ's being independent of the x's. a_m and A_m are by definition the same tensor; and a_1, a_2, a_3, a_4 or A_1, A_2, A_3, A_4 are called the components of this tensor. It follows from (25·1) that if one set of components of the tensor, a_m, are all zero, then any other set of components, A_m, derived from it by substituting X_m for x_m, are all zero. When all the components are zero, the tensor is said to be zero. The substitution of X_m for x_m is what we call a transformation; accordingly (25·1), which is the definition of a tensor and is supposed to be true whatever function X_m is of x_m, states that a tensor which is zero before transformation is zero after transformation. This property of a tensor is termed covariance or contravariance. It is a form of invariance; for it involves that something (namely the zero value) is unchanged by the transformation. But it is not the form of invariance we have considered so far; for that invariance involved the preservation of a numerical value during transformation. The only numerical value preserved here is 0, which is incidentally the only numerical value in which any right-thinking mathematician takes an interest.

The use of the term 'components' suggests that tensors are analogous to vectors. They are so analogous up to a certain point; in fact if there were only 3 components and 3 coordinates, which were orthogonal distances, the tensor defined by (25·1), subject to the conditions (22·3), (22·4), would be a vector. But it must be noted that there is not, as there is with vectors, any relation between the components and something that can be called the whole tensor. There is nothing corresponding to the relation that the sum of the squares of the components is equal to the scalar magnitude of the vector. A tensor simply is the group of the a's and A's associated by (25·1); it is essentially multiple not individual.

So far we have stated nothing about tensors; we have merely defined them. (In defining them we have implied something, namely that there is no essential self-contradiction in the definition; but we need not stop to examine this matter; the definitions can be shown to be self-consistent.) We have said nothing about the existence of such tensors as we have defined. But we can discover immediately that at least one tensor exists, namely the space-time interval dx_m or dX_m defined by (22·7). For, throughout any infinitesimal change, $\frac{\partial x_m}{\partial X_n}$ is a constant independent of x_m and may be identified with λ_{mn} in (25·1); (22·7) is then identical with (25·1) if we write a_m for dx_m and A_m for dX_m. Having thus found one tensor we guess that there may be other similar tensors defined by equations similar in form to (25·1). All such tensors we propose (as a mere definition) to call contravariant tensors of the first rank. The definition of such a tensor is therefore that

$$a_m = \sum_n \frac{\partial x_m}{\partial X_n} A_m \qquad (25\cdot2).$$

It now occurs to us that there may be tensors defined by

$$a_m = \sum_m \frac{\partial X_n}{\partial x_m} A_m \qquad (25\cdot3).$$

Such tensors, if they exist, are to be called covariant tensors of the first rank. (It is usual to distinguish in notation between contravariant and covariant tensors, writing the former with the suffix at the top; thus a_m covariant, a^m contravariant.)

Again it occurs to us that there may be tensors, defined by

$$a^{mn} = \sum_{o,p} \frac{\partial x_m}{\partial X_o} \cdot \frac{\partial x_n}{\partial X_p} \cdot A^{op} \quad \text{..............(25·4)},$$

$$a_{mn} = \sum_{o,p} \frac{\partial X_o}{\partial x_m} \cdot \frac{\partial X_p}{\partial x_n} \cdot A_{op} \quad \text{..............(25·5)},$$

$$a_n^m = \sum_{o,p} \frac{\partial x_m}{\partial X_o} \cdot \frac{\partial X_p}{\partial x_n} \cdot A_p^o \quad \text{..............(25·6)}.$$

These, if they exist, are to be called contravariant, covariant and mixed tensors of the second rank. They have 16 components; and so we may go on to define tensors of the third rank with 64 components, of the fourth rank with 256 and so on. But the formulae for these need not be set out.

But still we only know of one tensor, namely dx^m or dX^m. Can we find others? We can, by means of rules for the combination of tensors which are derived from the definitions of the tensors. It follows from these definitions, by ordinary mathematical reasoning, that certain combinations or functions of tensors are also tensors. Thus it is obvious to anyone that the sum of two tensors of the same kind, if they exist, is another tensor of the same kind, the components of the tensor which is the sum being the sums of the components of the original tensor, taken in pairs. Another rule that can be deduced almost as easily is that the product of two covariant or contravariant tensors of rank a is a covariant or contravariant tensor of rank $2a$, the components being the various terms in the products of the components of the original tensors. Similarly we get rules for determining whether we obtain another tensor and, if so, of what kind, by differentiating a previous tensor with respect to the coordinates. The general consequence of these rules is that, if only we can start with one or two tensors known to us, we can by multiplying them, adding them, differentiating them and so on, arrive at other tensors of other kinds and ranks in virtue of the mere fact that the original collections are known to be tensors; and just as we have to introduce a zero into arithmetic in order to make its rules perfectly general and can then prove from those rules what the properties of a zero are, so here we can prove that a tensor of zero rank, consisting of one 'component' only, is an invariant in the old sense, the numerical value of which is unchanged by transformation.

That is all we need of the general and very difficult analytical theory of tensors. Let us now apply it.

26. **The law of gravitation.** The ideas involved in the application are quite simple. If we can find a tensor G which is a function of the g's and the coordinates and which is zero for the special values of the g's characteristic of flat space (i.e. $g_{mm}=1$, $g_{mn}=0$), then the equations obtained by putting this tensor (i.e. all its components) equal to zero is the required law of gravitation. For since G is a tensor, the equations $G=0$ are unchanged by a transformation from x_m to X_m; the first condition of p. 76 is satisfied; and since $G=0$ for flat space, the law is true for flat space and the second condition is satisfied. We have assumed (p. 76) that there is only one law fulfilling the necessary condition; consequently $G=0$, if it is a law at all, must be the law we seek. Let us then try to find G.

We start from (22·8),

$$ds^2 = \sum_{m,n} g_{mn}\, dX_m \,.\, dX_n.$$

We know that ds is invariant; therefore it is a tensor of zero rank. We know that dX_m, dX_n are both contravariant tensors of the first rank. It follows from the rules of tensor combination that g_{mn} is a tensor, covariant of the second rank. It is called the 'fundamental tensor'. But g_{mn}, although a function of the g's, is not the tensor G, for it is not zero when $g_{mn}=1$; we must seek further. The process of search need not be detailed here. It suffices for our purpose that by differentiating the g's with respect to the coordinates, multiplying, adding and performing other algebraic operations on them we can arrive at a tensor B^p_{mno}, mixed, of the fourth rank, three parts covariant and one part contravariant, which is such that $B^p_{mno}=0$ is a necessary and sufficient condition that $g_{mm}=1$, $g_{mn}=0$; if and only if the g's have these values is B^p_{mno} zero, and it is the only tensor which satisfies this condition. It is called from its discoverers the Riemann-Christoffel tensor. It is immensely complex, having 256 components, each of which consists of 20 or more terms; typical terms (expressed as well as they can be with the notation developed so far) are such combinations as

$$g_{mn}\frac{\partial g_{op}}{\partial x_o}.$$

But B^p_{mno} is not the tensor G; for $B^p_{mno}=0$ is true only for flat space. If we took it as our law of gravitation, we should have to conclude that all space was flat and that there was no gravitational field anywhere. Such a conclusion is absurd. G must be such that while $G=0$ for flat space, there are other values of the g's, not those for flat space, which also make $G=0$; only if that condition is fulfilled will it be possible to have some coordinates for which there is no field and others for which there is a field. But having found B^p_{mno} it is easy to find another tensor for which $g_{mm}=1$, $g_{mn}=0$ is a sufficient but not a necessary condition of vanishing. In fact it is too easy, for there are many such tensors. If they are all laws, our assumption that there is only one law satisfying the necessary conditions is false. Accordingly, in order to maintain that assumption, we must pick out one of these tensors and state (or if possible prove) that this tensor G alone is such that $G=0$ is a law. Mathematicians have no difficulty in making such a choice; among all the alternatives there is one so sharply distinguished from the remainder that it appears to them the natural and obvious choice, if a choice has to be made. The reason for this view can be appreciated only by those thoroughly versed in the theory of tensors and possessed of the pure mathematician's instinctive feeling for symbols. But it must be observed that the reason is purely mathematical; it has nothing to do with the possibility of interpreting mathematical statements in terms of physical concepts; the choice is made because the alternatives are mathematically, not physically, unacceptable. At this point the argument becomes more difficult for the physicist to follow than at any other, although elsewhere the difficulties are more apparent; here he has simply to accept a judgement the reasons for which he cannot hope fully to understand.

The tensor B^o_{mno} or G_{mn} as it is usually written, which is thus selected as the expression of the law of gravitation, is a covariant tensor of the second rank. It has 16 components, but of these six vanish identically in virtue of the symmetry of dX_m, dX_n in the original tensor (22·8). $G_{mn}=0$ therefore consists of 10 significant equations; but of these again only six are independent. We are left finally with these six independent differential equations, involving the g's and the coordinates, as the law of gravitation for which we are seeking.

But in what sense is it a law? A law is a proposition which states relations between observations. The g's, involved in the 'law' $G_{mn}=0$, cannot be measured by direct observation; the analogy on which their introduction is based suggests only that they might be determined by measuring ds and the coordinates dX_m, and finding the relation between these measured magnitudes. Let us then get rid of the g's by solving the equations $G_{mn}=0$, obtaining thereby the g's as functions of the coordinates, and substituting these functions in (22·8); this process can be performed by the ordinary operations of mathematical analysis, and yields us a relation between ds and dX_m which is precisely equivalent logically to the original relations $G_{mn}=0$. Is this equation a law? No; it is still not a law; for we cannot measure ds directly. This is where the analogy breaks down, as we saw on p. 82; the analogy suggests that ds is a distance which, like the real distances of ordinary space measurable with straight rods or calipers, can be determined by direct measurement. But it is not; it is merely a name for the function of the dX_m on the right-hand side of (22·8). We cannot test the law by examining whether that relation is actually true; its truth is a matter of definition; there is no way of determining ds, even if we can measure dX_m, except by assuming that it is true. Much of the difficulty and obscurity of current mathematical expositions of the general theory of relativity arises from the mathematician's habit of treating ds as if it were something directly measurable.

And further, we cannot even measure dx_m. In (22·1) the δx's were measurable; three of them were differences between actually measured coordinates (which are the distances of a point from orthogonal axes), the fourth a time. (In (22·1) the fourth is an imaginary time, which is not measurable. But if we consider (16·2) instead of (22·1) δt is a measurable time. This further complication may be neglected for the moment.) But in making our transformations, we have substituted for these dx_m other quantities dX_m, which are or may be functions of all the dx_m; they may be functions of both distances and times, and we do not know what functions they are. We cannot possibly measure something which may be either a distance or a time or a combination of both; in order to set up the experiment whereby the measurement is made, we must know fully the physical nature of the magnitude we propose to measure.

Accordingly though we have obtained a relation which we believe expresses the law of gravitation, we have not obtained the law itself; we have obtained nothing that we can test by experiment. In order to obtain a true law, we must somehow get rid of ds, as we got rid of the g's, and turn dX_m into magnitudes (probably distances or times) of which we know the physical nature without ambiguity. For this purpose, we need a new assumption, which arises very directly out of the analogy on which the whole argument is based; if that analogy can be accepted, it is natural and obvious.

27. **The world-line of a particle.** The law of gravitation is usually expressed as a relation between the force and the distance from a gravitating particle. But if, as usual, we combine this law with the laws of dynamics, we might state it equally well by describing the path of a particle moving in the gravitational field. Thus, to say that two gravitating particles move about their common centre of gravity in an ellipse with a particle at one focus and in such a way that the angular momentum about the focus is constant, is precisely equivalent to saying that the force varies as the inverse square of the distance. If we add a proposition stating how the angular momentum is related to the masses, we are adding the proposition that the force varies as the product of the masses. We are going now to express the law of gravitation in this form by describing the path of a particle moving in a gravitational field. The argument runs thus. In the absence of any gravitational field the path of a particle is a straight line in flat space. By a straight line, we mean here a path defined by a linear relation between the coordinates, so that $\frac{dx_m}{dx_n}$ is independent of x_m. But a straight line, defined in this manner, has a further important property when the space is flat; it is the 'shortest distance' between any two points on it. This property may be expressed analytically in terms of ds. Substituting $\frac{dx_m}{dx_n}\, dx_n$ for dx_m, we write (22·6) in the form

$$ds = \left\{1 + \left(\frac{dx_2}{dx_1}\right)^2 + \left(\frac{dx_3}{dx_1}\right)^2 + \left(\frac{dx_4}{dx_1}\right)^2\right\}^{\frac{1}{2}} dx_1 \quad \ldots\ldots(27{\cdot}1).$$

We now consider $\int_{x_1=a}^{x_1=b} ds$, and inquire by means of the calculus of variations what must be the form of the functions $\frac{dx_m}{dx_n}$ within the range ab in order that the integral shall be less if the functions have that form than if they have any form differing infinitesimally from it; that is to say, we inquire what form the functions must have in order that

$$\delta \int ds = 0 \quad \text{.......................(27·2).}$$

The answer is that the functions $\frac{dx_m}{dx_n}$ must be constant, which is equivalent to saying that the line defined by (27·2) is straight.

If the space is not flat, the straight line is not the shortest distance between two points; that is to say, if we substitute (27·1) for (22·6), the functions $\frac{dx_m}{dx_n}$ which make $\delta \int ds = 0$ are not those which make $\frac{dx_m}{dx_n}$ constant, unless the g's have the values characteristic of flat space. In curved space we must distinguish between straight lines and 'geodesics', as the paths of shortest distance, for which $\delta \int ds = 0$, are called. Now in the experimental geometry of curved surfaces, on which our analogy is based, geodesics play much the same part as straight lines play for plane surfaces; thus the great circles on a sphere, which are closely analogous to straight lines on a plane, are the geodesics of the sphere. In fact geodesics are the lines in which a flexible tape pulled taut will lie; surface distances are measured along geodesics. Accordingly we should expect the path which, in curved space, corresponds to the straight line in flat space to be the geodesic. Consider now O_1, who observes no gravitational field, and O_2, accelerated relatively to him, who observes a field. O_1 says that his particles move in straight lines (straight world-lines, be it understood, which imply also constant velocity): O_2 says they do not. But O_2 also says that the space which O_1 believes to be flat is not really flat; it is curved. Surely it is natural to suppose that the lines which to O_1 appear to be straight lines in flat space will to O_2 appear to be geodesics in the curved space.

This is the new assumption. The path of a particle in the curved space of which the gravitational field is either the cause or the expression (as we choose to view the matter) is a geodesic in that space. Analytically the assumption is that the path is such that $\delta \int ds = 0$, when the integral is taken by writing ds in the form (27·1) and integrating between two values of x_1 on the path[1]. We can now take a further step towards the solution of our problem. Having solved $G_{mn} = 0$ and substituted the resulting g's in (22·8), obtaining ds as a function of the coordinates dx_m, we seek some forms for differential coefficients $\frac{dx_m}{dx_n}$ which are such that when these forms are inserted in (27·1), $\delta \int ds = 0$. These forms for the differential coefficients, when we know them, define a relation between the coordinates which is the equation of a world-line in our four-dimensional space. We then know that this world-line is the path of a particle in a gravitational field. By this method we can only obtain at one time the path in a particular kind of gravitational field, e.g. that round an attracting particle or that in a continuous distribution of matter. We have in fact to pick out one particular solution of $G_{mn} = 0$, carry the process through for that solution; then pick out another, carry through the process again and so on.

But we are not yet quite out of the wood. Two problems remain. One has been noticed already. When we have got the equation to the path in terms of some coordinates, we shall not know exactly what those coordinates are; we cannot test our result until we do know. Again all that we shall find is that a certain equation represents the path in some kind of gravitational field; we shall not know with what distribution of matter this field is associated. Strictly we should have to search experimentally through all possible distributions until we found one which agreed with the predicted law; if we found such a one we might hold that our law was confirmed.

[1] $\delta \int ds = 0$ may mean that $\int ds$ is a maximum or a minimum. For a geodesic in the original sense it is a minimum; but mathematicians think the difference between a maximum and a minimum too trivial for notice, and they apply the term geodesic indifferently in the two cases. In any particular kind of space only geodesics of one kind occur, at least in the neighbourhood of a particular point. In the curved space characteristic of the gravitational field the geodesics always correspond to a maximum. This is, of course, yet another difficulty for the physicist, but it need not be the last straw.

(27·6) differs from (27·3) only in the presence of the second term on the right; (27·6·2) differs from (27·3·2) only in the substitution of s for t. We have not completely eliminated the meaningless s, but since we concluded on p. 19 that the numerical value of s is always very near that of t, and since s occurs in a constant term of which only the numerical value is important, we overlook the second difference. The first then represents the only correction we have to make to the law of gravitation, if we identify the gravitational field corresponding to the solution of G_{mn} adopted with that round a gravitating centre, and identify also the constants of integration m and h with the mass of that centre and the angular momentum in the orbit.

28. **Comparison with experiment.** Such is the guess that we make, and now at last we can compare our conclusions with experiment. For whereas (27·3) represents motion in an ellipse, (27·5) represents motion in a rather different orbit. The difference will be small in all experiments we can try, for 'experiments' on gravitational orbits round attracting centres are practically confined to observations of the solar system. Let us insert numerical values. The radius of the earth's orbit a is $1{\cdot}49 \times 10^{13}$ cm. Its period of revolution is $3{\cdot}15 \times 10^{7}$ secs. But since we have been using throughout as unit of time the period in which light travels the unit of length, its period in our units is $3{\cdot}15 \times 10^{7} \times 3 \times 10^{10} = 9{\cdot}45 \times 10^{17}$. Consequently its angular velocity ω is $2\pi/9{\cdot}45 \times 10^{17} = 6{\cdot}64 \times 10^{-18}$. The ratio of the second to the first term on the right of (27·5·1) is $\dfrac{3h^2}{a^2}$, where $h = a^2\omega$. Consequently the ratio is

$$3a^2\omega^2 = 3 \times (1{\cdot}49)^2 \times (6{\cdot}64)^2 \times 10^{-10} = 3 \times 10^{-8} \text{ about} \quad \ldots(28{\cdot}1).$$

Accordingly we can treat the new term introduced by our new law as a small correction. The same is true if we consider any other planet.

If (27·6) is solved treating this term as small, we find

$$\frac{1}{r}\cdot\frac{h^2}{m} = (1 + e\cos\gamma\phi) \quad \ldots\ldots\ldots\ldots\ldots(28{\cdot}2{\cdot}1),$$

where

$$\gamma = 1 - \frac{3m^2}{h^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots(28{\cdot}2{\cdot}2),$$

and e is the eccentricity of the ellipse to which (28·2) reduces when $\gamma = 0$. m is here expressed in gravitational units, so that the gravitational constant is 1; it follows from the definition of h in (27·3·2), that

$$\frac{h^2}{m} = a(1 - e^2) \quad \text{.....................(28·3)},$$

where a is the major axis of the ellipse, so that

$$\frac{m^2}{h^2} = \frac{a^2\omega^2}{1 - e^2} = v_0^2 \quad \text{.....................(28·4)},$$

where v_0 may be termed the mean velocity of the planet in its orbit. Consequently (28·2) becomes

$$\frac{a(1 - e^2)}{r} = (1 + e\cos\gamma\phi) \quad \text{............(28·5·1)},$$

where

$$\gamma = 1 - 3v_0^2 \quad \text{.....................(28·5·2)}.$$

In Chap. XV, p. 84, we arrived at a very similar equation for the orbit of an electron round a charged nucleus; the only difference is that here γ differs from 1 by $3v_0^2$, whereas in that equation it differed from 1 by v_0^2. The c has vanished because we are now expressing v_0 as a fraction of c (p. 54). The result in Chap. XV was derived from the special theory of relativity, and applied to electrical forces obeying the inverse square law. If we had not developed the general theory, we might have expected the same result to be true for gravitational forces, which obey much the same laws as electrical; but we find that the general theory predicts a term in γ three times as large. It must be noted that since we have specifically confined the general theory to gravitational forces, we have no reason to believe that our previous result was not true for electrical forces. There is nothing in our new conclusion to invalidate those described in Chap. XV.

But the trebling of the term in γ is of extreme importance. In Chap. XV we saw that the presence of this term meant that the axis of the ellipse rotated through the angle $\delta\phi$ during each revolution of the particle round the ellipse, where

$$\delta\phi = 2\pi\frac{1 - \gamma}{\gamma} \quad \text{.....................(28·6)}.$$

Now the axes of the orbits of most of the planets are rotating slowly in some such way as this. But for all except one of them, the observed rotation can be accounted for by perturbations due to neighbouring planets calculated according to the Newtonian theory[1]. But there is one exception, the planet Mercury. It has long been known that there is a slow rotation of the perihelion of Mercury which cannot be wholly accounted for on the Newtonian theory. Newcomb estimated the outstanding discrepancy as a rotation of 43″ per century. Now for Mercury $a = 5{\cdot}8 \times 10^{12}$ cm.; the period T is 88 days or $2{\cdot}28 \times 10^{17}$ of our units; e is 0·2.

Therefore, from (28·4), (28·5·2),

$$1 - \gamma = \frac{12\pi^2 a^2}{T^2(1-e^2)} = 8 \times 10^{-8} \quad \text{...........(28·7)};$$

so that the rotation per revolution is by (28·6)

$$2\pi \times 8 \times 10^{-8} \text{ radians} = 0{\cdot}103''$$

and the rotation per century is

$$0{\cdot}103'' \times 100 \times \frac{365 \; 2}{88} = 42{\cdot}9''.$$

The value predicted by the general theory of relativity accounts exactly for the outstanding discrepancy: that predicted by the special theory is only one-third. That is the first of the striking confirmations of the general theory.

The same test is not easily applied to other planets, partly because their mean velocity is less, and therefore $d\phi$ less, but more because their orbits are so much more nearly circular ($e =$ 0·09—·01) that a large rotation would pass unnoticed; for it is clear that if the eccentricity were zero and the orbit circular, a rotation of it could not be observed. But the predicted values are not inconsistent with observation. It should be mentioned here that since so much attention has been paid to the matter, there has been careful inquiry into the outstanding discrepancy for Mercury. The

[1] The reader may be surprised that all 'corrections' to the theory of gravitation produce the same result, namely a rotation of the axis; this is the result of the special theory and of the general theory of relativity and of mutual perturbations. The explanation is simple. Any very small departure from an elliptical orbit can be represented as (1) a change in the size of the orbit, (2) a rotation of the axis, or rotation of perihelion as it is called. All three 'corrections' produce departures of both kinds, but the first cannot be tested experimentally. It appears in our calculations as a change in the gravitating mass of the sun, and since we have no way of measuring the sun's mass except from the dimensions of the planetary orbits, we have no way of deciding whether the correction is or is not of the right magnitude. If the wrong correction were applied, there might be a slight discrepancy in the values deduced for the sun's mass from the orbits of different planets; but we do not know the relative dimensions of the orbits with sufficient accuracy to apply this test.

value 43″ is not that directly observed; it is only a residue after many other corrections have been subtracted and is therefore affected by the possibility of considerable error; in fact the probable error is about half the value. It is certain therefore that the very close agreement between prediction and observation is a mere chance; on the other hand the majority of astronomers seem agreed that any alterations that can be produced in the observed values by recalculation or improvement of the observations cannot bring that value further from the predicted value than can easily be accounted for by experimental errors.

We pass now to another test. It is fundamental in our theory that we can treat a ray of light like a moving particle. The peculiarity of the world-line of a ray of light, according to the special theory, apparent directly from (16·2), is that for it $ds=0$; that is to say there is only one point on the line. According to the general theory, an observer who experiences no field, because he is falling freely, is working in flat space, characterised by (22·1); for him the special theory is still applicable. Consequently he will still find $ds=0$ for a ray of light. But since ds is invariant, any other observer will also find $ds=0$, which is therefore the general expression for the path of a ray, and gives us a relation, in place of (27·2), to eliminate the meaningless ds. $ds=0$ implies from (27·5·2) that h is infinite, but we cannot apply (28·2) by simply putting $\frac{m}{h^2}=0$, for in deducing (28·2) the assumption was made implicitly (in part of the analysis not set out) that h is finite. We must go back to (27·4). Starting thence and considering motion in a plane (which our treatment of the motion of a particle shows to be a possible motion), we may omit θ and find for the equation to the path of the ray (writing again r for X_1, ϕ for X_3, and it for X_4),

$$\left(1-\frac{2m}{r}\right)^{-1}\left(\frac{dr}{dt}\right)^2+r^2\left(\frac{d\phi}{dt}\right)^2=\left(1-\frac{2m}{r}\right)\quad\text{......(28·8)},$$

m/r is small compared with 1 in the strongest gravitational fields we know; even at the surface of the sun it is only $2{\cdot}11\times10^{-6}$. Neglecting squares of this quantity, we find that (28·8) is the equation of a hyperbola with the small angle $\frac{4m}{r_0}$ between its asymptotes where r_0 is the distance of the apex of the hyperbola from the attracting centre. This hyperbola, if we insert the appropriate value of m, will be the path of a ray of light passing within a distance r_0 of the sun. In other words the ray in passing from an infinite distance on one side of the sun to an infinite distance on the other will be deflected through an angle $\frac{4m}{r_0}$ radians. If it grazes the surface, so that r_0 is the sun's radius, the angle of deflection is $8{\cdot}22 \times 10^{-6}$ radian $= 1{\cdot}74''$.

The paths of rays of light passing near the sun's surface can be traced by observing stars very near the sun when it is eclipsed by the moon. If the stars round the sun are photographed during an eclipse and also at some later or earlier period when the sun is not in that region of the sky, a comparison of the photographs will show whether the star images are displaced by the sun's presence, and, if so, by how much. It is needless nowadays to give any account of the carrying out of this experiment; the daily press has been full of it. The first observations were made in May 1919 in conditions good astronomically, but bad meteorologically. They were decidedly favourable to the theory. There was no doubt of a displacement in the right direction, and its amount was perfectly consistent with the theory. The final figures as given by Eddington for the quantity which is predicted to be $1{\cdot}74''$ as determined at two independent stations are $1{\cdot}98 \pm 0{\cdot}12$ and $1{\cdot}61 \pm 0{\cdot}30$. In the absence of any rival theory, predicting a different deflection[1], this may be regarded as complete confirmation. The observations, which are extremely delicate—the displacements measured on the actual photographic plates are only a few thousandths of a millimetre—are to be repeated on Sept. 21, 1922: the results may be available before this chapter is published.

29. **Generalisation of the second principle.** At the beginning of this section (p. 76) it was observed that its problems would be

[1] If a ray of light were subject to the same deflection in passing near the sun as a material particle travelling with the velocity c, then the quantity corresponding to $1{\cdot}74''$ would be half this amount, or $0{\cdot}87''$. But since no theory covering any considerable range of phenomena predicts this value, it can hardly be regarded as an alternative. If it proved the true value, we should be in a worse position than if there were no deflection at all; there would be no theory applicable to the optics of moving systems.

soluble in a much more direct manner, if we knew some proposition, corresponding to the second principle, which would relate O_1's space and time measurements to O_2's. We have now solved the problems, or some of them, and we should therefore be able to reverse the argument and to arrive at the relation between these measurements. Let us inquire then how the measurements of O_2, at rest relative to an attracting centre and therefore experiencing its gravitational field, are related to those of O_1, falling freely to the centre and therefore experiencing no field.

The basis of the comparison is provided by the two formulae for ds. O_1 finds (22·6) appropriate, O_2 finds (27·4) appropriate. In (27·4) we now identify (X_1, X_2, X_3) with (r, θ, ϕ) and X_4 with it. In order to compare it with (22·6) we must transform the latter to polar coordinates, which, if X_4 is identified with it, gives

$$ds^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta \,.\, d\phi^2 - dt^2 \quad \text{......(22·6·1).}$$

The relation we seek arises from our fundamental assumption that ds, the space-time interval between the same two events, is the same for the two observers, so that, writing as usual O_1's magnitudes with small letters and O_2's with large,

$$dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta \,.\, d\phi^2 - dt^2$$
$$= \left(1 - \frac{2m}{R}\right)^{-1} dR^2 + R^2 d\Theta^2 + R^2 \sin^2\Theta \,.\, d\Phi^2 - \left(1 - \frac{2m}{R}\right) dT^2 \quad \text{.. (29·1).}$$

We shall leave out the terms in $d\theta$, $d\Theta$, assuming as before that everything is in one plane.

First consider length measurements. O_1 and O_2 measure lengths by observing the positions occupied by two points at the same time. We therefore put $dt = dT = 0$. If the lengths measured are radial from the attracting centre we also put $d\phi = d\Phi = 0$, and find

$$dr^2 = \left(1 - \frac{2m}{R}\right)^{-1} dR^2 \quad \text{..............(29·2).}$$

That is to say O_2, in the gravitational field, finds a radial length shorter than O_1, in no gravitational field, finds the corresponding length, the ratio of dR to dr being $\sqrt{1 - \frac{2m}{R}}$ and decreasing as the attracting centre is approached. On the other hand, circumferential lengths will be measured by $rd\phi$, $Rd\Phi$, we put $dr = 0$ and find

$$rd\phi = Rd\Phi \quad \text{....................(29·3).}$$

O_1 and O_2 agree about circumferential lengths. Accordingly, when the lengths are the circumference and radius of a circle, if one of the observers finds the ratio of these lengths to be π, the other must find it different from π. Euclidean geometry (as follows naturally from our argument) cannot be true for both of them; we naturally suppose that it is true from the observer who experiences no field; the other observer then will find the ratio of the circumference to the radius of a circle to be greater than π[1].

Now consider time measurements. Suppose that O_1, falling freely, is only just beginning his fall, so that he, as well as O_2, is at rest relative to the system observed. Then we put

$$dr = d\phi = dR = d\Phi = 0$$

and conclude that

$$dt = \sqrt{1 - \frac{2m}{R}}\, dT \quad \text{.................(29·4).}$$

If O_1 and O_2 measure the time intervals between the same two events, O_1 will find that interval shorter in the ratio $\sqrt{1 - \frac{2m}{R}}$ to 1.

Such are our conclusions. Can we test them experimentally? No; we cannot. As was insisted on p. 3 we cannot hope really to compare with any high accuracy the observations of two observers who are moving, or just preparing to move, relative to each other. The second principle, we must remember, was introduced into the special theory only in order that we might afterwards eliminate the conceptions involved in it. Since we have got our results in the general theory without any use of it, no useful purpose is served in bringing it in.

However this is not the conclusion usually drawn. Einstein has drawn from an argument concerning time measurements, very similar to that just stated, a conclusion about the frequency of spectral lines emitted from the sun as observed on the earth; this conclusion might possibly be tested by experiment and it is therefore necessary

[1] This result is very similar to one that has been deduced from the special theory and given rise to much discussion. It has been argued from that theory that the circumference of a rotating wheel, moving in the direction of its length, must be shorter in the ratio $1/\beta$ than 2π times the radius of the wheel, which is moving perpendicular to its length. (See p. 23.) But this conclusion does not really seem to follow; for the special theory applies only to uniform velocities and the velocities here are not really uniform.

to consider it carefully. No similar test of the theory has been based on the conclusions concerning length measurements, and we shall pay no further attention to them; but the difficulties we are about to discuss apply equally to the length measurements.

30. **The 'Einstein shift'.** Before considering Einstein's argument let us carry ours a little further. It would be a highly reasonable guess that, since O_1 is falling freely towards the attractive centre and experiences no field, it cannot matter at what distance from the centre he is; he may even be so distant that he would experience no appreciable field even if he were not falling. In that case, we may identify O_1 with an observer on the earth and O_2 with an observer on the sun, and our conclusion is that any time interval, such as the period characteristic of a vibrating atom, will appear shorter to an observer on the earth than to one on the sun. But that does not tell us anything about the frequency of lines emitted from the sun as seen from the earth. For in the first place, we have said nothing whatever about the position of the system of which the time interval is characteristic, and nothing about its motion, except that both observers are at rest relative to it. What our arguments would prove is that the estimates of the two observers of the period of an atom will be related in the same way, whether the atom is on the sun or on the earth or infinitely distant from either. Our conclusion depends wholly on the circumstances of the observers, not of the source. There is nothing in (29·4) which would justify us in applying its conclusions to an atom close to O_2 rather than to one close to O_1. But, secondly, even if some reason could be given for applying (29·4) only to atoms close to O_2, another obstacle to any experimental test remains. (29·4) tells us only what is the relation between the periods as measured by two observers, one on the earth and one on the sun. But we know nothing whatever about the experiences of observers on the sun; we can only compare the periods of atoms on the sun with those of atoms on the earth; and to arrive at any experimental conclusion we must know some relation between the period of an atom on the sun, as observed by an observer on the sun, and the period of a similar atom on the earth as observed by one on the earth. At first sight we might be inclined to guess that these two periods

were equal; but there is nothing convincing about that guess; for after all the strong gravitational field of the sun may alter the periods of atoms just as strong electric or magnetic fields alter them. Moreover, as we shall see in a moment, this assumption, if we made it, would lead to a conclusion opposite to Einstein's. The ideas that we have developed so far seem to permit of no deduction whatever being made concerning the periods, measured on the earth, of two atoms one on the earth and one on the sun. This failure of the principles is a consequence of their inability to deal with static forces, which was discussed on p. 75.

But Einstein has made such a deduction. He predicts that the atom on the sun ought to appear to vibrate with a period less than that of the similar atom on the earth, so that lines in the solar spectrum should be shifted slightly towards the rod as compared with those in the corresponding terrestrial spectrum.

I cannot follow his argument and will not therefore try to state it completely; the reader must refer to his works. But it seems to depend on the identification of dt with the real period of an atom in the field characterised. Since ds is equal to dt when there is no field, and to $\sqrt{1-\frac{2m}{R}}\,dt$ in the field we are considering, it would follow from the invariance of ds that the real period of vibration of an atom on the sun is greater than that of one on the earth. (Of course Einstein would not admit openly that there is a 'real' frequency or a 'real' field; the fields and the frequency change with the observer. But it is only by introducing this conception, foreign to relativity, that I can follow the argument at all.) But there are two difficulties. First, it seems to be inherent in all the argument that we have considered that ds is invariant, not for a change from one observed system to another, but for a change from one observer to another. The invariance of ds tells us nothing about the relation between the real magnitudes of two different systems; it tells us only about the apparent magnitudes of the same system. Secondly, even if the real frequencies of the atom on the sun and the earth are related as Einstein supposes, the question still remains what is the relation between the real frequency (or other magnitude) of a system on the sun and its apparent magnitude to an observer on the earth. If the two are the same, again all the argument that we have considered breaks down. If they are not the same, Einstein's conclusion does not follow, and we have no means of deducing another; for having used the invariance of ds to inform ourselves about real frequencies, we have nothing left to inform us about apparent frequencies.

Unfortunately the matter cannot be settled by an appeal to experiment, for the observational evidence is conflicting. The displacement in question amounts only to about 0·01 Å. in the visible spectrum, and its measurement, if it exists, is greatly complicated by the existence of other possible displacements, due (e.g.) to the barometric pressure at the sun's surface. Some observers assert, others deny, that a displacement of the predicted amount is consistent with observation; and spectroscopists of equal skill and experience are to be found on either side. Objections to Einstein's interpretation of his theory in this matter have been put forward from many sources, and some of them are the same in substance, though different in form, from that urged here. In face of these objections Einstein maintains that the existence of the displacement is a necessary consequence of his theory. In that position he is inexpugnable. For the whole theory is based on arguments suggested by analogy to the minds of particular persons, of whom Einstein is the chief; if they say that these arguments do suggest this result to them, it is useless to argue; for it is admittedly impossible to present the theory as a logically complete and consistent structure, appealing with equal force to all. If experiment ultimately decides in favour of the 'Einstein shift', there will once more be evidence how futile are the protests of lesser minds against the dictates of genius; but at the same time—I speak here for myself—all hope for lesser minds to understand this particular product of genius will vanish.

31. **Further developments.** There I propose to stop. We have not, of course, exhausted the general theory of relativity, but we have considered all those parts of it which lead to results which can at present be brought to the test of experiment. Some of the developments that we have omitted may be mentioned very briefly. (1) The theory can be applied to distributions of gravitating matter other than a simple attracting centre. (2) The theory can be

developed into a general, powerful and (to good analysts) attractive method of discussing all the familiar dynamical problems of gravitation, just as the 'geometrical' form of the special theory can be developed into a general method of discussing electromagnetic problems. (3) With the theory in its present form there are certain difficulties which arise when we inquire into the gravitational field at an infinite distance from all attracting bodies; it appears that it is not a necessary consequence of the theory that this field should be zero. Slight modifications of the theory are introduced which lead to results which can be interpreted as setting a limit to the possible dimensions of the universe. Two views, both startling and both attractive, have been put forward by Einstein and de Sitter; expert opinion appears on the whole to consider the latter the more plausible. But no account of them is proposed here. (4) Lastly there is the theory of Weyl, who by divorcing yet further the space of the mathematician from that of the physicist, is able to show that the electromagnetic as well as the gravitational laws are mere expressions of the kind of space in which we observe.

All this is very interesting, but very difficult. I am quite sure that anyone who finds difficulty in understanding thoroughly the simpler parts of the theory, which are considered here—in understanding them, not only when he reads about them, but when he tries to reproduce them or to work with them—would be wasting his time in attempting to grasp these higher developments. On the other hand, when he has so understood them, he will be far beyond any help that I can give; he can go with confidence to the original sources.

V. SUMMARY

32. **The special theory.** The question remains whether the theory of relativity is true. As I have explained in the preface, I am not going to attempt an answer to that question; but it is permissible to consider in a very general survey the features of the theory on which any answer must depend.

When we are considering the truth of any scientific doctrine two questions must be answered. Is it in accord with the facts? If so, is it the only doctrine in accord with the facts, or are there alternatives? Only when we have come to a decision on these matters, are we in a position to make up our minds about its truth. Let us try to decide them for the theory of relativity. In making the attempt it will be convenient to discuss separately the special and the general theory. Objection to this course may be taken by those who hold, as apparently Einstein holds, that the two parts of the theory are inseparable, and that if one of the two parts is rejected the other must necessarily share its fate. But that view can only be sustained if the geometrical interpretation of the special theory, due to Minkowski, is regarded as an essential part of it. The special theory was in existence before Minkowski interpreted it, and it is possible to discuss it apart from that interpretation. It will be so discussed here; and if it is said that what we are discussing is not properly called the 'special theory', that is a mere question of words which we may leave on one side.

Concerning the special theory, as we regard it, the two questions can be answered definitely and, I believe, with the assent of all those competent to judge. First, the theory[1] is in accord with the facts. There is nothing ascertainable by experiment predicted by the theory which is not admitted, even by the advocates of 'alter-

[1] I am not at all sure that the special theory, in the limited sense that we are attributing to it, is properly termed a theory; I should prefer to call it a principle, and so connect it with the principle of the conservation of energy or the principle of least action. But that is a rather abstruse matter unsuitable for discussion here; the reader can judge for himself whether the differences between the special theory and the general theory (which I admit to be a theory) are sufficient to justify a difference in nomenclature.

natives', to be actually true. It is actually impossible to detect the earth's motion through the aether; the Michelson-Morley and Trouton experiments do give negative results; aberration and the Doppler effect do occur as indicated; there is no reason to doubt the accuracy of Rowland's or Roentgen's or Wilson's experiments on the electric and magnetic properties of moving bodies; the Fizeau experiment, recently been repeated with great elaboration and accuracy, has given results wholly concordant with the special theory; and the mass of a moving electron does vary with its speed as relativity predicts. Further the success of the special theory combined with the Bohr-Sommerfeld theory of spectra, though a less direct, is almost a more striking confirmation. Against this evidence there is none to be set which throws any doubt on the theory.

Moreover the special theory really does predict these results which are so completely confirmed. Nobody disputes that once we have stated the two principles, or at least when we have converted them into an analytical form, the results follow by a rigid process of mathematical analysis which anyone can appreciate. The assumptions can be stated formally and made the basis of pure logical deduction; there are no vague guesses which cannot be put into words or symbols.

The only possible objection that can be urged against this answer which we have given to the first question is that the theory predicts absurd results for certain experiments which cannot actually be tried. None of these experiments have been mentioned here, though they bulk largely in most expositions of relativity; for I cannot see that the theory 'predicts' them. Maxwell's demon could make thermodynamics look very foolish if molecules were as large as footballs. But thermodynamics does not predict anything about the size of molecules; it predicts that a gas with molecules of less than a certain size will have certain properties; it says nothing whatsoever about what sizes will actually occur, and therefore nothing about sizes which do not actually occur. In the same way the special theory asserts that if systems are moving relative to each other with certain uniform velocities, certain results will follow. It does not say anything about what velocities occur and therefore nothing about what happens at velocities which do not occur. Indeed this is an understatement of the position; for though it does not assert, it does very distinctly imply, that the velocities which would have to occur in order that the absurd experiments might be made can *not* actually occur.

Secondly, there are no alternatives to the special theory; the question of alternatives does not arise; what are sometimes regarded as alternatives are really supplements. All that the special theory, as we are interpreting it, states is that if we make certain assumptions certain results follow; it is not the less true because the same results will follow from other assumptions. Thermodynamics is not the less true because some of its consequences can be deduced from the dynamical theory of gases; all that is thereby proved is that the dynamical theory is consistent with thermodynamics. All the supposed alternatives (which it is not the purpose of this chapter even to enumerate completely), the Maxwellian theory together with the Lorentz-Fitzgerald contraction, the manifold aethers of Lenard, the asymmetrically propagated light of Paschsky—all these theories, in so far as they predict true results, are merely special doctrines conforming to the fundamental principles of the special theory. They attempt something which the special theory, as we are regarding it now, does not attempt; they attempt explanation and their explanations are an essential part of them. Explanation, though it may be a bye-product, is not an essential part of the special theory. And if it is asked, What then is the use of relativity, the reply may be, What is the use of thermodynamics. The use is the coordination of a large number of separate laws which, if regarded apart from the principles, are unconnected. It is no less notable an achievement to show that the variation of the mass of an electron with its speed is a consequence of the same principles as the Fizeau effect than to show that the plastic flow of glaciers is a consequence of the same principles as the bursting of domestic water pipes; the achievement remains although we may not know *why* observers may differ in their time judgements or why the characteristic of water is abnormal in the neighbourhood of the triple point.

33. **The general theory.** Now let us turn to the general theory. The answer to the first question is here more difficult. There is no fact known to be inconsistent with the theory; on the other

hand there is only one fact, namely the deflection of light passing near the sun, which can be investigated so accurately that the agreement between theory and experiment is really striking. But one fact, if it is sufficiently definite, may be conclusive. If the observations at the forthcoming eclipse confirm those already obtained, many people will probably hold that the theory, or at least any part of it which is involved in the prediction, is definitely established; for the prediction has the great weight which naturally attaches to one made before the facts were known. The other predictions, namely the change in the perihelion of Mercury and (if it is really a consequence of the theory) the Einstein shift, will probably remain doubtful for many years to come; but we should note that these predictions, if confirmed, will have as much evidential value as the deflection of light. For though the irregularity of the orbit of Mercury was known, it cannot be disputed that, if the theory predicts anything about the orbit of the planets, it predicts without ambiguity the value given on p. 98; in the numerical work there is involved nothing arbitrary which could possibly have been adjusted to suit another value, if experiment had suggested that value as more plausible. The difference between present and future knowledge is important only in so far as the former and not the latter may influence our reasoning.

The difficulty with the general theory is not so much whether the facts predicted by it are true, but whether they are really predicted and whether there may not be other facts, equally predicted, which are not true. So long as the theory cannot be reduced to a form which all can understand equally, there must always be a doubt whether its implications have been exhausted. That is, I believe, one of the main obstacles to complete confidence. But there is another sense in which it is doubtful whether the theory predicts the results which experiment confirms. Can we admit as 'prediction' a process of reasoning in which there are gaps filled only by guesses, or at least by the selection of one out of several possible alternatives that cannot be distinguished by rigid logic? Some exponents of the theory seem to maintain that the question does not arise, because the theory is a completely logical whole; but that is not, I think, the view of Einstein himself, nor is it—that is much less important—the view taken here. And if the question is relevant, there will undoubtedly be some who answer it in the negative; for a great deal of the criticism which the theory has received has been directed simply to exposing the gaps in the logical sequence and to insisting that the solutions of the problems which Einstein proposes are not unique; so distinguished a mathematician as Painlevé maintains that Einstein's solution is only one of an infinite number all equally 'natural'. How far this criticism is valid must be left to mathematicians to judge; it would be presumptuous to attempt any decision here. But it may be suggested that this particular criticism of the theory, unlike that which was attempted in Section IV, is less important to physicists than to mathematicians.

For no experimental science is, or can be, purely logical. If we refused to believe any proposition for which there was not logical proof, there would cease to be any science at all. In establishing laws and, still more, in establishing theories, we do habitually choose between alternatives for reasons which are perfectly incapable of formal statement. To say that we always choose the simplest proposition, even if it were true, would not reduce the arbitrariness of our choice; for nobody can explain exactly what we mean by simplest. The arbitrariness of the choice is not apparent to us, because it appears to us natural and obvious to make the choice that is actually made; but to others it is not always obvious. And it ceases to be obvious to us when it is made by those with intellectual instincts rather different from our own—by mathematicians, for example. Have we any reason to deny to their choice the validity which we so freely assume for our own?

If we do deny it, we shall do so on the ground that physics is the business of physicists and that the shoemaker should stick to his last. It is therefore relevant to point out that in the past history of physics the instinctive judgement of mathematicians has played an important part. So far, I fear, I have spoken of mathematics and of those who study it with less than due respect; but we cannot really divorce physics from the purest of pure mathematics. Consider only one example which concerns the starting point of modern physics, the Maxwellian theory of electrodynamics. It is often said that Maxwell merely put into mathematical form the physical ideas of Faraday; but a great deal depended upon which mathematical form he adopted. Faraday's theory doubtless suggested to him that the 'displacement' current should be introduced into the theory on a parity with the conduction current. But in choosing to represent that current by the term $\frac{dE}{dt}$, Maxwell was doubtless influenced very largely by the symmetry thereby produced in the equations; he chose between various alternatives on the ground of 'form', the distinctively mathematical criterion of value. And yet it was only if that particular term was introduced that the electromagnetic theory of light became possible. Another example of a purely mathematical choice leading to important physical results is given in Chap. XV, pp. 60—63. So far as I can see, the choice made by Maxwell or by Sommerfeld differs in no important feature from the choice by Einstein of the tensor G_{mn} or from any of the other 'guesses' to which attention was drawn in our discussion of the general theory. It is unreasonable to refuse confidence merely because such guesses are a necessary part of the argument; it would be reasonable only if there were alternatives in which the guesses were more acceptable.

34. **Are there alternatives?** In considering alternatives to the general theory we may exclude any which are based on the same fundamental idea of generalised space. Several modifications of the original theory have been proposed with the object of avoiding some of the difficulties of its more advanced parts, but for our purpose these may be regarded as part of the same theory. Attempts have also been made to deduce the conclusions from some of the 'supplements' to the special theory, which have been mentioned; but none of these are likely to be taken very seriously; they are all artificial and obviously directed to deduce a result already known. The most serious is probably that of Painlevé, who proposes to modify the axioms of dynamics rather than those of time and space; but it must be remembered that, apart from relativity, there is no reason to believe that light acts dynamically as a massive particle. On the other hand alternatives to the general theory are possible. For the general theory, unlike the special theory in its pre-Minkowski form, cannot be regarded merely as a unifying principle with which many theories may be consistent;

it is not simply a method of arriving at conclusions and without any 'real' significance. For the results deduced in it are much less important than the process of deducing them; the results which it explains are obscure; the process of deduction is immensely complicated; nobody would think it worth while to set up an elaborate and difficult principle in order to unify two, or perhaps three, minute 'corrections', appreciable, if at all, only to the most delicate observations, unless the principle had intrinsic interest. The importance of the general theory lies in the ideas which it introduces in order to arrive at its conclusions, not in the conclusions themselves; and any theory which arrived at the same conclusions starting from different ideas would be a true alternative. But, at least until very recently, there has been no serious alternative proposed. If the eclipse observations really establish the predicted theory of light, we may have to choose between the Einstein explanation and no explanation at all.

But is the Einstein explanation an explanation? This is really the vital question. When we have answered completely for any theory the two questions with which we started this section, the final judgement of truth has still to be made. A theory may predict much that is true and nothing that is false; there may be no alternative proposed which achieves as much; and yet it is perfectly legitimate to reject the theory without hesitation because it fails to give that intellectual satisfaction which is the only end and aim of science. It may be that a situation so extreme and clear cut has never actually occurred; rejection of a theory has almost always been based on some failure to predict and accompanied by the acceptance of some alternative. But rejection for mere failure to satisfy our intellectual demands is possible; indeed it happens every day when we accept laws for the time being as inexplicable, rejecting all the absurd and fantastic theories which would 'predict' but not explain them. It is, beyond question, the mere strangeness and difficulty of its ideas which prevent everyone from acclaiming the theory of relativity as a satisfactory solution of all the problems with which it professes to deal; in such a manner none can judge for another.

But perhaps two considerations may be urged, one in favour of relativity' and one against it. It is absurd to reject the theory merely because it is novel and strange. The branch of science with which all the chapters of this book deal is a new branch. Some 30 years ago it sprang into existence suddenly through the genius of a few great men, and it has developed with a rapidity for which it is hard to find a parallel. The new science started, as all sciences must start, from old principles, but from its very earliest days it has propounded problems to which those principles did not and could not give an answer. The new science demands new principles. Anyone who refuses acceptance to all that does not accord with the old science had better leave the new alone and follow some less progressive study. If we can grasp the ideas of relativity at all, we should be prejudiced in its favour if it asks us to rebuild the very foundations of our knowledge.

Now of these foundations perhaps the most fundamental is our conception of 'time'. It is so fundamental that it has been accepted hitherto practically without examination. And yet it is becoming every day more suspect. Almost all the great difficulties to which attention has been drawn in this and other chapters can be interpreted without strain as difficulties concerning this concept. A theory which suggests that our view of 'time' is false, just because it strikes at the roots of all science, should receive our very earnest consideration.

If we regard the matter from this point of view, if, instead of trying to reconcile these obscure phenomena with our existing ideas, we deliberately regard them as slight clues which may lead us to new realms of knowledge, we shall find nothing but good in the special theory. It tells us that we are wrong to consider that we can fix the time of an event apart from its position; 'time' and 'space' are not independent; they are—if we care to put it so—forms of the same thing. This discovery is a negative achievement; it clears the ground of pre-conceptions for which there was never any sound foundation. But the general theory attempts a positive achievement; it tries to replace the ideas which have been destroyed by others; and we must be more critical. The conclusion that 'time' and 'space' are but forms of the same thing may attract us, but the suggestion that this thing is again 'space' may well repel; it confines us once more within the circle of familiar ideas from which we would be free; it may repel not because it is revolutionary, but because it is not revolutionary enough.

But where shall we find anything more revolutionary? The answer is beginning to appear. Contemporaneous with relativity, and following a curiously parallel course, is the quantum theory of energy. It also calls on us to revise our ideas of time, indeed to revise them almost beyond recognition, and with the change in 'time' is necessarily associated some change of 'space'. Surely the two streams must join into a wider flood. As these concluding paragraphs are written, two fresh and striking suggestions (for there have been several hints) are made for such a junction. No discussion of them would be appropriate here; they are too recent and they belong properly to a later chapter.

INDEX

Aberration **26**, 30, 32–34, 45, 108
Absolute 5, 7, 34–35
Acceleration 68
Aether 5, 29, 32–34, 36
Ampère 41, 42
Axes 52–58, 62
Bohr-Sommerfeld theory 48, 108
Causation 12
Charge, electric 42, 43
Clock **13**, 18, 39, 40
Coincidence **8**, 39, 40
Coordinates 80–81, 93–95
Coulomb 40, 42
D'Alembert 48
Distance 57, 58, 62, 63, 81, 90, 91
  surface- 81, 82
Doppler effect **25**, 30–34, 45, 108
Einstein 51, 70, 102–105, 106, 107, 111–113
  'shift' 31, 103–105, 110
Electromagnetism 36, 40–49, 106, 111, 112
Electronic theory 45, 46
Energy 47–49
Equivalence, principle of 69–75
Euclid 66, 102
Events **8**, 39, 40, 41, 58, 62
Faraday 41, 42, 112
Fizeau experiment **26**, 29, 46, 108, 109
Forces 68, 69, 83–84
Fourier 66
'*g*'s' **79**, 83
Galileo 73, 74
Geodesic **92**
Gravitation 69, 70, 76, 88, 89, 94–96
Huyghens 29
Independence (of systems) **5**, 71
Invariance 61, 77–80, 85
Laws 6, 37–39, 44, 59, 90
  of motion 1, 68–69
Lenard 109
Length 10, 15–17, 23, 101, 102
Light, gravitational deflection of 99, 100
  velocity of 22, 28, 30, 38, 39, **44**
Lorentz-Fitzgerald contraction 109
Mass 47–49, 96–98
Maxwell 42, 44–46, 108, 109, 111, 112
Mercury, orbit of 98, 110
Michelson-Morley experiment **25**, **27–29**, 32, 36, 45, 108
Minkowski 51, 54, 64, 66, 67, 107, **112**
Newton 1, 34, 47, 48, 66, 68–70, **74**, 76
Numerical relations 59–63
  values 38
Observers 2, 3, 7
Ohm's Law 46
Orbits, elliptical 94–99
Painlevé 111, 112
Paschsky 109
Physical relations 59–63
Plotting 51, 60, 61
Poisson 42, 44, 45
Quantum theory 115
Real 33–35, 104
Relativity, first principle of **4–8**, 37–39, 68, 69
  fundamental ideas of 2–7
  geometrical interpretation of 51–67
  mathematical developments of 65, 66
  second principle of 4, **8–15**, 15–19, 22–24, 39–40, 75, 76, 100–103
  special and general theories 50, 107, 108
Roentgen 45, 108
Rowland 45, 108
de Sitter 106
Sommerfeld 112
Space, flat **83**, 92
  four-dimensional 63–66
  varieties of 80–84
Space-time **58**
Tensors **86–88**
Thermodynamics 1, 108–109
Time 11–13, 17–18, 102
  and space 56–59, 114
  imaginary 58, 64
Transformation (algebraic) 18, 23, 53, 61, 77–78
Trouton 36
Universal constants 38
Universe, size of 106
Velocities, compounding of 19–21, **24**
Vectors 66, 86
Weyl 106
Wilson (H. A.) 108
World-line **65**, 91–93
  of light 99, 100

CAMBRIDGE: PRINTED BY J. B. PEACE, M.A., AT THE UNIVERSITY PRESS