ON GENERAL RELATIVITY

By Banesh Hoffmann
Princeton University

TABLE OF CONTENTS

§1. Introduction

IN THE present article we give an account of certain of the fundamental ideas and assumptions underlying the general relativity theory and proceed to a discussion of various points that are not easily to be found in the literature of the subject; such problems, for example, as how the gravitational potentials are to be measured experimentally and what meaning, if any, can be attached to such phrases as *the velocity of light, rotation, relative radial motion*, and the like.

Questions of this nature are likely to be of interest to the general worker and we shall therefore aim at keeping the discussion as free from mathematical symbolism as possible. A more technical article is being prepared for this journal by H. P. Robertson; the present article is supplementary to such a discussion since completeness is not our concern; we shall, for instance, have practically nothing to say concerning field equations or cosmology and shall, moreover, confine ourselves strictly to the general relativity theory; this latter implies two major omissions, namely of a consideration of quantum phenomena and of electromagnetic phenomena, since the theory of these phenomena does not form an intrinsic part of the general relativity scheme which is essentially concerned with gravitational and inertial effects in their macroscopic aspect.

The classical, Maxwell-Lorentz, theory of electromagnetic phenomena has, of course, been brought into conformity with the general relativity scheme, but in such a way as to make no difference to the fundamental ideas of the relativity theory. Numerous attempts have been made with varying degrees of success to give a fundamental significance within the relativity theory to the electromagnetic field; these attempts constitute the majority of the so-called unified field theories and are not our concern in this article.

The omission of any consideration of quantum phenomena is more serious; it is not that we merely refrain from discussing the equations of the

quantum theory—since we shall anyway not discuss field equations—but that, in all that follows, we must definitely ignore the implications of the uncertainty principle. The general relativity theory is based upon the assumption that measurements can be refined to any desired degree of accuracy, the effect of the operation of measurement upon the system under observation being ignored. General relativity is in this sense classical and the appropriate extension to embrace the ideas of the quantum theory is not yet known. It remains nevertheless the best theory of gravitational phenomena and avoids many of the difficulties of the Newtonian theory.

§2. Space-Time; Coordinate Systems; General Invariance

In the relativity theory we start out with the assumption that every physical observation consists essentially of a succession of sets of four measurements; that is we assume that the world can be thought of as four-dimensional. So far we have said nothing startling; precisely this assumption lay at the bottom of the whole of classical field physics; it was pointed out by Laplace, for example, that particle dynamics is geometry of four dimensions in which time is the fourth dimension.

Suppose we wish to specify the motion of an ideal particle; we must first decide upon some technique of measurement that will provide us with sets of four numbers representing the position of the particle at each instant. A position at an instant is called an event and if we specify all the events that constitute the motion, that is, if we specify the position of the particle at each instant, we shall have specified the motion; it will be represented by a set of relations between the sets of four numbers representing the events constituting the motion. As an example, if a typical set of four numbers be denoted by (x, y, z, t) we might find that the motion obeys the relations:

$$\begin{cases} x = 0 \\ y = 0 \\ z = t. \end{cases}$$

We might be tempted to say that this particular motion is in a straight line with unit velocity, but it is evident that no such conclusion can be drawn from the above relations since we have made no statements regarding the actual technique used in obtaining the numbers, (x, y, z, t), that represent an event. If we assert that the above motion is rectilinear with uniform velocity we are making several assumptions; we are assuming, for example, that the locus, $x=0$, $y=0$, is a straight line in space and that z, or t, represents, say, astronomical time. This may or may not make sense but at the moment we are certainly not far enough along in our argument to attempt to decide. Instead let us start out with as few prejudices as possible from the assumption of a four-dimensional world in order to see to what a view of space and time we are led by a consideration of some simple ideal experiments.

First of all, however, we wish to obtain a clearer idea of the methods by which sets of four numbers are to be assigned to events. The direct way is to

erect a scaffolding and to place a clock at each intersection; an arbitrary numbering may be given to the beams in the scaffolding so that each intersection has three numbers attached to it. The four numbers to be assigned to a given event are then the three numbers belonging to the nearest intersection and the reading of the clock *at that intersection* at the instant of the event.[1] In practice of course so ideal a technique would never be used. We have considered it here in order to bring out the point that there is a separate clock for each intersection of the scaffolding. We have not mentioned the question of synchronization of these clocks and furthermore it is usual in actual practice to employ only one, or at most only a few clocks. These two points are intimately related; for, let us consider a more practical technique from the point of view of our ideal scaffolding with its clocks at each intersection. To avoid the actual erection of a scaffolding a method of triangulation is employed and generally only one clock is used. The values of the four numbers that specify an event are obtained by some mode of computation from the readings of the surveying instruments and the clock, and in particular it would be quite in order from the relativisitic point of view merely to take these readings themselves as the four numbers in question. If this be done it is evident that the clocks of our hypothetical scaffolding are assumed to be synchronized by light signals from[2] the master clock actually used in making the observation. It is more customary, however, to make some correction for the time that is required for light to reach the master clock from the event under observation. In this case an assumption must be made concerning the manner in which light is propagated when expressed in terms of the scheme of labelling events that we have been using, and this, since it will be represented by a series of relationships between the readings of the clocks at the various intersections of our scaffolding, is nothing more than another system of synchronization—the system actually employed in our hypothetical scaffolding in the present case. The way in which light propagation is expressed in terms of a given hypothetical scaffolding can only be determined by experiment with this scaffolding; and the way in which light propagation is expressed is as much a property of the scaffolding as it is of light; and indeed, in general the result of any experiment contains information not only about the object upon which the experiment was supposedly performed but also about the hypothetical scaffolding and set of clocks in terms of which the measurements were expressed.

The main function of one of our scaffoldings with its clocks at every intersection is to provide a unique label to every event under consideration; there are, however, certain properties that we require of any scaffolding system—properties that are always possessed by the equivalent scaffolding system of any scheme of measurement used by the experimenter. We require, as we have already mentioned, that our given scaffolding system and its clocks provide a label of four numbers for every event under consideration;

[1] We have made no requirements as to rigidity or immobility of our scaffolding; the correct picture is of a scaffolding, as it were, idly floating and flapping in the breeze!

[2] Assuming that the forward and backward velocities of light are equal.

two distinct events are to acquire distinct labels, and any single event not more than one label. Again, we have stated that the numbering of the beams in our scaffolding may be arbitrary; in order to avoid mathematical complications, such, for example, as the appearance of functions that are not continuous or that do not possess derivatives, we require to modify this statement. The simplest way in which to express the modification is just to say that we shall deal only with scaffolding and clock systems in terms of which all functions that we shall encounter are as far as possible analytic. From the mathematical point of view this is of course a very serious restriction, but physically it amounts to nothing more than an expression of intuitive limitations ordinarily imposed by the experimenter. The beams of a scaffolding form three distinct families; the above requirements prevent such happenings as, for example, the intersecting of two beams belonging to the same family or the sudden reversal of the direction of motion of the hands of any of our clocks.

An event, then, is specified, in terms of a scaffolding system with clocks at the intersections, by means of four numbers. With the customary terminology we refer to these four numbers as the *coordinates* of the event. The system of scaffolding and clocks in terms of which sets of four numbers are assigned to events is thus equivalent to a system of coordinate lines in four dimensions; but the system of scaffolding and clocks is itself no more than a convenient way of talking about a complicated technique of measurement and computation whose net result is to assign coordinate numbers to events; it is therefore to this complicated technique that we shall always refer primarily when we speak of a coordinate system—although we shall often find it convenient to regard it as a system of scaffolding and clocks and also, especially in later sections, as equivalent to a four-dimensional network of coordinate lines.

It should be noted that space and time are interrelated according to the present scheme, since we deal always with the time indicated by that clock of our scaffolding that is situated nearest to the event being observed; later on we may or we may not find reason to separate space and time from each other but at present we have no excuse for so doing, and in this respect we depart from the Newtonian theory in which an assumption as to the separability of space and time is introduced at this point.

Now the mode of measuring and computing coordinates for events is not unique; we may change it at will; for example, we could multiply every number obtained by two, or could perform much more complicated feats with them; for us to choose some definite scheme of assigning coordinates in preference to all others, or some set of such schemes in preference to the rest, would imply the existence of an objective standard; if we attempted to define such a standard at the present juncture we should be arguing in a circle since we have not yet made any use of objective experiments which alone could determine the matter for us. So at the present stage we must treat all modes of assigning coordinates to events—that is, all coordinate systems—within the limitations imposed upon them, as on an equal footing. But it is evident that

an alteration in the mode of describing objective phenomena makes no difference to these phenomena; it follows that the phenomena of physics although only expressible, by present methods, in terms of coordinate systems are yet to be completely independent of them. We must therefore require that the mathematical expression of physical phenomena be of such a nature that, although it necessarily makes use of a coordinate system, what is expressed is not dependent upon the coordinate system that happens to have been used. This is the general principle of relativity and holds in exactly the same words for the Newtonian theory; the difference between the two theories lies in what is expressed and not in the mode of expression, as we shall see in the course of the argument. Meanwhile we have seen that all physical phenomena are to be expressible in terms that do not require a specification of a definite coordinate system; such a requirement is not a difficult one to fulfill; the ordinary vector calculus, for example, performs precisely this function. In general relativity an extension of the vector calculus is employed which makes use of entities called *tensors* which have components in a given coordinate system and whose components change in a definite manner as the coordinate system is changed. It is not our intention to develop here the general invariantive theory of Ricci and his followers, and in what follows the question of invariance will be treated dogmatically without any attempt at proof.

We have discussed the mechanism of observation; our next step is to see what can be learned from an application of our technique to certain simple phenomena.

§3. Discussion of an Experiment

The most suitable phenomenon for our consideration happens to be the explosion of a bomb; we are particularly interested in two aspects of the explosion; several particles will be projected from (approximately) the same point at (approximately) the same instant, in various directions and with various speeds and furthermore a pulse of light will spread out as a wave-front from the place at which the explosion occurred. We shall idealise the experiment in the usual manner by looking on the particles as mass-points, by leaving out the "approximately's" and so on.

Let us attempt to picture these two aspects of the explosion in terms of the four-dimensional world of our observational technique; that is to say, let us regard our technique of labelling events as providing a four-dimensional coordinate network into which all the events that constitute that part of the explosion in which our interest lies may be imbedded. This four-dimensional world represents the actual world of physics insofar as physics deals with observations of the type we are considering in the present article.

In order to be able to draw adequate pictures of what is happening in the four-dimensional world we must neglect one of its dimensions and since we cannot spare the time it is a space-like dimension that must be omitted. We are now in a position to draw a picture of the explosion and the result is shown in Fig. 1. The line, AB, denotes the bomb before the explosion; the

conical surface, BCD, (remembering that we are really dealing with a conical surface of three dimensions in a four-dimensional space) represents the history, or at least the initial stages of the history, of the pulse of light sent out from the position of the bomb at the time of the explosion, i.e., sent out from the event, B. We know that this is a conical locus since we know light has a definite speed in any given direction in terms of a given coordinate system. The pencil of lines, BEF, denotes the histories of the various particles into which the bomb was broken by the explosion; the speeds of the particles are not determined when the direction is given as was the case with light and therefore the trajectories representing the histories of the particles do not lie on a (three-dimensional) conical surface through the event B but occupy a four-dimensional region. The velocities of these particles are usually small

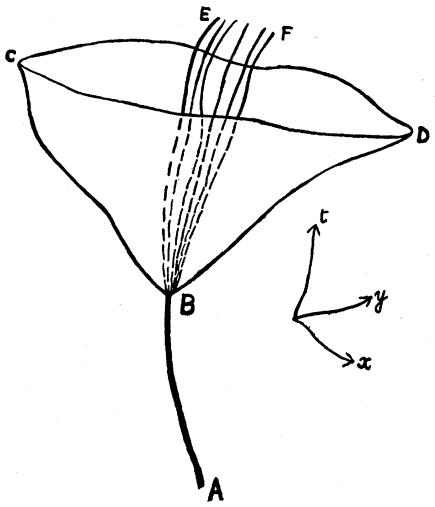

Fig. 1.

compared with that of light and this is represented by drawing BEF as a pencil of small angle compared with that of the cone. We have not drawn any of the lines to look straight since so far no meaning has been attached to the term "straight" and even if some intuitive meaning were ascribed to it our drawing would not necessarily have to be without distortion so long as the coordinate network were compensatingly distorted, and this has not been drawn in the figure; the situation is completely analogous to that which arises when a country is mapped according to an unspecified projection.

There are two immediate remarks to be made concerning the explosion; the first is the fact that light rays propagated from a definite point at a definite instant in a definite direction always move so that their tips coincide; in terms of the four-dimensional picture this means that the light-cone, BCD, determines only two velocities for any spatial direction through the position of B—the velocities in the positive and negative senses of the given spatial direction. Hence the cone is one that has ovaloidal spatial cross-sections.

The second point is due to Galileo and is to the effect that once the direction and speed of projection from a given point at a given time are known the complete motion of a particle is determined irrespective of the mass of the particle performing the motion.[3] In the Newtonian theory this fact is expressed by saying that gravitational and inertial mass are proportional; let us see what it means in terms of our four-dimensional world. If we know the coordinates of the event B and of an event, B', that lies near B on a given trajectory we have a knowledge of the initial direction and initial speed of the projection of the particle under consideration in terms of the coordinate system used and we also have a knowledge of the initial position and initial instant of the projection. Thus by Galileo's principle we have sufficient data to specify the whole motion completely. It follows, therefore, that all four-dimensional trajectories of material particles (with the reservation of the last foot-note) are curves that are determined when two events near to each other belonging to the trajectory are given.

It is to be noted that we are not using the term "trajectory of a particle" to denote the spatial locus of its motion — but to denote the four-dimensional curve that gives a more complete characterization of the motion by including a record of the instant at which each spatial position was occupied by the particle.

§4. Equations of the Light-Cone and of Trajectories; Existence of g and Γ Fields

Our next step is to consider the mathematical significance of the two points brought out in the previous section.

The first shows that at every point of the four-dimensional world there is determined a unique conical surface of three dimensions having ovaloidal spatial cross-sections; the simplest surface of this nature is a quadratic cone of three dimensions, and in the relativity theory it is assumed that it is a quadratic cone that is determined at each point of space-time by the behaviour of light.

Let the coordinates of an event be denoted by (x^1, x^2, x^3, x^4), or by (x^α), where α takes on the values 1, 2, 3 and 4. Let $(x^\alpha+dx^\alpha)$ by the coordinates of an event near (x^α). Then the general equation of a quadratic cone through (x^α) is

$$\sum_{\alpha,\beta}^{1,2,3,4} g_{\alpha\beta}dx^\alpha dx^\beta = 0,$$

where $g_{\alpha\beta}$ are sixteen functions of the coordinates, x^α. We shall use the summation convention introduced by Einstein whereby the summation sign is

[3] In general this is not true; to be more accurate we should state that it is necessary to know in addition to the above data the value of the initial component of acceleration in some direction. If we find that the trajectory is independent of this additional quantity, as assumed in the text, we characterise the situation by saying that there is no electromagnetic field present but only a gravitational field, or else that the particles carry no charge; we are omitting any consideration of electromagnetic phenomena and are therefore justified in leaving the statement in the text as it stands.

omitted whenever summation over a repeated suffix is intended. Our cone thus becomes

$$g_{\alpha\beta}dx^{\alpha}dx^{\beta} = 0 \tag{4.1}$$

or, if we prefer to have a differential equation,

$$g_{\alpha\beta}\frac{dx^{\alpha}}{ds}\frac{dx^{\beta}}{ds} = 0, \tag{4.2}$$

where s is some parameter, as yet not specified.

We have seen that all physical phenomena are to be expressible in terms that do not depend upon the particular coordinate system being used. It can be shown that this requires the sixteen functions $g_{\alpha\beta}(x)$ to be the components of a tensor. Further, it is only the symmetric part, $\frac{1}{2}(g_{\alpha\beta}+g_{\beta\alpha})$, of $g_{\alpha\beta}$ that enters into (4.1) so that we may look upon $g_{\alpha\beta}$ as symmetrical without loss of generality; in this case, since $g_{\alpha\beta}=g_{\beta\alpha}$ by hypothesis, there are only ten independent functions, $g_{\alpha\beta}$.

Now (4.1), or (4.2), is unaltered if we multiply throughout by any quantity whatsoever, other than zero; thus we can only infer that a tensor $Xg_{\alpha\beta}$ is determined at each point of space-time, where X is completely arbitrary and may involve the quantities dx^{α}, the distance between New York City and Berlin, or any other quantity. Nevertheless in the relativity theory it is assumed, as being the straightforward thing to do, that X involves only the x^{α} and may therefore be absorbed into the $g_{\alpha\beta}$ which are themselves functions only of the x^{α}.

So the first point has as a consequence the assumption that the four-dimensional world is the seat of a symmetric tensor field, $g_{\alpha\beta}$; we shall see later how this is related to the theory of gravitational and inertial effects.

Let us proceed to a consideration of the mathematical expression of the second point of the previous section, namely the principle of Galileo. It can be shown that the differential equations to a trajectory fulfilling the conditions required by the principle of Galileo must be of the form

$$\frac{d^2x^{\alpha}}{ds^2} + \Gamma^{\alpha}_{\beta\gamma}\frac{dx^{\beta}}{ds}\frac{dx^{\gamma}}{ds} = 0, \tag{4.3}$$

where the quantities $\Gamma^{\alpha}_{\beta\gamma}$ satisfy $\Gamma^{\alpha}_{\beta\gamma}=\Gamma^{\alpha}_{\gamma\beta}$, are functions of the x^{α} and of the dx^{α}/ds, and are homogeneous of degree zero[4] in the dx^{α}/ds. Furthermore, in order that the equations (4.3) have a significance that is independent of the particular coordinate system used the quantities $\Gamma^{\alpha}_{\beta\gamma}$, of which forty are independent of each other, must transform, under a transformation of coordinates, according to a certain law; we characterise this by saying that the Γ's transform like the components of an affine connection. Thus Galileo's principle leads to the assumption that there exists in space-time a field of Γ's that transform like the components of an affine connection.

[4] This comes from the additional assumption that the trajectory is unaltered if we substitute $as+b$ for s where a and b are constants.

Space-time has thus been endowed with two definite characteristics as a result of our study of the explosion of a bomb.

Before we proceed to a discussion of the assumption that is made in the relativity theory regarding the relationship between the g and Γ fields and of the method by which the g field can be determined experimentally we have one important remark to make concerning the g's regarded as coefficients in the equation of the light cone; for the light cone is a real locus and this imposes a restriction upon the g's. The left-hand side of (4.2) is a quadratic form in the four variables dx^α/ds; by a well-known theorem any quadratic form involving real variables can be reduced by a real transformation to the canonical form in which it becomes the sum or difference of the squares of the new variables, and moreover the number of positive and the number of negative coefficients is independent of the mode of reduction. We may thus consider (4.2) in its canonical form to be

$$\pm\left(\frac{dx^1}{ds}\right)^2 \pm \left(\frac{dx^2}{ds}\right)^2 \pm \left(\frac{dx^3}{ds}\right)^2 \pm \left(\frac{dx^4}{ds}\right)^2 = 0,$$

where a definite number of positive and a definite number of negative signs are to be taken. Now we know from experience that three dimensions of the four-dimensional world are of the same nature and hence must enter into our equations in the same way; thus three of the signs must be the same. If the fourth sign were the same as the other three the locus would be imaginary. Hence we must take the sign of the fourth square to be opposite to that of the other three, and this is essentially the restriction that the g's must obey. It is possible to state this restriction in general terms but nothing is gained by so doing since the conditions are not expressible in a simple manner; it is sufficient for our purposes to state that the g's must be such that under a suitable real transformation of coordinates the equation of the light cone can be reduced to the form

$$-\left(\frac{dx^1}{ds}\right)^2 - \left(\frac{dx^2}{ds}\right)^2 - \left(\frac{dx^3}{ds}\right)^2 + \left(\frac{dx^4}{ds}\right)^2 = 0. \tag{4.4}$$

This is a restriction upon the behaviour of the g's in the neighbourhood of a given event, and nothing is implied as to the possibility of such a reduction being possible over a large region of the four-dimensional world.

§5. Experimental Determination of the g and Γ Fields

Let us now consider the problem of determining the g's and the Γ's experimentally. In the relativity theory a further assumption is made that leads to an expression for the Γ's in terms of the g's; at the moment we shall not make use of this assumption but, for the sake of simplicity, we shall tentatively assume that the Γ's are actually independent of the dx^α/ds — this being the simplest way of fulfilling the condition that they be homogeneous of degree zero in the dx^α/ds and being in keeping with the assumption we shall introduce later.

The Γ's and the g's are functions of the coordinates in two senses, they involve the x^α in their mathematical expression and they also undergo somewhat complicated intermixing when the coordinate system is changed so that their functional form becomes altered under a transformation of coordinates; for example, in one coordinate system we might find that g_{11} is represented by (x^1+2x^3) and in another coordinate system by $x'^2\cos x'^3$. Our experiments are designed to provide a knowledge of the functional form of the g's and Γ's so that it is necessary to keep to one definite coordinate system throughout the experiment and the subsequent computations; the result will then be the functional forms of the g's and Γ's *in terms of the coordinate system used.*

We consider the determination of the Γ's first; the Eqs. (4.3) may be rewritten in the following manner; let x^4 represent the time-like coordinate and let us replace it for convenience by t; let us furthermore use Latin suffixes to to take on only the values 1, 2 and 3. then we have

$$\frac{dx^\alpha}{ds} = \frac{dx^\alpha}{dt}\frac{dt}{ds}$$

and

$$\frac{d^2x^\alpha}{ds^2} = \frac{d^2x^\alpha}{dt^2}\left(\frac{dt}{ds}\right)^2 + \frac{dx^\alpha}{dt}\frac{d^2t}{ds^2};$$

the Eqs. (4.3) are therefore the same as

$$\frac{d^2x^\alpha}{dt^2} + \Gamma^\alpha_{\beta\gamma}\frac{dx^\beta}{dt}\frac{dx^\gamma}{dt} + \left\{\frac{d^2t}{ds^2}\Big/\frac{dt}{ds}\right\}\frac{dx^\alpha}{dt} = 0. \tag{5.1}$$

But we have written t for x^4 so that, for $\alpha=4$, we have

$$0 + \Gamma^4_{\beta\gamma}\frac{dx^\beta}{dt}\frac{dx^\gamma}{dt} + \left\{\frac{d^2t}{ds^2}\Big/\frac{dt}{ds}\right\} = 0,$$

and, substituting in (5.1) for $\alpha=1, 2, 3$ and making use of our convention regarding Latin suffixes, we easily obtain the three equations

$$\begin{aligned}\frac{d^2x^a}{dt^2} + \Gamma^a_{44} - \frac{dx^a}{dt}\{\Gamma^4_{44} - 2\Gamma^a_{a4}\} - \left(\frac{dx^a}{dt}\right)^2\{2\Gamma^4_{a4} - \Gamma^a_{aa}\} \\ - \left(\frac{dx^a}{dt}\right)^3\Gamma^4_{aa} + 2\Gamma^a_{b'4}\frac{dx^{b'}}{dt} - 2\Gamma^4_{b'4}\frac{dx^a}{dt}\frac{dx^{b'}}{dt} \\ + \Gamma^a_{b'c'}\frac{dx^{b'}}{dt}\frac{dx^{c'}}{dt} - \Gamma^4_{b'c'}\frac{dx^a}{dt}\frac{dx^{b'}}{dt}\frac{dx^{c'}}{dt} = 0,\end{aligned} \tag{5.2}$$

where the primes denote that the suffix does not take on the value that a has, so that a primed Latin suffix takes on only two of the three values 1, 2, 3 depending on which particular value a happens to have; and furthermore the summation convention is suspended for unprimed indices that occur in (5.2). Let us now return to our explosion; this furnishes several trajectories emanating from a given event in various initial space-time directions. The act of ob-

servation provides us with a table of values of the coordinates, x^α, of events lying on these trajectories; from these tables we can compute, for a given trajectory, the initial values of the quantities dx^a/dt, d^2x^a/dt^2, d^3x^a/dt^3, $\cdots$, and so on, and thus dx^a/dt and d^2x^a/dt^2 can be expressed as Taylor series about the event at which the explosion occurred; if these values of dx^a/dt and d^2x^a/dt^2 belonging to a given trajectory be substituted in (5.2) we obtain three equations involving the Γ's as unknowns. Each trajectory provides three such equations and by choosing a sufficiently violent explosion we can obtain as many different trajectories as we please emanating from the same event. The number of independent terms in (5.2) is thirty-six. Thus by observing twelve trajectories, substituting for the various dx^a/dt and d^2x^a/dt^2 in (5.2), and solving the resulting algebraic equations we can obtain values of the coefficients,

$$\Gamma^a_{44},\ \{\Gamma^4_{44} - 2\Gamma^a_{a4}\},\ \{2\Gamma^4_{a4} - \Gamma^a_{aa}\},\ \Gamma^4_{aa},$$
$$\Gamma^a_{b'4},\ \Gamma^4_{b'4},\ \Gamma^a_{b'c'} \text{ and } \Gamma^4_{b'c'},$$

as functions of the x^α; that is to say, we can obtain functional expressions for the quantities

$$\left.\begin{matrix}\Gamma^4_{44} - 2\Gamma^a_{a4}\\ \Gamma^a_{aa} - 2\Gamma^4_{a4}\end{matrix}\right\}, \tag{5.3}$$

(where the summation convention is, of course, still suspended), and for every Γ not entering the above expressions.

Now, starting once more from Eq. (4.3), we can change the independent variable to, say, x^3 instead of to $t=x^4$; by using the readings obtained from our previous explosion[5] we can now compute the dx^1/dx^3, dx^2/dx^3 and dx^4/dx^3 as functions of the x^α. Proceeding as before we shall now be able to obtain expressions for

$$\left.\begin{matrix}\Gamma^3_{3} - 2\Gamma^e_{e}\\ \Gamma^e_{ee} - 2\Gamma^4_{e}\end{matrix}\right\}$$

and every Γ not entering the above, where e takes on the values 1, 2 and 4, and the summation convention is suspended. But now, for example we have an expression for

$$\Gamma^3_{33} - 2\Gamma^1_{13}$$

as a function of the x^α, and we have already obtained a value for Γ^1_{13} from our previous work; we thus have information enough to determine Γ^3_{33}, and in a similar manner Γ^1_{11}, Γ^2_{22} and Γ^4_{44} may be found, so that, knowing (5.3), we are able to determine every $\Gamma^\alpha_{\beta\gamma}$ as a function of the x^α.

Let us now turn our attention to the problem of determining the $g_{\alpha\beta}$ as functions of the x^α; we make use of our explosion once more, this time to observe the behaviour of the light pulse that is emitted. The Eq. (4.2) which expresses the way in which light is propagated may be written as

[5] It is important that we keep to one explosion all the while, since a second explosion would give information of the state of the four-dimensional world at a different four-dimensional point and would therefore be of no use to us at the moment.

$$g_{\alpha\beta}\frac{dx^\alpha}{dt}\frac{dx^\beta}{dt} = 0,$$

where t is written for x^4 as before; this is the same as

$$g_{ab}\frac{dx^a}{dt}\frac{dx^b}{dt} + 2g_{a4}\frac{dx^a}{dt} + g_{44} = 0, \tag{5.4}$$

where we are once more using Latin suffixes to take on the values 1, 2 and 3, but the summation convention is no longer suspended. To return now to our experiment, we can obtain a table of coordinates of events on the light-cone through the event of the explosion and from this we can compute for a given initial direction the values of the initial velocity, initial acceleration, initial rate of change of acceleration, $\cdots$, and so on of light along a ray in that direction in terms of the coordinate system used! Postulating for the moment that an experiment of such accuracy could be devised we thus have the initial values of dx^a/dt, d^2x^a/dt^2, d^3x^a/dt^3, $\cdots$, and so on, and can therefore obtain the quantities dx^a/dt for a given initial direction as Taylor series about the event of the explosion. Substituting these values in (5.4) and repeating the process for the tips of eight other rays we obtain nine algebraic equations from which the *ratios* of the ten $g_{\alpha\beta}$ may be determined as functions of the x's. We cannot discover more than this from (5.4) since only the ratios of the g's enter into it.

However, it is quite impossible for us to measure even the acceleration of light in a given direction and so the determination of the g's from experiments on the propagation of light in conjunction with (5.4) is extremely inaccurate. The method however furnishes an accurate means of determining the *initial* values of the ratios of the g's at the event of the explosion and we shall see that this will be of considerable use for the more accurate determination of the g's that we shall now discuss which is based upon the relationship that is assumed to exist between the g and the Γ fields.

§6. Assumption Relating the Γ and g Fields; More Accurate Experimental Determination of the g Field

The consideration of a bomb explosion has led us to postulate that space-time is the seat of a tensor field, $g_{\alpha\beta}$, and of a field of Γ's which transform like an affine connection. We have seen also that by a suitable choice of coordinate system we can cause the light-cone to take the form (4.4), that is, the tensor $g_{\alpha\beta}$ to take the values

$$\begin{vmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix}$$

in the neighborhood of a given event. The form (4.4) is reminiscent of the special relativity theory and we are led to postulate that the quantity,

$$-(dx^1)^2-(dx^2)^2-(dx^3)^2+(dx^4)^2,$$

that enters (4.4) has a metrical significance similar to that which it has in the special relativity theory; that is to say we regard the g field as possessing a metrical significance and generalise the special relativity concept of the proper time or proper length of a vector, V^α, to be the scalar $(g_{\alpha\beta}V^\alpha V^\beta)^{1/2}$. Thus for two neighbouring points, (x^α) and $(x^\alpha+dx^\alpha)$, we have the invariant

$$(ds)^2 = g_{\alpha\beta}dx^\alpha dx^\beta \tag{6.1}$$

which represents the square of the interval between them. Light is therefore propagated according to the law

$$(ds)^2 = 0.$$

Furthermore, a ray of light, or more accurately the motion of the tip of a ray of light, is completely characterized when an initial event and initial four-dimensional direction[6] are given, and this is precisely the situation that arises in the case of material trajectories in a gravitational field. We must therefore conclude that there is a Γ field determined by the behaviour of light in addition to the Γ field determined by the behaviour of material particles; it is tempting to assume that these two Γ fields are actually identical and if we make this assumption we are at once led to look for a relationship between the Γ's and the g's since both are now theoretically determined by the behaviour of light. Light propagation is characterised by having $ds=0$; the question is, can we obtain a means of expressing the Γ's in terms of the g's in such a way that the only difference between the characterizations of the motion of a material particle and of the motion of the tip of a light ray shall be that in the latter case $ds=0$? The answer is that we can attain this object if we assume that each of these types of trajectory satisfies the condition that it is that curve for which between any two events on it the integral

$$\int\left(g_{\alpha\beta}\frac{dx^\alpha}{dt}\frac{dx^\beta}{dt}\right)^{1/2}dt$$

taken along it is stationary. This means that we assume that the trajectories of moving particles and the trajectories of the tips of light rays are geodesics with regard to the metrical field, $g_{\alpha\beta}$. The mathematical consequence of this is that $\Gamma^\alpha_{\beta\gamma}$ is now replaced by $\left\{{\alpha \atop \beta\gamma}\right\}$, where

$$\left\{{\alpha \atop \beta\gamma}\right\} = \frac{1}{2}g^{\alpha\sigma}\left(\frac{\partial g_{\gamma\sigma}}{\partial x^\beta}+\frac{\partial g_{\beta\sigma}}{\partial x^\gamma}-\frac{\partial g_{\beta\gamma}}{\partial x^\sigma}\right), \tag{6.2}$$

$g^{\alpha\sigma}$ being the normalized cofactor of $g_{\alpha\sigma}$ in the determinant $|\,g_{\alpha\beta}\,|$.

The assumption that

$$\Gamma^\alpha_{\beta\gamma} = \left\{{\alpha \atop \beta\gamma}\right\} \tag{6.3}$$

[6] For light a three-dimensional direction automatically determines its four-dimensional direction but this does not spoil the argument.

cannot be satisfied by arbitrary Γ's, as is evident since the forty independent Γ's are here expressed in terms of only ten functions, $g_{\alpha\beta}$; the above identification implies that the Γ's are independent of the dx^α/ds, as we have already tentatively assumed, and that they satisfy certain integrability conditions. Assuming that the values of the Γ's found from observations upon an explosion are capable of satisfying (6.3), we have from (6.3) and (6.2),

$$\frac{\partial g_{\alpha\beta}}{\partial x^\gamma} = \Gamma^\epsilon_{\beta\gamma} g_{\alpha\epsilon} + \Gamma^\epsilon_{\alpha\gamma} g_{\epsilon\beta}$$

or, multiplying by dx^γ and employing the summation convention,

$$dg_{\alpha\beta} = (\Gamma^\epsilon_{\beta\gamma} g_{\alpha\epsilon} + \Gamma^\epsilon_{\alpha\gamma} g_{\epsilon\beta}) dx^\gamma$$

which must be integrable according to our assumptions. Now this equation expresses the increment of $g_{\alpha\beta}$ that accompanies an increment dx^γ of the x^γ in terms of the values of the g's at the point (x^γ), the Γ's being known from experiment. It is thus evident that, subject to certain conditions of convergence, single-valuedness, etc., which we have assumed to be satisfied, we can obtain a Taylor expansion solution for the g's which will involve as arbitrary constants their ten initial values at the event of the explosion. But these initial values are, except for an arbitrary multiplying constant, precisely what can be accurately found from observations upon light propagation. Thus in general relativity it is possible to measure experimentally the values of the $g_{\alpha\beta}$ to within a single arbitrary constant; that is to say, $\gamma g_{\alpha\beta}$ can be determined where γ is an arbitrary constant depending on the scale used in our measurements.

§7. Statement of the Relativistic View-point

In previous sections we have been led to postulate that the four-dimensional world is the seat of a tensor field, $g_{\alpha\beta}$, and have discussed the way in which this field may be experimentally determined. The general relativity theory aims at the use of the g field, and of the g field only, for an explanation of all macroscopic gravitational and inertial phenomena; the detailed discussion of field equations and the like does not belong to the present article as we have already pointed out, but certain more general matters are of great importance. Our problem may be stated in the following way; for various reasons it has been found necessary or convenient to introduce certain terms such as "inertial frame", "acceleration", "rotation", "Doppler effect", and so on; the motive for the introduction of these terms must have been the existence of certain corresponding physical effects that have been noticed; very often, however, our choice of a word and the implications we have come to attribute to it have been greatly influenced by the Newtonian philosophy—the term "simultaneity" is an excellent case in point. Although the concepts denoted by the terms we ordinarily use may prove to have no unambiguous meaning from the point of view of general relativity, nevertheless every such term must actually refer to some physical phenomenon or assumption since

otherwise it would not have been introduced; it is our purpose to attempt to recognise the phenomenon that is referred to when a given term is employed, to examine whether the usual meaning attached to this term can be justified from the new standpoint, and, if it cannot, to discover the true extent of its significance from the point of view of the general relativity theory.

The equations of motion of a material particle are

$$\frac{d^2x^\alpha}{ds^2} + \left\{ \begin{matrix} \alpha \\ \beta\gamma \end{matrix} \right\} \frac{dx^\beta}{ds} \frac{dx^\gamma}{ds} = 0. \tag{7.1}$$

These express a fact that is independent of the particular coordinate system used, namely that the trajectories of moving particles are assumed to be geodesics in a space-time whose metrical properties are governed by the $g_{\alpha\beta}$. If the coordinate system be changed the functional forms of the $g_{\alpha\beta}$ are changed and therefore the explicit form of the above equation is altered; hence the experiments designed to measure the g's are actually also experiments for the determination of the idiosyncracies of the coordinate system used; coordinate systems in terms of which the Eqs. (7.1) take on particularly simple forms are evidently in some sense privileged systems and if the theory is to stand we must be able to point to the way in which their recognition as such is related to our experience.

The simplest form that (7.1) might take is undoubtedly

$$\frac{d^2x^\alpha}{ds^2} = 0, \tag{7.2}$$

and this will hold for all trajectories only if a coordinate system can be found in terms of which the $g_{\alpha\beta}$ become constants. In general it is not possible to reduce arbitrary $g_{\alpha\beta}$ to constants by a suitable choice of coordinate system and the condition that this be possible is that the so-called curvature tensor, or more accurately the Riemann-Christoffel tensor,

$$R^\alpha_{\beta\gamma\delta} \equiv \frac{\partial \left\{ \begin{matrix} \alpha \\ \beta\gamma \end{matrix} \right\}}{\partial x^\delta} - \frac{\partial \left\{ \begin{matrix} \alpha \\ \beta\delta \end{matrix} \right\}}{\partial x^\gamma} + \left\{ \begin{matrix} \epsilon \\ \beta\gamma \end{matrix} \right\} \left\{ \begin{matrix} \alpha \\ \epsilon\delta \end{matrix} \right\} - \left\{ \begin{matrix} \epsilon \\ \beta\delta \end{matrix} \right\} \left\{ \begin{matrix} \alpha \\ \epsilon\gamma \end{matrix} \right\},$$

vanish. This gives us a means of determining whether the complexity of the equations of motion is entirely due to an unwise choice of coordinate system or whether it is impossible for us to effect a complete simplification to the form (7.2).

Now for all motion to be expressible in the form (7.2) is the nearest approach in the relativity theory to the state in which motion is in a straight line with constant speed, and therefore, using the Newtonian terminology, under no forces; it actually corresponds to special relativity. By changing our coordinate system we change the g's and therefore, in general, the form of the equations of motion; this will be interpreted as the introduction of a fictitious force, as, for example, a centrifugal force; but if the curvature tensor do not vanish we shall be unable to express all motions in the form (7.1) and shall

therefore regard the situation as corresponding to the existence of a field of force which cannot have been entirely due to our choice of coordinates; we interpret it as the gravitational influence of other bodies. Now this influence is not arbitrary but is spread out in a definite manner, as we know from observation as approximately embodied in Newton's inverse square law; it follows that the g's must satisfy certain extra requirements, and we thus realise the necessity for field equations to express the relativistic analogue of the Newtonian law. We shall not pursue this matter further; we wish only to point out that the sole difference between gravitational and inertial effects from the point of view of general relativity is that the complexity of the latter can be regarded as entirely our own fault whilst the former remain always to a certain extent outside the influence of our coordinate systems.

In general, then, on account of the curvature of space-time, there does not exist a coordinate system in terms of which the $g_{\alpha\beta}$ become constants.

§8. Inertial Frames; Acceleration and Rotation of Coordinate Systems

Let us consider first of all what we mean by the term "inertial system". The Newtonian laws of motion hold in an inertial system and this is the only definition of the term that can be given from the classical standpoint without a circular argument. Insomuch as the Newtonian laws have to be modified even in the special relativity theory it seems that the term "inertial system" is meaningless. It is nevertheless important that we discover the reason for the existence of this term since an inertial system is characterised by the facts that it does not "rotate" and does not undergo "acceleration", and we shall therefore find an indication as to the meaning of these terms when we decide upon the significance of the term "inertial system".

The Newtonian laws of motion under no forces may be expressed as

$$\frac{d^2x^\alpha}{dt^2} = 0.$$

The nearest approach to these in the special relativity theory is

$$\frac{d^2x^\alpha}{ds^2} = 0.$$

In the general theory we have seen that it is usually impossible to find a coordinate system in terms of which *all* trajectories take the above form; what is the best we can do?

It can be shown that coordinate systems exist in terms of which *all trajectories*[7] *through the origin* take on the above form; they are called *normal coordinates*, and for distinction we use y^α for the coordinates of an event referred to such a system. In addition to the above they possess the following characteristics:—

[7] By the term "trajectory" we always refer to the trajectory of a particle acted upon by nothing except the gravitational influence of other bodies; such bodies we refer to as free bodies.

At the origin

$$\frac{\partial g_{\alpha\beta}}{\partial y^{\gamma}} = 0; \tag{8.1}$$

Corresponding to a given origin there is an infinite number of normal coordinate systems;

These are related by general Lorentz transformations;[8]

Not only all trajectories but all geodesics[9] *through the origin take on the simple form*

$$\frac{d^2y^{\alpha}}{ds^2} = 0, \tag{8.2}$$

and conversely.

We shall look upon the normal coordinate systems as the entities that the term inertial system was intended to describe; let us see the consequences of this. We have, immediately, the result that special relativity holds in the neighbourhood of the origin of coordinates. Again, it is evident that the time-like coordinate axis is a trajectory, since it satisfies (8.2), and moreover all the coordinate axes are geodesics; the time-like axis is such that for events on it the spatial coordinates remain zero; it therefore represents the trajectory of a body permanently at the origin of the spatial part of the system.

Integrating the equations of motion of trajectories through the origin we obtain

$$y^{\alpha} = p^{\alpha}s \tag{8.3}$$

where p^{α} are constants depending on the velocity of the particle and s is the "proper time." The time-like axis, in integrated form, is

$$\left.\begin{aligned} y^4 &= p^4 s \\ y^{\alpha} &= 0 \end{aligned}\right\}$$

and it follows, therefore, that the indications of the clock used in the coordinate system at the origin provide a "time that flows uniformly" for all trajectories through the origin.

Inertial frames are looked upon as being unaccelerated and non-rotating; that is to say the words "accelerated" and "rotating" are applied to a coordinate system to describe properties whose symptoms are a failure of the Newtonian laws of motion in that system; and furthermore there is considered to be an essential difference between a rotating system and one that is merely linearly accelerated. When we go over to the relativity theory it seems that we must regard the general term "accelerated", when applied to a coordinate system, as implying that it is not a normal coordinate system; but not every coordinate system that does not happen to be a normal coordinate system can be regarded as "accelerated" in any intuitive sense; for example, let us take a normal coordinate system and substitute for its set of clocks a

[8] We always refer to "orthogonal" normal coordinate systems, for which $g_{\alpha\beta}=0$ if $\alpha \neq \beta$ at the origin.

[9] For a trajectory is merely a particular type of geodesic since it must lie within the light cone.

new set of clocks the indication of each of which is related to that of the clock it displaced by, say,

$$t' = t^{1.003};$$

the resulting coordinate system will be such that trajectories through the origin will no longer "look like" straight lines —i.e. will no longer have the form

$$x^\alpha = p^\alpha s$$

—with respect to it; it is nevertheless not an "accelerated" system, neither does it rotate. We need not pause over the classification of this type of departure from normalcy since it evidently corresponds to the effect of using an ir-

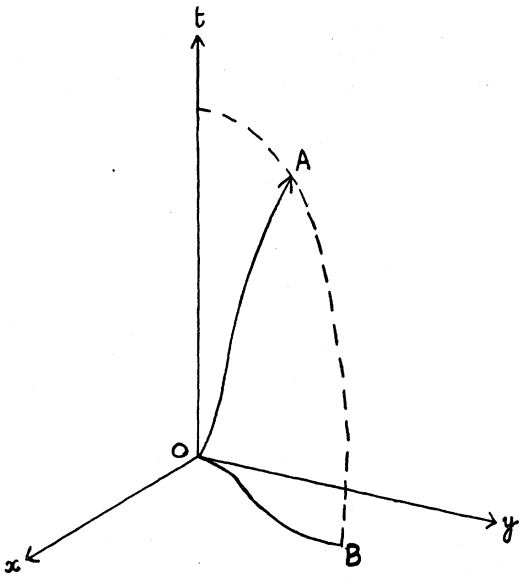

Fig. 2.

regular watch in the Newtonian theory; and similar remarks hold for coordinate systems that correspond to curvilinear coordinates in the Newtonian case. Let us rather attack the problem of characterising a truly accelerated coordinate system in the relativity theory; as a definition we can hardly object to the following: "a coordinate system is not accelerated if its spatial origin moves with uniform velocity", provided this really has meaning. We can supply a meaning to it very simply; the motion of the spatial origin is easily seen to be the curve denoted parametrically by the coordinates (o, o, o, t) of points on it; for uniformity of velocity we have only one criterion, namely the motion of a free particle. Hence we can at last define an accelerated coordinate system as one whose time-axis is not a trajectory.

The case of rotation is a little different; let us try to find a type of coordinate system that can be characterised as unaccelerated but rotating. It must evidently be such that its time axis is a trajectory. Speaking classically,

we recognise a coordinate system as rotating when all particles projected from the origin trace out spatial loci that, instead of being straight lines, are, in general, spirals described in the same sense about an axis—the axis of rotation—the only exceptions being those particles that were projected along the axis of rotation. This way of regarding the symptoms that betray the rotation of a coordinate system may be taken over with very little change into the relativity theory; in Fig. 2 let Ox, Oy and Ot represent respectively two spatial coordinate axes and the time axis, the third spatial dimension being, as usual, omitted. Let OA be any trajectory through the origin; denoting its initial direction by D_1 and that of the time axis (which is, by hypothesis, a trajectory) by D_2 we can define a two-dimensional surface uniquely determined by Ot and OA as the surface traced out by the geodesics through O whose initial directions are given by

$$D = D_1 + \lambda D_2$$

for different λ's; we shall refer to this as the geodesic surface determined by the intersecting geodesics Ot and OA. It will cut the three-dimensional space, $t=0$, in a curve, OB say. In general OB will not be a geodesic but it is doubtful if a satisfactory definition of rotation of a coordinate system can be obtained except in the cases when OB is a geodesic for every OA; such an assumption amounts to our taking Ox, Oy and Oz as geodesics and defining the space $t=0$ as depending on Ox, Oy and Oz in precisely the same way as the surface tOA depended on Ot and OA; physically this is the nearest approach to the concept of rectilinear spatial coordinates that can be made in the relativity theory; we shall therefore assume that OB is a geodesic; it is thus the analogue of the initial spatial tangent radius vector of the motion, OA. The equations representing OB will be of the form

$$\left.\begin{aligned} f_1(x, y, z) &= 0 \\ f_2(x, y, z) &= 0 \\ t &= 0 \end{aligned}\right\}$$

in terms of our coordinate system. The subsequent behaviour of this line in space will be represented by the two-dimensional surface

$$\left.\begin{aligned} f_1(x, y, z) &= 0 \\ f_2(x, y, z) &= 0 \end{aligned}\right\}$$

and this will therefore be the relativistic analogue of the motion of a direction marked on our coordinate system which originally coincided with the initial spatial direction of the motion of our free particle. Hence if OA lie wholly in this surface we can consider the spatial direction, OB, to be non-rotating; if, on the other hand, OA should curl out of this surface, say, to the left we regard the coordinate system as rotating in such a way that the spatial direction, OB, rotates to the right. We have, therefore, a test as to whether a coordinate system of the type to which we have restricted the discussion

has a rotation or not. The point is more easily seen if we consider a trajectory, OA, which determines an OB that happens to be a coordinate axis, say Ox; then our criterion for a rotation of which Ox is not the axis is that the geodesic plane, tOx,—whose points have coordinates of the form (x, o, o, t)—do not contain OA. The above does not imply that the coordinate system rotates as a whole; to verify this we must make our test for all possible spatial directions, OB, and even then we cannot, in general, obtain a satisfactory significance for the concept of rotation as a rigid body.

We have treated acceleration and rotation separately; whether they can be recognised and distinguished without ambiguity in a coordinate system that possesses both characteristics is a much more difficult question and one which, to the best of the author's knowledge, has not been investigated.

§9. Relative Motion of Two Bodies; The Velocity of Light

In the previous section we were concerned with properties ascribable to coordinate systems; we shall now consider what can be said regarding certain phenomena irrespective of the coordinate system used in their description. We shall begin with the relative motion of two bodies a finite distance apart, the problem being to discover to what extent we are justified in saying that they are moving towards or away from each other, or that they are rotating around each other. Let the curves AC and BD in Fig. 3 represent the motion of the

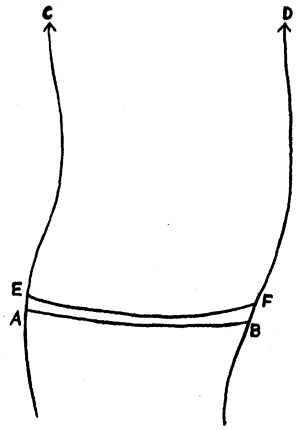

Fig. 3.

two bodies in terms of the four-dimensional world, and let us consider their relative radial motion first. With A as origin we can erect a normal coordinate system relative to which the particle AC is momentarily at rest—that is, we erect a normal coordinate system whose time-axis at its origin, A, is tangent to AC.[10]

[10] We have not stipulated that the two bodies be free bodies, acted upon only by gravitational influences, so that we cannot assume that AC and BD are trajectories and it follows therefore that the time-axis of a normal coordinate system will in general not coincide with AC over a finite interval.

Once a coordinate system has been decided upon we are able to talk about simultaneity; let B be the event on BD simultaneous with A— that is, let B be the event on BD whose t-coordinate in our coordinate system is zero. The three-space, $t=0$, will be a geodesic three-space perpendicular to AC at A; it will contain the geodesic joining A to B and hence this geodesic is perpendicular to AC at A. Let E be an event on AC near to A and let F be the event on BD simultaneous with it relative to our coordinate system; then there is only one geodesic joining E and F. If the length of EF be greater than AB we may think of the particles as receding from each other, if it be less as approaching, and if the same as having no relative radial motion—but this criterion is relative to a particular coordinate system associated with a particular event on a particular one of the particles! To what extent can we make it more objective? It can be stated without reference to coordinate systems essentially as follows; we choose an event, A, on one trajectory, draw the

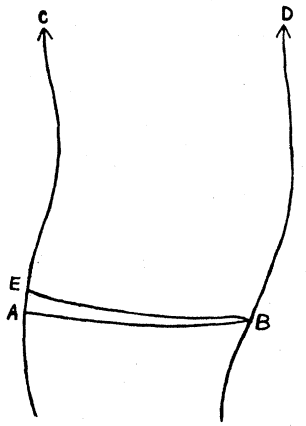

Fig. 4.

geodesic through A and perpendicular to AC that intersects the other trajectory and then look on the two particles as receding from each other, approaching each other, or having no relative radial motion, according as the angle between the geodesic and the second trajectory is obtuse, acute, or a right angle. We shall show that this criterion is a relative one in general; for in Fig. 4 let AC and BD represent the motion as before; let AB be the geodesic perpendicular to AC at A and let BE be the geodesic perpendicular to BD at B. Then if we take A as our initial event we must regard B as simultaneous with it, whilst if we look on B as the initial event we find that E is the event on AC simultaneous with it. In general A and E do not coincide, for a geodesic is uniquely determined by an event and a four-dimensional direction at that event, and it follows that A and E will coincide only if AB is perpendicular to both AC and BD. Hence relative radial motion can be defined only relatively to one of the particles; but in the special case in which at the instant under consideration the two particles have no relative radial

motion our criterion is independent of which trajectory is taken as the standard of synchronization.[11]

The case of the relative rotation of two bodies is closely parallel to the above discussion; it is hardly necessary to repeat many of the ideas employed in the case of relative radial motion and we shall therefore merely state the results; in Fig. 5 let AC and BD represent the motions of the two particles; let Σ represent the geodesic three-space perpendicular to AC at A; it will contain the geodesic, AB, that we have discussed above, the radial motion of the second particle relative to the first being decided by the angle ABD; the relative rotation of the second particle around the first at the instant of the event A is decided by the angle that BD makes with Σ in directions perpendicular to AB. If BD be perpendicular to Σ the two particles have no relative

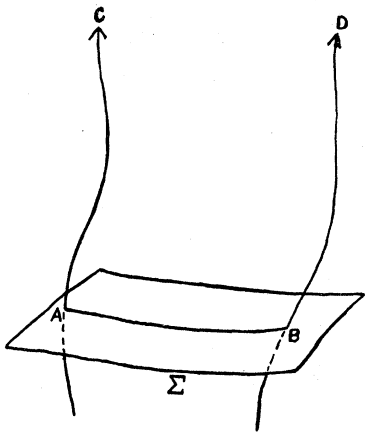

Fig. 5.

motion, either radial or transverse, and in this case only is the criterion independent of which particle is taken as stationary. In gen ral the motion of particle BD relative to particle AC is not the same as the motion of particle AC relative to BD.

[If we required that space-time be of such a sort that the relative motion of two bodies have an absolute significance independent of everything except the instant for which the motion is considered, so as to avoid, for instance, the possibility that although particle A is rotating around particle B nevertheless particle B is not rotating around particle A, we should find it necessary to endow space-time with the property of so-called distant parallelism.]

The results of the above discussion are in no way dependent upon the fact

[11] It would be possible to take as initial event some quite arbitrary event and to erect an arbitrary trajectory at that event to provide our synchronization; a definition of relative radial velocity could easily be obtained in terms of this construction, but it would involve a great deal of useless arbitrariness from our present point of view since our object here is to discover how far concepts like the relative motion of two particles can be made independent of arbitrary modes of observation.

that we have chosen two particles moving in an arbitrary manner instead of two free particles; in the special relativity theory there is a clear distinction between the two cases as the reader may easily see for himself by drawing the appropriate figures and remembering that the term free particle in special relativity means a particle acted upon by no influences whatsoever whilst in general relativity, where the gravitational field has been given a geometrical significance, the term implies a particle acted upon by the gravitational effect of other bodie but by nothing else.

The complications of the problem of the relative motion of two bodies in general relativity arise from the fact that we have to deal with events that do not lie infinitesimally close to one another; the situation is simpler if we consider the relative motion of two bodies at an instant at which they happen to coincide since, of course, the definitions of the various relative motions no longer require the specification of a particular particle as standard; it is

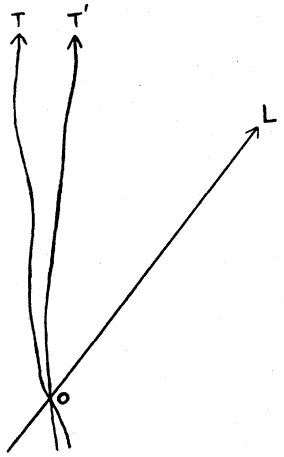

Fig. 6.

easily seen, for example, that the relative radial velocity is represented by the angent of the angle at which the two curves representing the particles' motions intersect. Let us consider the velocity of light from this point of view; our first notion is to consider the "velocity of light relative to" rather than the velocity of light; in Fig. 6 let OT represent the motion of a body that we shall consider as being our standard of rest and let OL represent the motion of the tip of a light ray containing the event O. The velocity of light relative to the body at the event O is represented by the tangent of the angle at which the curves OT and OL intersect; this is an objective definition of the relative velocity and is therefore independent of particular coordinate systems, nevertheless it is not this fact that gives to the velocity of light its characteristic property, completely analogous to that which it possesses in the special theory of relativity; for let us consider another body containing the event O, whose motion may be represented by OT', say. If bodies OT and OT' have a relative motion at O their tangent geodesics will not coincide

there; however the velocity of light relative to OT' is the same as that relative to OT as is at once seen to be the case since OL is characterised by the fact that its geodesic length is zero and hence $\tan^2\widehat{TOL} = -1 = \tan^2\widehat{T'OL}$;[12] it thus follows that the velocity of light at an event O relative to any standard body containing O is a constant, irrespective of the coordinate system and irrespective of the particular standard of rest employed. The first point implies that our definition of relative velocity is objective,[13] the second that the velocity of light is the same for all observers, just as in the special theory.

We have found that the velocity of light is $(-1)^{1/2}$; by a change of units we can not only make this real but also give it any numerical value we choose; it follows therefore, that the phrase "velocity of light" has significance in that it is independent of the observer, but has no definite numerical value notwithstanding; the conclusion to be drawn is that there is a natural relation between the units of length and time and that we do not employ it in the c.g.s. system. And, furthermore, whether the velocity of light remains a constant or not when expressed as cm/sec., say, is not the concern of the motion of light but is a reflection of the fact that our efforts at spreading a holonomic four-dimensional web of cm/sec. coordinate lines are unsuccessful in the presence of matter. The motion of light is the standard in terms of which everything else is expressed.

§10. The Doppler Effect

In this final section we discuss the term "Doppler effect"; our pres ntation will be by the aid of diagrams, since not only does this method avoid the use of mathematical symbolism but, once grasped, it provides a vivid comprehension of the relativity outlook and its points of difference from the Newtonian system.

We commence with a diagrammatic account of the classical theory of the Doppler effect. The Newtonian world is four-dimensional and we represent two of these dimensions in Fig. 7, omitting two of the spatial dimensions for the sake of simplicity. Let AC be the trajectory of a body at rest emitting monochromatic radiation; we are interested in two aspects of this radiation, its period and its velocity; the former we represent by dots evenly spaced on AC, which may be looked upon as marking successive moments of similar phase, and the latter we denote by the definite slope of the lines AA', BB',

[12] It is more accurate to say that the notion of angle breaks down when one of the lines in the definition is of zero length; the argument in the text is valid if we regard it as employing a convenient means for talking about analytical processes, the relative velocity being defined as the analytical expression which in ordinary cases represents the tangent of the angle between the two curves representing the motions of the two bodies, this definition being retained in the limiting case.

[13] In previous sections, notably the third, we have been concerned with "coordinate velocities" which may be defined as entities having as components the limiting values of $\delta x/\delta t$, $\delta y/\delta t$ and $\delta z/\delta t$ for an event on a trajectory; this entity is not independent of the coordinate system and moreover no *magnitude* is defined for it; it is, however, not without significance since, in some cases, certain coordinate systems appear to be of importance as being those used intuitively by the astronomer; see, for example, the next section.

etc., that represent the propagation of the disturbance through space. By the assumptions underlying the Newtonian theory, the three-dimensional spaces, t = constant, are all Euclidean and independent of the value of t, matter being considered as having no influence upon the structure of space or time; our figure is in fact the sort of diagram that is used in the compilation of railway time-tables, stationary bodies being represented by trajectories parallel to the time-axis and therefore perpendicular to the space-axes. Let OT denote an observer having no radial motion relative to the emitter, so that it is parallel to AC; the disturbance originating at the event A reaches our observer at the event, O, which is the intersection of AA' and OT; we may, without loss of generality, take O as the origin of space and time coordinates; the disturbance of similar phase, originating at B, will reach the observer at the event, P, which is on BB' and OT; the time interval between these two events is

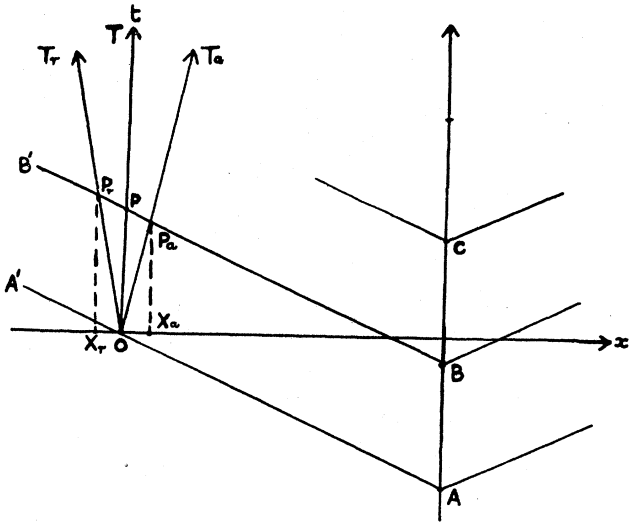

Fig. 7.

the distance PO, and this is the same as AB so that the frequency of the light received is the same as that of the light emitted. Now consider an observer approaching the emitter; he is represented by the trajectory, OT_a, inclined towards AC, the inclination depending on the velocity of approach; let AA' and BB' meet OT_a in O and P_a respectively; then the time interval between the reception of the disturbances from the events A and B is the distance between P_a and O *parallel to the time-axis*, i.e., it is the perpendicular distance, P_aX_a, from P_a to Ox; this is less than AB and therefore the frequency of the light received is higher than that of the light emitted. For a receding observer we have the trajectory OT_r and we see that there is an apparent lowering of the frequency. When the observer is fixed but the emitter moving the Doppler shift, as is well known, is slightly different in its dependence on the relative velocity of the emitter and observer from its value in the above cases; the

construction of the appropriate figure is not difficult. To be accurate we should point out that the measure of the alteration in frequency is performed against a standard emitter in the laboratory of the observer; in the present case on account of the Euclidean nature of our spaces, $t =$ constant, this point has little significance, but we shall see later, especially in the general relativity case, that it is of great importance.

We take up now the special relativity theory of the Doppler effect; this has sometimes caused a slight confusion for the following reason; it is known that in special relativity the addition of any velocity to the velocity of light has no effect, and yet the Doppler effect is often explained in the Newtonian theory by talking of the increase or decrease in the velocity of light relative to an observer due to his motion; the inference is that apparently a Doppler effect due to relative radial motion cannot exist in the special relativity the-

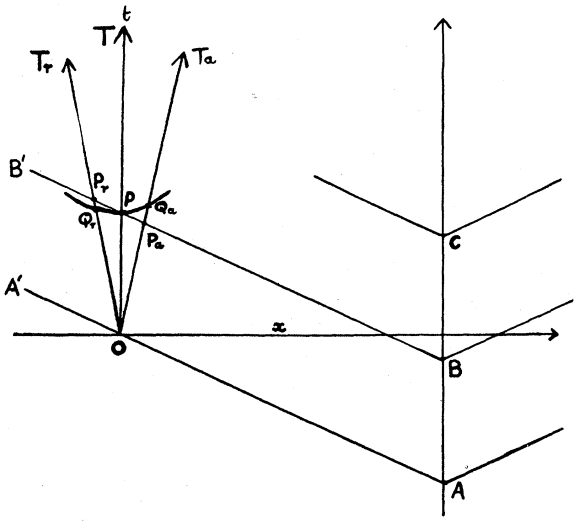

Fig. 8.

ory. The fallacy lies in the assumption that the Doppler effect depends upon an alteration in the relative velocity of light; it depends actually upon the alteration in the number of wave-lengths passing an observer in a second caused by the observer's motion relative to the emitter. In Fig. 8 we give the special relativity equivalent of Fig. 7. As before AC represents the emitter and AA', BB' etc., the propagation of the light emitted at the events A, B, etc., of similar phase. For the observer having no motion relative to the emitter the time interval between the reception of the disturbances sent out from the events A and B is PO and this is the same as AB so that there is no Doppler shift; however let us consider the case in which a transverse relative velocity exists; instead of having OT to represent our observer we must now take a trajectory OT' (Fig. 9) inclined to OT but lying in the plane through OT perpendicular to Ox. The interval between the reception of the disturbances at A and B is now $P'O$ and in special relativity we have to compare the interval between

P' and O with the interval representing a cycle of an emitter at rest relative to the observer; the assumption is made that the pulsations of any emitter beat out intervals of local time, ds, defined by

$$(ds)^2 = c^2(dt)^2 - (dx)^2 - (dy)^2 - (dz)^2,$$

since this has absolute significance whilst dt has not. There are two influences at work; firstly OP' is not the same length as OP in our diagram and secondly the unit of local time along OT' looks larger on our figure than the unit along OT, since the figure is drawn so that the coordinate system in which AC and OT are at rest actually looks rectangular. The curve PQ in the figure repre-

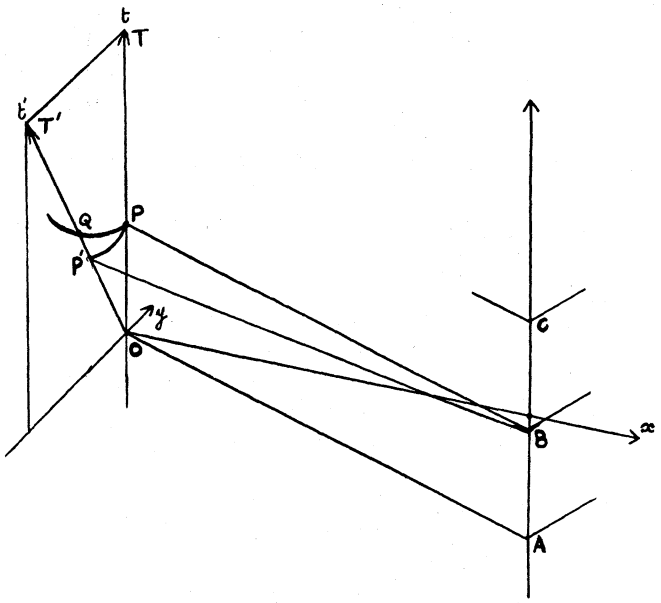

Fig. 9.

sents the locus of events in the plane TOT' that lie at a constant interval, OP, of local-time from O; it is, as is well known, the hyperbola

$$c^2t^2 - y^2 = \sigma^2,$$

where σ is the interval OP; it turns out that OQ is always larger than OP', that is, the interval OP' is less than the corresponding interval, OQ, ticked out by a local emitter similar to AC, and hence there is a Doppler shift to the violet due to the transverse motion; this effect is actually negligibly small for ordinary velocities; the figure exaggerates its usual magnitude since we have for the sake of clearness made the angle TOT' much larger than it should be for a relative velocity of ordinary magnitude.

Let us return to Fig. 8 and consider the Doppler effect caused by relative radial motion; for an approaching observer we take the trajectory OT_a. The disturbances originated at the events A and B reach him at the events O and

P_a; the curve Q_rPQ_a is part of the hyperbola of events in the plane TOA lying at a constant interval, OP, from O; let it cut OT_a at Q_a. We see that OP_a is less than OQ_a and this indicates a shift towards the violet. Similarly for the case of a receding observer, OT_r, since OP_r is greater than OQ_r we find a Doppler shift towards the red.

When the relative velocity has both transverse and radial components the situation can be investigated by drawing a suitable figure showing two space dimensions and the time dimension; we leave this for the amusement of the reader.

A comparison of Figs. 7 and 8 shows, when it is remembered that the angle between OT and OT_a or OT_r should be extremely small, that for ordinary velocities there is practically no difference between the spectral shifts predicted, for a given relative velocity, by the Newtonian and the special relativity theories; such difference as there is between the two formulae has the effect that whereas the Newtonian formula involves the "velocity of the emitter in space" the relativistic formula depends only upon the relative velocity of emitter and observer.

And finally, the Doppler effect in general relativity; the shift in spectral lines need no longer be considered as wholly due to relative motion—the existence of gravitating bodies can cause an apparent change of frequency; and anyway we have seen that the idea of relative velocity undergoes modification in the general relativity theory. We propose to consider a special case—the case of the most importance, as it happens; we shall consider the Doppler effect in the gravitational field of a body possessing spherical symmetry in space.[14] Let us assume that this body is not radiating away energy at an appreciable rate and that it is not changing at all rapidly in any other way; what may we expect concerning the gravitational field that accompanies it? Since the body, to a high degree of approximation, does not change we expect that its field is likewise unchanging; expressing this mathematically we say that the field must be such that it is possible to find a coordinate system in terms of which all the three-spaces, $t=$constant, are congruent—except for the minute disturbances arising from the motions etc., of what we shall consider as test-bodies. If such coordinate systems, then, are to be assumed to exist, it is natural for us to deal entirely with them to the exclusion of all others; a gravitational field capable of such coordinate systems is referred to as a statical field. There is another attribute of our field—its spherical symmetry; this evidently means that each one of our congruent three-spaces—they are of course non-Euclidean—possess spherical symmetry about the position of the field-producing body. With these two conditions, that the field be static and spherically symmetric, it was shown, by K. Schwarzschild, that there is only one type of space-time allowed by the limitations imposed by the field equations;[15] this space-time is such that, in terms of a suitable coordinate system, the $g_{\alpha\beta}$ take on such a form that the invariant $(ds)^2$ becomes

[14] The phrase actually has a meaning.

[15] It happens that the condition of spherical symmetry in conjunction with the field equations implies that the field be static.

$$(ds)^2 = c^2(1 - K/r)(dt)^2 - \frac{1}{(1 - K/r)}(dr)^2 - r^2(d\theta)^2 - r^2 \sin^2\theta(d\phi)^2, \quad (10.1)$$

where K is a constant proportional to the mass of the body producing the field and c is a constant relating the units of t and r. The quantities r, θ and ϕ are the nearest analogue possible to spherical polar coordinates in our non-Euclidean congruent three-spaces. The form (10.1) for $(ds)^2$ is referred to as the Schwarzschild line-element; it is for a space-time possessing such a line-element that we shall consider the significance of a shift in spectral lines. In Fig. 10 we attempt to picture certain features of such a space-time, but it is evident that great success will not attend our efforts if we aim at anything more than an indication of the state of affairs; we suppress two spatial

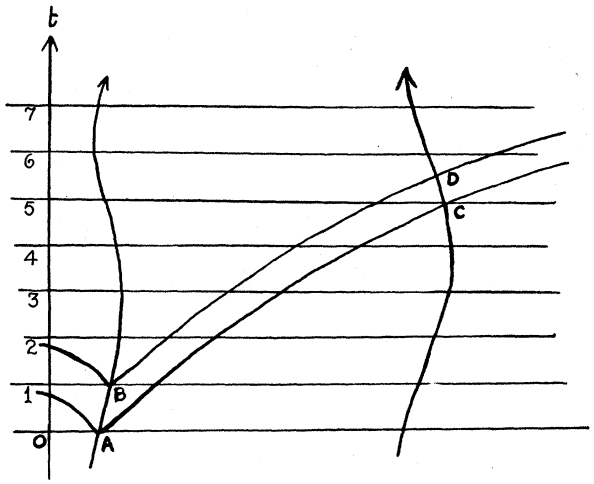

Fig. 10.

dimensions and represent our curved three-spaces by straight lines perpendicular to the time axis. Let Ot be the trajectory of the field-producing body; it is also the time axis in the coordinate system of (10.1).[16] Let the events 1, 2, 3 $\cdots$ represent events on Ot occurring at equal intervals of *coordinate time;* the spaces represented by the parallel lines through these events are thus separated by equal intervals of coordinate time, and all events in a given such space are simultaneous with regard to the coordinate system used. Let A and

[16] It should be realized that this is *not* a normal coordinate system; it is owing to this fact that light is "bent" by the field of the sun; it is assumed that when we deal with the sun's field experimentally we intuitively set up a statical coordinate system having pseudo-spherical-polar spatial coordinates; relative to these the trajectory of the tip of a light ray does not take a form such as, say,

$$\left.\begin{aligned} r\cos\theta &= \text{const.} \\ \phi &= \text{const.} \\ \theta/t &= \text{const.} \end{aligned}\right\},$$

that is, it does not "look" straight, and it is only in this sense that we may talk of light being bent by a gravitational field.

B be two events on the emitter's trajectory representing the beginning and end of one cycle of the emitting mechanism; let the disturbances originating at A and B be propagated along AC and BD, respectively, to arrive at the trajectory, CD, of the observer at C and D respectively. On account of the static nature of our space-time the propagation of light relative to our coordinate system is independent of the particular instant at which it takes place; this means that the curves BD and AC on our diagram are, except for the end pieces, congruent and can be brought into coincidence by a bodily shift of one of them parallel to the time axis; of the significance of this point we shall have more to say in a moment; meanwhile we consider the general Doppler effect; the events C and D mark the reception of the beginning and end of a single wave-length of the light sent out by the emitter; we have to

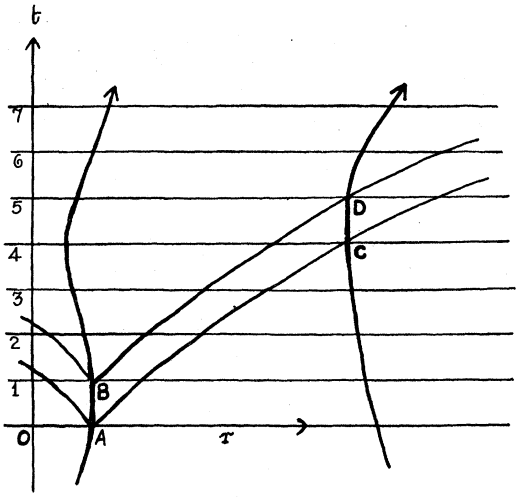

Fig. 11.

compare the interval of local-time separating the two events C and D with the interval of local time separating the events of the beginning and end of a cycle by an emitter local to the observer; let C and E be such events (we do not draw the event E in the figure since we do not know yet whether it ought to lie between C and D or beyond D); then if CD be greater than CE, that is if E fall between C and D, it will follow that in a given amount of coordinate time the local emitter provides more cycles at CD than does the distant emitter AB, so that there is an apparent shift to the red in the light of the emitter AB compared with that of the local emitter; if CD turn out to be less than CE we shall obtain a shift towards the violet. There remains the difficulty of computing where E should fall on our diagram; we make use of the same assumption as we employed in the special relativity case; that is, we assume that, since the atoms of the emitter AB are of the same kind as those of the local emitter, the interval, AB, of local-time beat out by it is the same as the interval, CE, of local-time beat out by the standard emitter of the ob-

server. The computation is straightforward and we shall not give the details; but before we discuss the factors upon which the total Doppler shift depends it is advisable that we consider a special case; this is illustrated in Fig. 11, which differs from Fig. 10 in that the element AB is parallel to the time axis as also is the element CD; this signifies that at the moments of emission from the events A and B the emitter has no coordinate velocity and at the moments of reception of the disturbances sent out from A and B the observer has no coordinate velocity; with this simplification we are able to see more clearly what is taking place in the mathematical theory. Since the curves of propagation AC and BD are congruent and are separated through a shift parallel to the time-axis it follows that the coordinate time between A and B is the same as the coordinate time between C and D; that is the value of dt between A and B is the same as its value between C and D. But by (10.1) we see that the corresponding value of ds for AB is given by

$$(ds_1)^2 = c^2(1 - K/r_1)(dt)^2,$$

where r_1 is the coordinate distance from O to A, since dr, $d\theta$, and $d\phi$ are all zero for AB; and for CD we find

$$(ds_2)^2 = c^2(1 - K/r_2)(dt)^2,$$

where r_2 is the coordinate distance from 4 to C. Since r_2 is greater than r_1 and dt is the same for AB and CD we see that ds_1 is smaller than ds_2. But a local emitter belonging to the observer will beat out intervals ds_1, according to our assumption, and therefore in a given amount of time the local emitter will beat out more cycles than are received from the distant emitter, AB, that is, there will be a shift to the red in the spectrum of the distant emitter when compared with a similar emitter at CD. If the observer were nearer to the time-axis, that is, nearer to the field-producing body, than the emitter the shift would be towards the violet. The point to note is that the above represents the nearest approach that can occur to the case of no relative motion between emitter and observer and yet a Doppler shift occurs; in the general case of Fig. 10 instead of having dr, $d\theta$ and $d\phi$ all vanish for AB and CD, and dt the same for both, we now have two different sets of values for dt, dr, $d\theta$ and $d\phi$ and for these two sets the corresponding values of ds must be computed from (10.1) and the results employed in the manner already explained.

The general Doppler effect for the statical spherically-symmetric field thus depends on several causes such as a particular choice of coordinate system, the positions of emitter and observer, and their respective velocities relative to this coordinate system at the instants of emission and reception. Nevertheless the Doppler effect is an objective phenomenon having nothing to do with any particular choice of coordinate system since it is no more than a comparison of the frequency of light received from a distant source with that given out by a local emitter, and a statement that, for example, five cycles are received from the distant emitter in the time taken for the local emitter to give out six cycles is one that involves no mention whatsoever of coordinate systems. We must therefore expect to be able to express the causes

of the shift in objective terms without reference to a particular coordinate system; now the curves AB, CD, AC, BD are objective, being trajectories of particles and light and the only reference to the coordinate system of (10.1) remains in the quantities r_1 and r_2; these quantities have objective significance, for it can be shown that, although the coordinate system of (10.1) is not a normal coordinate system, the radii vectores,

$$\left.\begin{aligned} t &= \text{const.} \\ \theta &= \text{const.} \\ \phi &= \text{const.} \end{aligned}\right\},$$

are actually geodesics, and the quantity r belonging to an event is related in a definite, though quite complicated, manner to the geodesic distance, $\int ds$, along the radius vector from Ot to this event; moreover, since the radius vector is the geodesic through this event that cuts Ot perpendicularly it is uniquely determined in an objective manner by the event.

Thus we see that the Doppler effect is an objective effect in the general relativity theory, but its significance cannot be adequately described in terms of ordinary concepts like relative velocity since such concepts lose their objective character in the general theory.

In the non-statical case the Doppler effect depends not only upon the position and motion of the emitter and observer and of all matter exerting gravitational influence but also upon the instant at which the observation is made, and when we remember that the finiteness of the universe may be, in some cases, a further large contributory cause it is evident that a complete disentanglement of all the elements that have gone to the production of a given Doppler shift would require far more mathematics than we have permitted ourselves during this article.