AAPT PHYSICS EDUCATION | AMERICAN JOURNAL of PHYSICS

General relativity primer

Richard H. Price

Citation: American Journal of Physics **50**, 300 (1982); doi: 10.1119/1.12853
View online: http://dx.doi.org/10.1119/1.12853
View Table of Contents: http://scitation.aip.org/content/aapt/journal/ajp/50/4?ver=pdfcov
Published by the American Association of Physics Teachers

Editor's note

This review article, the second in a series, is made possible by a grant from the National Science Foundation.

General relativity primer

Richard H. Price
Department of Physics, University of Utah, Salt Lake City, Utah 84112

(Received 14 July 1980; accepted for publication 2 December 1981)

In this tutorial article the physical ideas underlying general relativity theory are discussed and the basic mathematical techniques (tensor calculus, Riemann curvature) needed to describe them are developed. The general relativity field equations are presented and are used in several applications including a discussion of black holes.

I. INTRODUCTION

A. Purpose and outline

Special relativity theory (SRT) is a part of the intellectual toolbox of all physicists and a feature of the physicist's education even at the undergraduate level. The novel concepts of SRT, so shocking in 1905, hold no special terror now. The same, regrettably, cannot be said for the general relativity theory (GRT), Einstein's relativistic theory of gravity. The imagery of space-time curvature, and such exotica as black holes, give GRT such a recondite aura that it is too often regarded as hopelessly mystical, even by students and teachers who accept quantum mechanics as a perfectly reasonable description of the world. It is my goal in this article to show that this viewpoint on GRT is unjustified, that relativistic gravity is intuitively accessible and that space-time curvature is a natural conceptual basis for it. More specifically this article presents the mathematical and physical structure of GRT for a student or a teacher of physics, or a physicist in another field, in such a way that these readers can understand how calculations are done in GRT and what they mean. This article then is intended to present in a fairly small number of pages a subject usually dealt with in full length textbooks.[1] This is not, however, a "popular" introduction limited to metaphors, analogies, and word pictures. Such introductions (and in fact Sec. II below could stand alone as one) are of value but they do not teach the theory. They give some answers but not the general scheme for finding answers. This article presents GRT as a physically motivated *mathematical* theory of gravity. The distinction between such a presentation and a popular one is particularly sharp for GRT since the necessary mathematics of tensors is not part of the background of most physicists. To avoid tensor calculus would be to avoid a meaningful expression of the ideas of GRT. To include the mathematics, unfortunately, engenders a great danger, that of the mathematical trees obscuring the physical forest. It is much too easy to forget that all the formulas and transformations, and all the mathematical symbols dripping with subscripts, are part of a description of the physical world. The necessary mathematics of tensors and of curvature are introduced in this article in the most painless and the most physically motivated way I could manage, but I still feel it necessary to urge the reader in the strongest possible terms never to lose sight of the simple underlying physical and geometrical principles.

Clearly in a small article covering a large subject, sacrifices must be made. The most regrettable sacrifice will be the omission of all but a cursory discussion of the stress-energy tensor, the "source" of the gravitational field. Also omitted will be many mathematical details, some of them formal and elegant, some of them tricky and technical, some of them useful for reducing very difficult calculations to merely difficult ones. Missing too will be most of the applications of GRT to problems of current interest. A useful discussion is given, however, of that aspect of GRT that stimulates the most interest and confusion: black holes.

I assume that the reader comes to this article with two prerequisites: First, a familiarity is required with partial differential equations and their application in physics, as would certainly result from, say, a junior- or senior-level course in electrodynamics. Experience with partial differential equations will be necessary for an appreciation of the meaning of the GRT field equations; specific techniques for solving such equations will not be of importance. The second requirement is a comfortable familiarity with SRT, with the Minkowski space-time description of SRT, and with the usual SRT jargon (worldlines, proper time, four-vectors, etc.). The conventions used in this article are outlined in Sec. I B.

This article is organized as follows: Section II presents a heuristic overview of gravity and space-time geometry. The physical ideas that are the basis of the theory are presented without their full mathematical realizations. This section serves as the motivation for the mathematical development in the subsequent two sections. It can also stand alone as a nonmathematical description of the structure of GRT. For some readers the phenomenological picture given in Sec. II may be enough. Others I hope will be inspired to read further in order to understand more quantitatively the ideas of Sec. II. The mathematical tools necessary for a quantitative understanding are developed in Secs. III and IV. In Sec. IV the field equations of GRT are presented and discussed, bringing to fruition the seeds planted in Sec. II. Section V contains a brief discussion of the weak field limit of GRT, in which the connection is made with Newtonian gravity and some features of gravitational waves are discussed. The use of the field equations and the interpretation of the result is exhibited in Sec. VI for the simplest but astrophysically most important case of a spherically symmetric source. In this section the Schwarzschild space-time geometry is derived from the basic starting point of symmetry principles

and the field equations, and is investigated up to and including the nature of the Schwarzschild black hole.

B. Notations and conventions

In this article we use for the most part the conventions of Misner *et al.*[1] except that we shall bother to include explicitly all factors of c (speed of light) and G (universal gravitational constant). To minimize the frustrations of unclear notation, space is taken, here at the outset, to spell out most of these conventions.

Some of the statements in this article will apply specifically to the four-dimensional space-time of relativity; others will apply more generally to a space (with or without timelike directions) of any (finite) dimension N. To help with this distinction the term "space(time)" will be coined to denote the general case; "space-time" will be reserved for the specific $N = 4$ case with one timelike direction.

A point in N-dimensional space(time) is specified by the values of N coordinates. Convenience will require the use of several different schemes of naming coordinates. As an example, spherical polar coordinates for ordinary $N = 3$ space can be written (r,θ,ϕ) or (x^r,x^θ,x^ϕ) or (x^1,x^2,x^3). The general symbol for a coordinate will use a lower case Greek letter, as in x^μ, with μ taking any value (e.g., 1,2,3 or r,θ,ϕ in our example).

In the Minkowski space-time of SRT, measurements in a particular inertial reference frame are most simply described with the Minkowski coordinates (t,x,y,z) for that frame. In this case a few notational peculiarities apply. First, the numerical indices run from 0 to 3 and x^0 will always have the meaning ct, that is, "speed of light $\times$ timelike coordinate". When it is necessary to single out the spatial coordinates in space-time, Latin indices will be used, as in x^i. As examples of these conventions, we have, for Minkowski coordinates,

$$\begin{aligned} x^\mu &= \text{any of } x^0,x^1,x^2,x^3 \\ &= \text{any of } ct,x,y,z, \\ x^i &= \text{any of } x^1,x^2,x^3 \\ &= \text{any of } x,y,z\,. \end{aligned}$$

Numerical indices are usually less clear than the others and will usually be avoided in this tutorial article except for x^0, which is useful in place of ct.

When two different coordinate systems are used for the same space(time) it is conventional to distinguish one by a prime as in $x^{\mu'}$. A familiar example is the Lorentz transformation, the transformation between two Minkowski coordinate systems. In the simple case of relative motion at speed v, purely in the x direction, the transformation is

$$\begin{aligned} &x' = \gamma(x - vt), \quad t' = \gamma(t - vx/c^2)\,, \\ &y' = y, \quad z' = z, \quad \gamma \equiv (1 - v^2/c^2)^{-1/2}\,. \end{aligned} \tag{1.1}$$

The more general Lorentz transformation is more easily given in the form

$$x^{\mu'} = \sum_\nu \Lambda^{\mu'}{}_\nu x^\nu\,. \tag{1.2}$$

The summation here is taken to be over all allowed values of ν. For the simple case of pure x motion the Lorentz matrices include, for example,

$$\Lambda^{0'}{}_x = -v\gamma/c\,.$$

An extremely useful and universal convenience is the "summation convention," which simply says that when a literal index is repeated in a term, both as a subscript and a superscript, summation is assumed. Equation (1.2) can then be written as

$$x^{\mu'} = \Lambda^{\mu'}{}_\nu x^\nu\,. \tag{1.3}$$

In this equation ν (index summed over) is called a dummy index and μ' (not summed over) a free index. The symbols used for free and for dummy variables are, of course, arbitrary so that

$$\Lambda^{\alpha'} = \Lambda^{\alpha'}{}_\beta x^\beta$$

means exactly the same as Eq. (1.3). The appearance in a term of triply repeated index, or different free indices in different terms of an equation, will not occur.

For two infinitesimally separated points ("events") in space-time, with coordinate separation dt, dx, dy, dz the "interval" between them is defined by

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

We now introduce very important symbols $\eta_{\mu\nu}$ and $\eta^{\mu\nu}$ defined by

$$\eta^{\mu\nu} = \eta_{\mu\nu} = \begin{cases} 0 & \text{if } \mu \neq \nu \\ -1 & \text{if } \mu = \nu = 0 \\ +1 & \text{if } \mu = \nu = 1,2, \text{ or } 3\,. \end{cases} \tag{1.4}$$

Note that $\eta^{\mu\nu}$ and $\eta_{\mu\nu}$ can be viewed as matrix inverses, since (with the summation convention)

$$\eta_{\mu\alpha}\eta^{\alpha\nu} = \delta^\nu_\mu\,, \tag{1.5}$$

where δ is the Kronecker delta. With the $\eta_{\mu\nu}$ symbol the expression for the interval can be written as

$$(ds)^2 = \eta_{\mu\nu}\, dx^\mu\, dx^\nu\,. \tag{1.6}$$

The displacement between the two events is said to be spacelike, null (lightlike), or timelike, if $(ds)^2$ is, respectively, positive, zero, or negative. For timelike displacements $(ds)^2$ is often replaced by $(d\tau)^2 = -(ds)^2/c^2$, where $d\tau$ is the "proper time" between the events, i.e., the clock time measured by an observer whose worldline passes through both events.

The crucial property of Eq. (1.6) and Lorentz transformations is that the interval for a given pair of events, has the same value when evaluated in any Minkowski coordinate system [just as $(ds)^2 = (dx)^2 + (dy)^2$ has the same value in any two-dimensional Cartesian coordinate system].

Vectors (and later, tensors) will be represented by boldface sans serif symbols. Of particular importance in SRT is the four-velocity of a particle, a tangent to the particle's worldline, defined by

$$\mathbf{U} \equiv \frac{\mathbf{ds}}{d\tau}\,,$$

where $\mathbf{ds}$ is the differential displacement along the particle worldline and $d\tau$ the proper time required for this displacement. The components of $\mathbf{U}$ in a Minkowski coordinate system are

$$U^0 = c\frac{dt}{d\tau}\,, \quad U^1 = U^x = \frac{dx}{d\tau}\,, \quad \text{etc.}$$

For a Lorentz transformation the components of $\mathbf{U}$ (or any vector) transform [see Eq. (1.3)] according to

$$U^{\mu'} = \Lambda^{\mu'}{}_\nu U^\nu\,. \tag{1.7}$$

The inner product ("dot product") of two vectors in

space-time is defined by

$$\mathbf{A}\cdot\mathbf{B} = -A^0B^0 + A^1B^1 + A^2B^2 + A^3B^3 = -c^2A^tB^t + A^xB^x + A^yB^y + A^zB^z .$$

With the $\eta_{\mu\nu}$ symbol this is simply

$$\mathbf{A}\cdot\mathbf{B} = \eta_{\mu\nu}A^\mu B^\nu . \tag{1.8}$$

Note that the dot product of the four-velocity with itself is always

$$\mathbf{U}\cdot\mathbf{U} = \eta_{\mu\nu}U^\mu U^\nu = \eta_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = -c^2 . \tag{1.9}$$

The idea of velocity-dependent mass, presented in some treatments of SRT, will be avoided. The "mass" of a particle will always denote its rest mass, a fixed velocity independent constant. The four-momentum of a particle is taken as

$$\mathbf{p} \equiv m\mathbf{U} . \tag{1.10}$$

For photons $\mathbf{U}$ is not defined (since $d\tau = 0$ along a photon path) but photon four-momentum remains a valid concept. The components of particle four-momentum in a given Minkowski coordinate system are

$$p^0 = mc\frac{dt}{d\tau} = mc\gamma ,$$

$$p^i = m\frac{dx^i}{d\tau} = m\gamma\frac{dx^i}{dt} ,$$

where i denotes a spatial index (1,2,3 or x,y,z) and γ is the Lorentz factor. It is often useful, in space-time, to retain the concept of an ordinary three-dimensional spatial vector, such as the ordinary velocity. When it is necessary to distinguish such three-dimensional vectors boldface roman symbols will be used as in

$$\mathbf{v} = \frac{d\mathbf{x}}{dt}$$

for the ordinary three-velocity. With this notation we have, for example,

$$\gamma = (1 - \mathbf{v}\cdot\mathbf{v}/c^2)^{-1/2} ,$$

$$p^0 = mc\gamma, \quad \mathbf{p} = m\gamma\mathbf{v} .$$

II. RELATIVITY, GRAVITY, AND GEOMETRY

A. SRT, forces, and gravity

There is a common misconception that SRT is adequate only for particles moving at a constant velocity. This is totally false; accelerations and forces *except for gravity* are completely within the scope of SRT. We can best understand what is so special about gravity by comparing it to a simpler physical interaction, classical electrodynamics. The example will be particularly clear if we can imagine that physics developed drastically differently than it actually did. Imagine that we knew and believed SRT beyond question, but had only very fragmentary knowledge of electrodynamics.[2] In particular let us suppose that experiments had been performed only with very slowly moving charges, and with charge densities that varied only very slowly. These experiments then would suggest that there was a potential Φ_e that was related to electric charge density ρ_e according to (esu units)

$$\nabla^2\Phi_e = -4\pi\rho_e , \tag{2.1}$$

and that a particle of charge q experienced a force associated with Φ_e, given by

$$\mathbf{F} = \frac{d\mathbf{p}}{dt} = -q\nabla\Phi_e . \tag{2.2}$$

(The boldface type here denotes ordinary three-vectors.)

Now SRT puts severe constraints on any viable theory of a physical interaction: the interaction must not allow us to distinguish one inertial reference frame (i.e., one Minkowski coordinate system) from another. That this requirement is violated by the above theory is most immediately evident in Eq. (2.1) that says that a change in ρ_e instantaneously results in a change in Φ_e everywhere. In SRT, due to the relativity of simultaneity, "instantaneous" is not an absolute judgment. Information can be propagated instantaneously only in one inertial frame so Eq. (2.1) can apply only in one inertial frame. The theory, then, distinguishes one frame from all others and hence violates SRT. This difficulty can easily be patched up by replacing Eq. (2.1) with

$$\nabla^2\Phi_e - \frac{1}{c^2}\frac{\partial^2\Phi_e}{\partial t^2} \equiv \Box^2\Phi_e = -4\pi\rho_e , \tag{2.3}$$

according to which changes in Φ_e propagate at the speed of light. The extra time derivative term, we could argue, is just too small to have been noted in any experiments performed.

This result is still not acceptable if we consider Φ_e to be a scalar field, i.e., a field measured in all inertial reference frames to have the same value at any particular space-time point. If Φ_e is a scalar then it is straightforward to show that the left-hand side of Eq. (2.3) is a scalar. What about the right-hand side? Electric charge is a scalar, i.e., measured to be the same in all frames. (See Ref. 2 for an argument for this.) But *charge density* cannot be a scalar. Let a group of charges, with total charge Q, be at rest in frame S and occupy a volume V in that frame. In a relatively moving frame S' the volume occupied will be smaller due to Lorentz contraction: $V' = V/\gamma$ where γ is the Lorentz factor. Since charge is a scalar $Q' = Q$ and hence the charge density measured in S' is

$$\rho_e' = Q'/V' = Q/(V/\gamma) = \gamma\rho_e . \tag{2.4}$$

The proper treatment of charge density in fact is to consider it as a part of a vector, the charge-current four-vector $\mathbf{J}$, with components as measured in any particular inertial frame given by

$$J^0 = c\rho_e \quad J^i = i\text{th component of current density } (i = x, y, z). \tag{2.5}$$

(Note that $J^{0'} = \gamma J^0$ if $J^i = 0$.)

We are forced to conclude that the right-hand side of Eq. (2.3) is a component of a vector. But the principles of SRT do not allow us to accept an equation of the form "scalar = component of a vector." Such an equation would make different predictions, in different coordinate systems, about absolute things. There is only one way out of this dilemma. Both Φ_e and ρ_e must be the "time components" of vectors and Eq. (2.3) is only a part of the description of the relationship between those vectors. More specifically we must construct a four-vector $\mathbf{A}$ such that $A^0 = \Phi_e$. Equation (2.3) then reads

$$\Box^2 A^0 = -4\pi c^{-1}J^0 \tag{2.6}$$

and it must only be one of four such equations that can be summarized as[3]

$$\Box^2 A^\mu = -4\pi c^{-1} J^\mu . \tag{2.7}$$

We should now play a similar game with the force equation (2.2). We must replace it with an equation that tells us, in terms of **A**, how to compute $\mathbf{dp}/d\tau$ and that, for slow particles (the only kind our imaginary experimentalists have studied), reduces to Eq. (2.2). A first halting attempt at the spatial components ($i = x, y, z$) of this equation might be

$$\frac{dp^i}{d\tau} = -\frac{q}{c}\frac{\partial}{\partial x^i}(A^0) . \tag{2.8}$$

This particular choice does not quite work; it turns out, for example, that it predicts changing particle mass. Rather than delaying a consideration of gravity by playing this game out in detail let us just write down the well-known answer:

$$\frac{dp_\mu}{d\tau} = \frac{q}{c}\left(\frac{\partial A_\nu}{\partial x^\mu} - \frac{\partial A_\mu}{\partial x^\nu}\right) U^\nu , \tag{2.9}$$

where

$$p_\mu \equiv \eta_{\mu\alpha} p^\alpha , \quad A_\mu \equiv \eta_{\mu\alpha} A^\alpha ,$$

and **U** is the four-velocity of the charged particle.

The result of this intellectual exercise is rather impressive. We have started from a very sketchy description of a physical law, based on very limited experimental data, and by satisfying the requirements of SRT have arrived at a much richer and more complicated theory, one that predicts new observable phenomena (magnetic forces, etc.) for which, according to the suppositions of our intellectual exercise, there was no experimental evidence. There is, however, no alternative. If relativistic ideas are correct the simple original theory cannot be correct.

The above exercise is in no way a summary of the history of the development of electrodynamics. Because large electric currents can be generated easily, magnetic forces were known to be part of electrical phenomena rather early, and it was electrodynamics that led to SRT rather than the other way around. The situation for gravity early in this century was drastically different and remains so even now. It is not possible to generate large mass currents and to do easily the experiments that reveal the relativistic nature of gravitation. Nothing makes this clearer than the fact that Newtonian gravity theory, a rather close analog of electrostatics, still suffices for almost all applications. For gravity then we would seem to be in a position to repeat our intellectual exercise in earnest. This after all (not curved space-time!) seems the common sense approach: to construct a theory of gravity compatible with SRT.

Our starting point is almost the same as in the case of electrodynamics. We have fragmentary experimental evidence about a physical interaction and the evidence suggests a potential Φ, satisfying

$$\nabla^2 \Phi = 4\pi G \rho \tag{2.10}$$

(where ρ is the mass density) and a force on a particle of mass m given by

$$\mathbf{F} = -m\nabla\Phi . \tag{2.11}$$

These two equations constitute the classical Newtonian theory. They are not, of course, compatible with SRT (e.g., instantaneous propagation of changes in Φ) but the equations are so strikingly similar to those of electrostatics that the path to modifying the theory seems clear. We will, however, run into two obstacles on this path: a minor one that introduces some technical complications, and a major one that suggests a new conceptual framework for gravity.

To see the minor problem let us investigate ρ in Eq. (2.10) as we investigated charge density ρ_e in Eq. (2.4). The source of the Newtonian gravitational field is

ρ = "mass"/volume.

But it is unclear in Newtonian theory what "mass" is to be included. It could be only rest mass, e.g., the sum of the rest masses of the fundamental particles making up a star. It might, on the other hand, mean "mass-energy," e.g., the sum of $c^{-1} p^0$ for all the particles, including photons, in the star; in this case kinetic energy, radiation energy, etc. contribute as well as rest mass. The difference between the two is tiny for ordinary matter and ordinary astrophysical objects. We shall accept for now and justify presently that it must be the latter choice, "mass-energy," that is the source of Φ. It is better, then, to write ρ as

$$\rho = c^{-2}(\text{sum of all mass-energy})/\text{volume} . \tag{2.12}$$

We can now follow the same pattern of argument that led to Eq. (2.4). Let a group of particles be at rest in frame S and occupy volume V as measured in that frame. Since the particles are all at rest only their rest mass contributes to mass-energy and hence the mass-energy is the sum $c^2 \Sigma m_i$ of the particle rest masses, and

$$\rho = \sum m_i / V .$$

In another frame S' a particle of rest mass m_i will have energy $\gamma m_i c^2$ and the particles will occupy a volume $V' = V/\gamma$. The source of the gravitational field in this frame will then be

$$\rho' = \sum m_i \gamma / V' = \gamma^2 \left(\sum m_i / V\right) = \gamma^2 \rho . \tag{2.13}$$

What are we to do with this? The γ^2 cannot appear in the Lorentz transformation of a scalar or of vector components, so what sort of mathematical object is ρ? The answer is to treat ρ as a component of a second-rank tensor and, as required by SRT, to treat Φ also as the component of a second-rank tensor, and to continue with arguments similar to those we used for electrodynamics (e.g., replace ∇^2 by $\Box^2$, etc.). To work out the details would be distractingly technical here because tensors have yet to be discussed or introduced. At this time it is sufficient for us to know simply that the details *can* be filled in and that it seems we can arrive at a theory that differs from electrodynamics in complexity but not in its conceptual framework (i.e., there is no hint of "curved space-time"). This theory, in fact, is simply the linearized gravitational theory to be described in Sec. V.

Before we look at the major problem with such a theory it will be necessary to reconsider why mass-energy rather than mass was used in Eq. (2.12). If we change this interpretation the mathematical form of our theory changes: rest mass is a scalar (like electric charge) so Eq. (2.13) is replaced by $\rho' = \gamma\rho$, and we no longer need to deal with tensors. The decision on the meaning of ρ will be crucial not only to the resolution of the ambiguity already mentioned, but also to the issue of the nonlinear nature of gravity.

For our purposes we shall need to consider an object at rest, say a small container of gas, and to distinguish three operationally defined types of mass for it: First is "inertial mass" M_I, the resistance to acceleration. This is the usual mass assigned to the object in SRT. It has contributions

from all the forms of mass-energy within the object. The container of gas, for example, certainly has a contribution to its inertial mass due to the kinetic energy of the thermal motion of the gas. A second type of mass is "passive gravitational mass" M_P. This is the mass that determines how much pull gravity has on the object; it is the mass that multiplies $\nabla\Phi$ in Eq. (2.11). The third type of mass is "active gravitational mass" M_A, the mass that generates the gravitational field. The density of this type of mass is the "ρ" that is needed in Eq. (2.10).

An experimental result that is of fundamental importance to gravity theory must now be introduced: All objects, whether containers of gas, rocks or feathers, beggars or kings, experience the same acceleration in a gravitational field, independent of their internal constitution. This basic fact, called the "weak equivalence principle," has been experimentally verified[4] with ever increasing precision by scientists, starting with Galileo who around 1610 confirmed it to one part in 10^3. The weak equivalence principle tells us that in a gravitational field the force on an object must always be proportional the object's resistance to acceleration, and therefore that the ratio M_P/M_I must be the same for every object. Convenience dictates choosing constants (e.g., the gravitational constant G) to make this ratio unity.

To justify Eq. (2.12) we must show that M_A is the same as M_I or equivalently M_P. Many interesting thought experiments can be performed to illustrate this point but we shall satisfy ourselves here with a simple one. Let us imagine that two containers 1 and 2, of the type we have been discussing, are connected by a rigid rod as in Fig. 1. Each container has N particles of mass m and the particles are at rest as if they constituted a zero temperature gas. Let us suppose that the whole contraption (containers, rod, particles) is initially at rest. The gravitational force on container 1 caused by 2 is equal and opposite to that on 2 due to 1, so the net force on the system is zero. Suppose now that the rest mass of a single particle in container 2 is completely converted to energy and that this energy shows up as thermal motions of the remaining particles in 2. The rest mass content of container 2 decreases but no energy enters or leaves container 2 so, SRT tell us, the inertial mass of container 2 is unchanged. If active gravitational mass were simply the sum of constituent rest masses then the gravitational pull of 2 on 1 would decrease while the gravitational pull of 1 on 2 would remain the same. The system would then begin to accelerate to the left although no external force is acting on it! Unless one is willing to accept this there is no alternative but to conclude that the active gravitational mass of an object is the same as its inertial mass and passive gravitational mass.

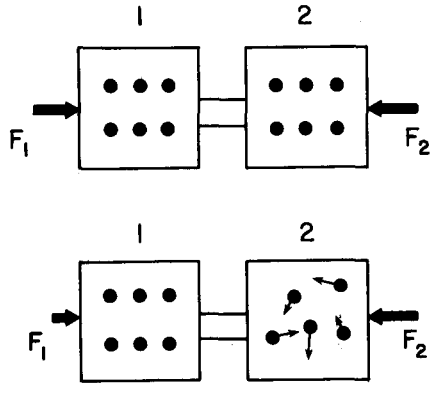

Fig. 1. Consequence of gravity generated only by rest mass. Initially the particles in both containers 1 and 2 are at rest. The force $\mathbf{F}_1$ on container 1, due to container 2 is balanced by $\mathbf{F}_2$. One of the particles in container 2 is then completely converted to kinetic energy of the remaining particles. According to the weak equivalence principle, the passive gravitational mass in the containers is unchanged so if the active gravitational mass of container 2 decreases, $\mathbf{F}_1$ will be less than $\mathbf{F}_2$ and the system will accelerate to the left.

What we are discovering here is an aspect of the universal character of gravity. All forms of energy contribute not only to the inertial and passive gravitational mass but also to the generation of the gravitational field. This tells us that Eq. (2.13) is indeed correct and that a tensor description is needed. This universality of gravity, however, raises also a disturbing question that proves the downfall of a "simple" tensor theory of gravity. The disturbing question is this: If all forms of energy are to be included on the right-hand side of Eq. (2.10)—or its tensor equivalent—how are we to include gravitational energy? Gravitational energy is certainly a real and not very subtle thing. Hydroelectric power plants would not conserve energy if there were no gravitational energy associated with water on the high side of a dam. If we do not include such energy as part of the total mass-energy of a gravitational source we will arrive at the same sort of unacceptable possibilities we encountered in our previous thought experiment. (Example: replace thermal energy in container 2 by gravitational binding energy and consider the consequences.) We *must* include gravitational energy as part of ρ. Let us not worry about the technical details of how we calculate gravitational energy from the gravitational field tensor but rather let us note the essential dilemma we face: We must know ρ to calculate the gravitational field; we need to know the gravitational field to calculate gravitational energy; but we must know the gravitational field energy to know ρ.

The mathematics of gravity, we see, is nonlinear since gravity itself is a source of gravity. No such difficulty arises, for example, in classical electrodynamics since the electromagnetic field is not itself a source of an electromagnetic field. Although the nonlinearity of gravity is a rather unwelcome complication it would not seem to present insurmountable difficulties. We might try, for example, to solve gravitational problems by an iterative procedure since gravitational energy is usually small compared to, say, rest mass energy. (Example: The rest mass energy of the sun in on the order of 10^{54} ergs while its gravitational binding energy is less than 10^{49} ergs.) In an iterative procedure we would ignore gravitational energy in ρ, solve for the gravitational field, compute the resulting gravitational energy, use it to find an improved estimate of ρ, solve for an improved estimate of the gravitational field, etc., etc.

Such a procedure can be and has been carried out[5] to all orders of iteration. A meaningfully detailed description of the results of this procedure would be out of place here, but the results are far too interesting not at least to sketch out: A gravitational field tensor is found from a highly nonlinear field equation. The gravitational field is not only generated by all forms of mass-energy but it affects all forms of mass-energy. In particular, any clock or meter stick, or any device for measuring spatial distances and time differences is necessarily and unavoidably affected by the gravitational field (even in the linearized approximation to this theory).

Any attempt to demonstrate by measurement that the space-time of events has the SRT geometry will fail if gravitational fields are present. The relationship of distances and time differences will be found by measurement to be more complicated than that of Eq. (1.6).

This theory is identical in content to Einstein's GRT but somewhat different in viewpoint. In the more common interpretation of GRT the SRT space-time geometry has no place. Since flat space-time has no direct relationship to measurement, it is considered irrelevant. The "real" geometrical structure of space-time is taken to be that measured by rods and clocks.

Einstein did not follow the above procedure of starting with a more or less standard field theory in the context of SRT. He started with space-time curvature as an *a priori* concept and so shall we.

B. Geometry

A space(time) is said to have a metric geometry if we have a way of computing a measure $(ds)^2$ of the separation of nearby points. If x and y are the coordinates in a two-dimensional space, a formula for $(ds)^2$ might be, for example,

$$(ds)^2 = (dx)^2 + (dy)^2 . \tag{2.14}$$

We immediately recognize this as flat two-dimensional space described in Cartesian coordinates, just as we recognize

$$(ds)^2 = (dx)^2 + (dy)^2 + (dz)^2$$

as three-dimensional flat space and

$$(ds)^2 = -c^2(dt)^2 + (dx)^2 + (dy)^2 + (dz)^2 \tag{2.15}$$

as the Minkowski space-time of SRT. Formulas such as these are called the line element or the metric formula.

The general formula for $(ds)^2$ is written (with the summation convention of Sec. I B) as

$$(ds)^2 = g_{\mu\nu}\, dx^\mu\, dx^\nu . \tag{2.16}$$

The "metric coefficients" $g_{\mu\nu}$ are in general not constants [as in Eqs. (2.14)–(2.16)] but are functions of position. We shall always assume that the metric is symmetric,[6] i.e., that $g_{\mu\nu} = g_{\nu\mu}$. Some examples of metric formulas, and of notation [e.g., suppressed parentheses on $(dx^\mu)^2$] are

Euclidean two-space in Cartesians
$$ds^2 = dx^2 + dy^2, \quad g_{xx} = g_{yy} = 1, \quad g_{xy} = 0; \tag{2.17a}$$

Euclidean two-space in polar coordinates
$$ds^2 = dr^2 + r^2\, d\phi^2, \quad g_{rr} = 1, \quad g_{\phi\phi} = r^2, \quad g_{r\phi} = 0; \tag{2.17b}$$

Surface of a sphere of radius a
$$ds^2 = a^2(d\theta^2 + \sin^2\theta\, d\phi^2), \quad g_{\theta\theta} = a^2, \quad g_{\phi\phi} = a^2\sin^2\theta, \quad g_{\theta\phi} = 0; \tag{2.17c}$$

SRT space-time Minkowski coordinates [see Eq. (16)]
$$ds^2 = -c^2\, dt^2 + dx^2 + dy^2 + dz^2, \quad g_{\mu\nu} = \eta_{\mu\nu}; \tag{2.17d}$$

Schwarzschild space-time with parameter r_g (see Sec. VI)
$$ds^2 = -(1 - r_g/r)c^2\, dt^2 + (1 - r_g/r)^{-1}\, dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2),$$
$$g_{00} = -(1 - r_g/r), \quad g_{rr} = (1 - r_g/r)^{-1}, \quad g_{\theta\theta} = r^2, \quad g_{\phi\phi} = r^2\sin^2\theta, \quad g_{\mu\nu} = 0 \ \text{ if } \mu \neq \nu . \tag{2.17e}$$

In addition to the set of metric coefficients $g_{\mu\nu}$, it is useful and traditional to define a second set of indexed quantities $g^{\mu\nu}$ by the requirement [cf. Eq. (1.5)]

$$g_{\mu\alpha}g^{\alpha\nu} = \delta^\nu_\mu \equiv \begin{cases} 1 & \text{if } \mu = \nu \\ 0 & \text{if } \mu \neq \nu. \end{cases} \tag{2.18}$$

A simple way of computing $g^{\mu\nu}$ is to consider the metric coefficients $g_{\mu\nu}$ arranged in a matrix. The quantities $g^{\mu\nu}$ than can be found, according to Eq. (2.18), as elements of the matrix inverse. If the metric is "diagonal" ($g_{\mu\nu} = 0$ if $\mu \neq \nu$), as are all the examples in Eq. (2.17), this is particularly simple. The result for flat two-space in polar coordinates is, for example,

$$g^{rr} = 1, \quad g^{\theta\theta} = r^{-2}, \quad g^{r\theta} = g^{\theta r} = 0 .$$

If we can find coordinates $(x, y, z, w, \ldots)$ in which the metric formula takes the simple form

$$ds^2 = \pm dx^2 \pm dy^2 \pm dz^2 \pm dw^2 \pm \cdots, \tag{2.19}$$

the space(time) and the coordinates are said to be "flat." The space(time) is said to be Euclidean if all the signs are plus (or all minus) and pseudo-Euclidean (or a space-time) if there are some pluses and some minuses. Flat coordinates are not unique. A transformation of coordinates of the type

$$x' = x\cos\alpha + y\sin\alpha, \quad y' = -x\sin\alpha + y\cos\alpha , \tag{2.20}$$

for example, changes Eq. (2.17a) to

$$dx^2 = dx'^2 + dy'^2,$$

so that both x, y and x', y' are flat coordinates. As another example, the transformations among different sets of Minkowski (flat) coordinates in space-time are the Lorentz transformations. Coordinates that are not flat are often called curvilinear. A space(time) is "curved" if no coordinates exist in which the metric formula has the form in Eq. (2.19); any coordinate system in a curved space(time) is curvilinear.

The reference to coordinate choice in the distinction between flat and curved space(time) raises the important question of what coordinates "mean." It may appear that we must specify what the coordinates "mean" before we write down a metric formula. (We know, for example, what spherical polar coordinates "mean" visually in ordinary three-dimensional space.) This is not true. Coordinates by themselves are only labels and have no geometric meaning until we know the distance structure associated with them. In the geometry of Eq. (2.17e), as an example, all points with the same r and t coordinate ($dr = dt = 0$) have the same distance structure as that of the surface of a sphere of radius r. We see then that the geometry must be spherically symmetric and that the r coordinate is related to the size of the spherical surfaces in the geometry. In this case, as in all cases, *the metric formula itself defines the geometric meaning of the coordinates.*

Very different looking metric formulas may correspond to the same space(time) described in different coordinates. This is the case if a change of coordinates transforms one metric formula into another. A familiar example is the transformation

$$r = (x^2 + y^2)^{1/2}, \quad \phi = \tan^{-1}(y/x),$$

which changes the metric formula of Eq. (2.17b) to that of Eq. (2.17a). A less familiar example is provided by the exotic appearing metric formula

$$ds^2 = -dv^2 + v^2\,du^2 + dy^2 + dz^2 . \tag{2.21}$$

New coordinates t,x may be introduced in place of v,u by the transformation

$$v = [(ct)^2 - x^2]^{1/2}, \quad u = \tanh^{-1}(x/ct).$$

When the expressions

$$dv = \frac{c^2 t\,dt - x\,dx}{[c^2t^2 - x^2]^{1/2}}, \quad du = \frac{ct\,dx - cx\,dt}{[c^2t^2 - x^2]^{1/2}}$$

are used in (Eq. 2.21) the result is

$$ds^2 = -c^2\,dt^2 + dx^2 + dy^2 + dz^2,$$

so that Eq. (2.21) is simply Minkowski space-time described in unusual curvilinear coordinates, just as Eq. (2.17b) is flat two-dimensional space in somewhat less unusual coordinates.

Clearly, coordinate transformations allow us considerable freedom to change the appearance of a metric formula. But they do not allow complete freedom. No matter how hard we search we will never find a change of coordinates that reduces Eq. (2.17c) to (2.17a). This is completely consistent with our intuition that the two-dimensional geometry—the relationship of distances—on the surface of a sphere is genuinely different from the flat two-dimensional geometry of a plane. The lesson to be learned from this is that the metric formula contains a confusing mixture of information both about the coordinates in which the formula is presented, and about the actual geometry it describes.

In Sec. IV we will develop the mathematics of curvature and will find, among other things, that we need not explicitly consider all possible coordinate transformations to conclude that Eq. (2.17c), for example, describes a curved space. For now let us just note that there must be quantitative measures of the curvature of a space(time). Suppose that Eq. (2.17c) represents the surface of the Earth (a = Earth radius, θ = colatitude, ϕ = longitude) and imagine the Earth to be a perfect sphere, bereft of all its interesting topographic features. An engineer designing a pipeline from Alaska to Los Angeles will make a terribly costly mistake if he ignores the curvature of the Earth's surface. On the other hand, a man laying out a parking lot can easily afford to forget the Earth's curvature. It is all a matter of scale. A small enough region of a curved geometry is essentially flat, but small compared to what? There must be a length scale—the Earth radius in our example—that determines the magnitude of curvature effects. The smaller this length scale is, the more the geometry is curved. We shall, for now, be very sloppy and call this length scale the "radius of curvature" although we shall see in Sec. IV that it is an oversimplification to characterize curvature by a single number.

Since, on a small scale, any geometry looks flat (except perhaps at a pathological point) it seems reasonable that we should be able to introduce coordinates that on a small scale are almost flat. Specifically, at any (nonpathological) point P we can choose coordinates in which

$$g_{\mu\nu} = \begin{cases} \pm 1 & \text{if } \mu = \nu \\ 0 & \text{if } \mu \neq \nu \end{cases} \text{ at } P \tag{2.22a}$$

and

$$\partial g_{\mu\nu}/\partial x^\alpha = 0 \quad \text{at } P. \tag{2.22b}$$

These coordinates will be called "locally flat," or LF, coordinates at point P. With LF coordinates at some point P of a curved space-time (e.g., Eq. [2.17e]) the metric coefficients at that point can be put in the form

$$g_{\mu\nu} = \eta_{\mu\nu} + O\left[(x^\alpha - x_P^\alpha)^2\right], \tag{2.23}$$

where $x^\alpha - x_P^\alpha$ is the coordinates separation from the chosen point P. The proportionality constants in the $O([x^\alpha - x_P^\alpha]^2)$ correction will be related to the radius of curvature of the space-time in the neighborhood of P.

As another, more specific, example we examine Eq. (2.17c) with point P taken to be $\theta = \pi/2$, $\phi = 0$. If we choose

$$x = a(\theta - \pi/2), \quad y = a\phi,$$

so that P is at $x = 0, y = 0$, the metric formula becomes

$$ds^2 = dx^2 + \cos^2(x/a)dy^2,$$

and therefore x, y are LF coordinates, since

$$ds^2 = dx^2 + dy^2 + O\left[(x/a)^2, (y/a)^2\right]. \tag{2.24}$$

LF coordinates at a point are not unique. We could, for example, perform the linear transformation of Eq. (2.20) for any rotational angle α and thereby get a whole class of coordinate systems satisfying Eq. (2.24). Similarly, in space-time a Lorentz transformation of a coordinate system that satisfies Eq. (2.22) produces another coordinate systems that also satisfies it. In flat space(time) such linear transformations (rotations, Lorentz) are the unique transformations that change one set of flat coordinates to another. Here, however, we are dealing with curved space(time)s and with the less stringent requirement that coordinates be *locally* flat. This means that nonlinear transformations do not ruin the LF property of a coordinate system if they affect the coordinates weakly enough near P. For example, the transformation

$$x = x' + (y')^3/a^2, \quad y = y' + (x')^2\,y'/a^2$$

defines new coordinates x', y' which also satisfy Eq. (2.24). In general, transformations of the type

$$x^\mu = x^{\mu'} + O\left[(x^{\nu'} - x_P^{\nu'})^3\right] \tag{2.25}$$

do not change the LF nature of coordinates.

The actual use of LF coordinates is almost always too tedius to be of use in calculations, but the concept of LF coordinates will be of crucial importance. They will be important to geometric considerations because we shall be interested in establishing certain "absolute" statements about geometric quantities. For such statements we shall know in advance that their truth or falsity is independent of coordinate system. We can then allow ourselves the convenience of proving, or disproving, these statements in the most convenient coordinates, usually LF coordinates. LF coordinates will be important to physical considerations in space-time because they will be the locally Minkowskian coordinates [Eq. (2.22)] corresponding to a local inertial reference frame.

The final point to be made here about geometry concerns a special class of curves called geodesics. In a flat space(time) there exist special coordinates, the flat coordinates. A straight line in a flat space(time) is simply a linear relationship amongst the flat coordinates, which can be put in the parameterized form

$$x^\mu = a^\mu \sigma + b^\mu , \tag{2.26}$$

where σ is the parameter along the curve and a^μ, b^μ are constants. In a curved space-time there is no preferred set of coordinates—no flat coordinates—in which Eq. (2.26) can be used. Rather, the concept of a straight line is replaced by that of a geodesic, a curve that is locally "as straight as possible." More specifically, let P be a point on a geodesic, and let x^μ be LF coordinates at P. In the neighborhood of P the geodesic can be put in the form

$$x^\mu = a^\mu \sigma + b^\mu + O\left[(\sigma - \sigma_P)^3\right] , \tag{2.27}$$

where $\sigma = \sigma_P$ at point P. The sloppiness in this equation due to the $O\left[(\sigma - \sigma_P)^3\right]$ term is unavoidable. If we demand that this term be zero in one LF coordinate system at P, an allowed transformation [Eq. (2.25)] will generate such a term in another LF system. Equation (2.27) then is the best we can do; it describes a curve that at point P is as straight as the coordinates are flat. A geodesic is, in this sense, as straight as possible at each of its points. In this sense, for example, it turns out that the meridians on the Earth (idealized as a perfect sphere) are geodesics, while the lines of latitude are not. See Fig. 2.

The development of the mathematical description of geodesics from this point of view requires a few new mathematical tools and symbols, and is put off until Sec. IV. For now we will have to make do with a less satisfactory conceptual starting point that turns out to be equivalent: geodesics, like straight lines in flat space(time), are the curves of extremal (minimum or maximum) length between two points.[7] Because this criterion for a geodesic is secondary the mathematical development here will only be sketched.

For definiteness (to eliminate ambiguities of sign and symbols) let us focus on the case of timelike curves between two points, P_A and P_B, in a curved space-time. Such a curve can be specified by giving the functions $x^\mu(\lambda)$, where λ is a parameter that runs from 0 at P_A to 1 at P_B. Along this curve proper time increases according to

$$d\tau = (-c^{-2} g_{\mu\nu}\, dx^\mu\, dx^\nu)^{1/2} = \left(-c^{-2} g_{\mu\nu} \frac{dx^\mu}{d\lambda}\frac{dx^\nu}{d\lambda}\right)^{1/2} d\lambda$$

$$\equiv F\, d\lambda , \tag{2.28}$$

and the total proper time is

$$\tau = \int_0^1 F\, d\lambda . \tag{2.29}$$

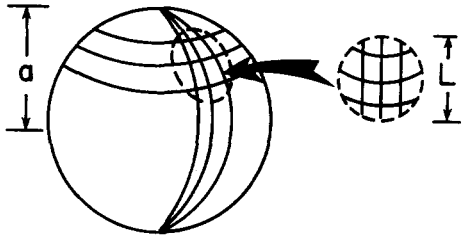

Fig. 2. Geodesic and nongeodesic curves on the surface of the Earth (assumed to be a perfect sphere). The meridians on the Earth are geodesics but the lines of latitude, except for the equator, are not. When a small region of the Earth's surface is examined it appears essentially flat. If this region is plotted on a graph made with locally flat coordinates (see text) the meridians appear almost straight. If the Earth radius is a, and the region is of size $L \ll a$, the tangent vectors to the meridians bend by an angle of order $(L/a)^2$ while the tangent vectors to the lines of latitude bend by a much larger angle, of order (L/a).

Since the proper time for a timelike curve is what we mean by its length, the integral in Eq. (2.29) must be extremized. With F considered as a functional of the functions $x^\mu(\lambda)$ and the derivatives $dx^\mu/d\lambda$, the necessary condition for Eq. (2.29) to be extremal is found by the usual Euler–Lagrange variation, familiar from Lagrangian mechanics,

$$\frac{d}{d\lambda}\left(\frac{\partial F}{\partial(dx^\alpha/d\lambda)}\right) = \frac{\partial F}{\partial x^\alpha} .$$

The result is a set of differential equations for the functions $x^\mu(\lambda)$:

$$g_{\mu\alpha}\frac{d^2x^\mu}{d\lambda^2} + \frac{1}{2}\frac{dx^\nu}{d\lambda}\frac{dx^\beta}{d\lambda}\left(\frac{\partial g_{\beta\alpha}}{\partial x^\nu} + \frac{\partial g_{\nu\alpha}}{\partial x^\beta} - \frac{\partial g_{\beta\nu}}{\partial x^\alpha}\right)$$

$$= g_{\mu\alpha}\frac{dx^\mu}{d\lambda} F^{-1}\frac{dF}{d\lambda} . \tag{2.30}$$

The curve $x^\mu(\lambda)$ can be parametrized in many different ways, that is, λ can be taken to be any parameter along the curve. The simplest choice for λ is simply τ, proper time along the curve. Since $F = d\tau/d\lambda$ in general, $F = 1$ for this choice and the right-hand side of Eq. (2.30) vanishes leaving us with

$$g_{\mu\alpha}\frac{d^2x^\mu}{d\tau^2} + \frac{1}{2}\frac{dx^\nu}{d\tau}\frac{dx^\beta}{d\tau}\left(\frac{\partial g_{\beta\alpha}}{\partial x^\nu} + \frac{\partial g_{\nu\alpha}}{\partial x^\beta} - \frac{\partial g_{\beta\nu}}{\partial x^\alpha}\right) = 0 , \tag{2.31}$$

which is the equation for a timelike goedesic, in its more-or-less common form. If $d\tau$ is replaced by ds, Eq. (2.31) is the equation for a spacelike geodesic parameterized with arc length. Equation (2.31) also describes null (i.e., lightlike) geodesics if τ is replaced by a parameter appropriate to a null curve.

Notice that Eq. (2.31) is, at least, compatible with our local viewpoint. In an LF coordinate system at point P the partial derivatives of the metric coefficients vanish at P, so Eq. (2.31) just says

$$\frac{d^2x^\alpha}{d\tau^2} = 0 ,$$

that is, the curve is locally straight in LF coordinates. That, of course, does not tell us why the partial derivatives of $g_{\mu\nu}$ must be arranged in the manner of Eq. (2.31) to describe a geodesic in non-LF coordinates.

C. Gravity and geometry

We come now to an important question: Why should we suspect *a priori* that gravity has anything to do with geometry? The guiding principle here is the weak equivalence principle. Let a point particle be "free" in the sense that the only influence on its motion is gravity. (We can say the particle is force-free since we shall not view gravity as a force.) At some moment—call it $t = 0$—we can specify the motion of the particle, for example, by specifying its velocity and position in some inertial reference frame. We are told by the weak equivalence principle that the subsequent motion of the particle is independent of the nature (mass, charge, etc.) of the particle. In other words, *the dynamics of the free particle is completely specified by a single worldline.* In Fig. 3(a) this is heuristically shown, on a space-time diagram, for a gravitational acceleration in the $+x$ direction. Figure 3(b) shows the very different situation for charged particles in an electric field. For electromagnetic interactions there is no weak equivalence principle; particles of

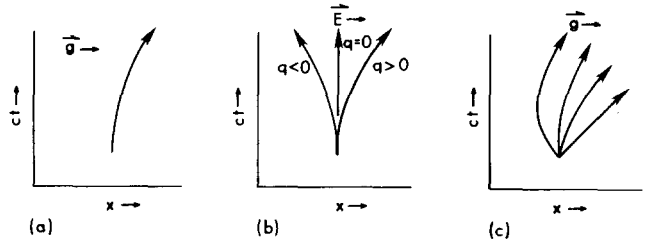

Fig. 3. Particle dynamics described by worldlines. In (a) a gravitational acceleration acts to the right and a single worldline suffices to show the effect on any particle starting from rest. In (b) an electric vector points to the right; the worldline of a particle starting from rest depends on the particle's electric charge q. A set of worldlines for different initial velocities summarizes, in (c), the dynamical effects of a gravitational acceleration acting to the right.

different charge follow different worldlines. The complete description of the dynamical effects of gravity at a point in space-time may be considered to be summarized by all possible free-particle worldlines, that is, by the worldlines corresponding to all initial velocities at the point [Fig. 3(c)].

This is an interesting and suggestive graphical device for picturing gravity, but it is not a geometrical one unless we can find some geometrical way of determining the special world lines. In the gravity-free case of Minkowski space-time we already know the answer: The "special" worldlines are the straight timelike worldlines; unaccelerated particles move with constant velocity. To find the geometrical basis of these curves in the presence of gravity we return yet again to the weak equivalence principle. According to this principle it would seem that we can *locally* banish gravity from physics by using a freely falling reference frame. The standard example of this is a freely falling elevator. Since the elevator and all the objects in it experience the same gravitational acceleration they all fall together without relative acceleration. An absent-minded physicist trapped in such a falling elevator would observe all objects in the elevator moving at a constant velocity with respect to the elevator, and might conclude that he is in outer space, away from all gravitational influences. For this physicist the freely falling frame of the elevator would seem to be a gravity-free inertial reference frame. A change of reference frame has eliminated gravity. Gravity must therefore not be a real force but a *pseudoforce* like centrifugal force, an apparent force that arises in a noninertial reference frame. Centrifugal forces appear in rotating frames and gravity forces appear in frames that are not freely falling.

Indeed it is true that the most common aspect of gravity, the weight force, vanishes in a freely falling frame and we *will* consider weight to be a pseudoforce, not a real gravitational force.[8] But it is not true that *all* effects of gravity disappear in the freely falling frame. If the trapped physicist is painstaking enough he will, at least in principle, be able to detect some gravitational effects. He will find, for example, that there is a miniscule relative acceleration (typical magnitude $\sim 4\times 10^{-4}$ cm/sec^2) between his eyeglasses, floating near the top of the elevator compartment, and his pen, near the bottom. In the standard viewpoint, based on a reference frame fixed with respect to the Earth, this is to be expected because the pen, slightly closer to the center of the Earth, experiences a greater gravitational acceleration than the eyeglasses. Because the gravitational field of the Earth is not uniform, gravity cannot be completely banished; there are then real gravitational forces (in contrast to the weight pseudoforce). These real aspects of gravity are often called tidal forces since it is the nonuniformity of the gravitational fields of spherical bodies (the sun and the moon) that drives the tides.

The effects produced by real (i.e., tidal) gravitational forces in the falling elevator will depend on the size of the elevator; the bigger the elevator is the larger will be the relative acceleration of objects at distant corners. If the elevator is small enough these effects may be judged to be unimportant and the elevator may be considered a good approximation to a gravity-free inertial frame. We shall in general call such a frame a "locally inertial frame."

In the locally inertial frame gravity-free SRT only approximately applies. We can try to construct flat coordinates x^μ for this frame with the usual construction techniques (e.g., perfect rigid rods, perfect clocks, light beams, etc.) but we know that the result will fall short of perfection. We cannot construct coordinates x^μ in which *all* free-particle worldlines have the straight-line form of Eq. (2.26). If we could, there would be no relative accelerations of any particles, and no effects whatever of gravity. The best we can do is to construct coordinates in which all free particles have no acceleration at a single point, call it P. For these coordinates it is fairly clear that all worldlines will have the form of Eq. (2.27), since this requires only that $d^2x^\mu/d\sigma^2 = 0$ at P.

The implications of the above considerations should be clear. Space-time is to be considered as curved. Small regions of space-time described in LF coordinates appear locally flat, so these LF coordinates must correspond to the locally inertial freely falling frames. Free particles follow worldlines that are locally straight, in the sense of Eq. (2.27), in these coordinates, hence free-particle worldlines must be geodesics. In its most succinct form the union of geometry and gravity is *Space-time is curved; free particles follow geodesics.*

We now have, at least conceptually, half of the structure of GRT. We have replaced the Newtonian "force" equation

$$\mathbf{F} = -m\nabla\Phi$$

by the dynamical principle

$$\text{force-free worldline} = \text{geodesic.} \tag{2.32}$$

Even with this much of the theory some answers are clarified. It is often asked, for example, whether the bending of the path of starlight passing near the sun makes a statement about light having an effective mass. In our picture of gravity, and in GRT, the question is irrelevant; light simply follows a null geodesic and in the neighborhood of the sun geodesics, null and otherwise, are bent. To answer other questions, however, we need the other half of the theory: What determines space-time curvature?

In the Newtonian theory the "source equation" relating the gravitational field to its sources is simply

$$\nabla^2\Phi = 4\pi G\rho\,. \tag{2.33}$$

For our geometric gravity theory we would want to replace this by something like

$$\text{curvature} = \text{source of gravity (mass-energy density, etc.)}\,. \tag{2.34}$$

The discovery of a mathematical representation of the left-hand side of Eq. (2.34) will be the task of Sec. IV, but some

preliminary observations can already be made. Just as Eq. (2.33) tells us that $\nabla^2\Phi$ vanishes outside a gravitating source, Eq. (2.34) seems to require that curvature vanish outside a source. This would imply, for example, that there is no gravitational influence near the Earth due to the sun, a conclusion in striking disagreement with observation. In Newtonian theory this problem is avoided in that there are two "measures" of gravity: $\nabla^2\Phi$, which couples to the source of gravity, and $\nabla\Phi$, which describes the effect of gravity. Clearly $\nabla\Phi$ needn't vanish everywhere that $\nabla^2\Phi$ vanishes, so that gravitational effects extend outside of gravitating sources. Similarly in a space-time description of gravity there must be two measures of curvature. The analog of $\nabla\Phi$, the measure of curvature appropriate to describing the effects of gravity, is called Riemann curvature. Just as $\nabla\Phi = 0$ means there is no gravity (Φ = constant) the vanishing of Riemann curvature guarantees that there are no gravitational effects and that there is no curvature, i.e., space-time is flat. The analog of $\nabla^2\Phi$, the curvature we want for the left-hand side of Eq. (2.34), is called Ricci curvature. Most important, the Ricci curvature can vanish without the vanishing of the Riemann curvature, so there *can* be gravity outside gravitating sources.[9] The structure of relativistic gravitation theories then can be summarized as

$$\begin{pmatrix}\text{dynamics}\\ \text{(worldlines of}\\ \text{particles, photons)}\end{pmatrix} \text{ governed by } \begin{matrix}\text{Riemann}\\ \text{curvature}\end{matrix}, \tag{2.35}$$

$$\text{Ricci curvature} = \text{gravitational sources}. \tag{2.36}$$

These equations apply not only to Einstein's gravitation theory, general relativity, but also to most alternative theories—other so-called metric theories of gravity. The differences among the theories lie in the details of Eq. (2.36). Einstein's theory, GRT, is in a sense the simplest of these theories since the sources of the Ricci curvature involve only *nongravitational* energy density, momentum flux, etc. In most alternative theories, extra nongeometric "gravitational fields" are proposed that have no effect in Eq. (2.35) but that enter into the right-hand side of Eq. (2.36).

It remains only to develop the mathematics necessary to find the quantitative meaning of the words in Eqs. (2.35) and (2.36).

III. VECTORS AND TENSORS IN CURVED SPACE-TIME

A. Vectors and bases

A vector in flat space(time) offers no conceptual challenge. It is simply a displacement, a directed straight line segment from one point to another. In a curved space or space-time such a concept has no clear meaning since there are no straight lines. We can salvage, however, the concept for a differential displacement **ds**, the displacement between two *infinitesimally* separated points. Heuristically, such a differential displacement is too small to care that the geometry is curved. (See Fig. 4.) Now if **ds** is a vector at some point we can do with it all the things we usually do with vectors in linear algebra. At a given point, in an N-dimensional space(time) differential displacement vectors form an N-dimensional vector space. This is, as usual, particularly clear from the locally flat viewpoint. *There is in fact nothing about the vector algebra at a point that depends on whether the space(time) is curved or flat.* We can thus add

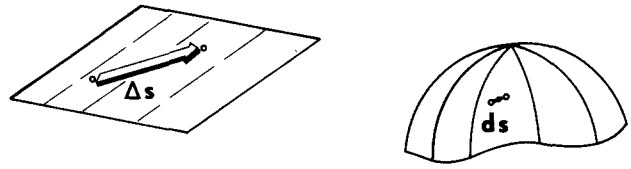

Fig. 4. Finite displacement as a vector in a two-dimensional flat space, and an infinitesimal displacement as a vector in a two-dimensional curved space.

these vectors, multiply them by scalars, etc. We can in fact multiply them by infinitely large scalars and thus create finite vectors.

For an important example we can start with the three-dimensional flat space of classical physics. If a particle undergoes a displacement $d\mathbf{s}$ in time dt we define its velocity vector to be $\mathbf{v} \equiv d\mathbf{s}/dt$. In space-time, curved or flat, the analogous vector is the four-velocity

$$\mathbf{U} \equiv d\mathbf{s}/d\tau\,. \tag{3.1}$$

Here $d\tau$ is the proper time the particle requires for displacement **ds**, and is a scalar. We can multiply now by another scalar m, the particle rest mass, to form another important vector, the particle's four-momentum

$$\mathbf{p} \equiv m\mathbf{U}\,. \tag{3.2}$$

Locally these vectors have their usual SRT meaning. Specifically, the vector **p**, for example, when viewed in a locally Minkowskian system, corresponding to a freely falling frame, is precisely the four-momentum that would be assigned to the particle by an observer in the freely falling frame, in blissful, but appropriate, disregard of gravity. (The technical procedure for computations based on this idea is simple and will be given below.) For a photon, in fact, Eq. (3.2) is inapplicable and it is best to start with the above point of view for the definition of four-momentum. That is, the four-momentum of the photon is the vector that in a locally inertial frame is the usual SRT four-momentum.

As in SRT and in classical mechanics the importance of vectors is based on the importance of **ds**, the displacement. Displacements are the starting point of kinematics, and hence of dynamics, and vectors such as velocity and momentum necessarily arise in a description of motion. Laws of motion—such as the familiar $\mathbf{v} = \mathbf{v}_0 + \frac{1}{2}\mathbf{a}t^2$ for motion with constant acceleration— are necessarily vector laws.

An important operation is the inner or "dot" product, a product of two vectors **A** and **B**, linear in each vector, which produces a scalar **A·B**. Since there is nothing about the algebra of vectors at a point in space-time that depends on curvature we already know in principle, at least, how to compute inner products in curved space-time: we can go to a locally Minkowskian coordinate system and compute **A·B** as in SRT. It must therefore be true that

$$\mathbf{U}\cdot\mathbf{U} = -c^2, \quad \mathbf{p}\cdot\mathbf{p} = -m^2c^2\,, \tag{3.3}$$

since it is true in SRT. As in SRT we call a vector spacelike if $\mathbf{A}\cdot\mathbf{A} > 0$, timelike if $\mathbf{A}\cdot\mathbf{A} < 0$, and null (or lightlike) if $\mathbf{A}\cdot\mathbf{A} = 0$.

Vectors are inherently geometric objects, but computations often require working with the components of vectors. As is familiar from linear algebra and classical physics, a vector basis is a complete set of vectors [N linearly independent vectors for an N-dimensional space(time)] at a

point. We shall denote a basis by $\mathbf{e}_\mu$. (Two points are worth emphasizing here. First, vectors and basis vectors—so far—have meaning only at given point; the relation of basis vectors at different points will be taken up in Sec. III C. Second, there is no necessary connection between basis vectors and components, on the one hand, and coordinates, on the other.) With a basis $\mathbf{e}_\mu$, a vector $\mathbf{W}$ can be written as

$$\mathbf{W} = W^\mu \mathbf{e}_\mu \tag{3.4}$$

(summation convection!). The numbers W^μ are called the components of $\mathbf{W}$ in the basis $\mathbf{e}_\mu$. It is common, and only mildly misrepresentative, to refer to the set of components W^μ as "the vector" and to a component relation (really the components of a vector equation), such as

$$v^\mu = v_0{}^\mu + \tfrac{1}{2}a^\mu t^2 ,$$

as a "vector equation."

It is important to keep in mind the arbitrariness of basis vectors. They can be chosen in any way that is convenient for the purpose at hand. The components of a vector depend, of course, on the basis vectors used. If the basis vectors are changed the components W^μ of a vector must change in such a way that the sum $W^\mu \mathbf{e}_\mu$ does not change.

A simple example can be given in Minkowski space-time. If the Minkowski coordinates are t,x,y,z then the most convenient basis at any point is the basis $\mathbf{e}_0,\mathbf{e}_x,\mathbf{e}_y,\mathbf{e}_z$ constructed as follows: The basis vector $\mathbf{e}_0$ is taken to be purely timelike, i.e., it is proportional to a displacement $\mathbf{ds}$ for which $dt \neq 0$ but $dx = dy = dz = 0$. In the same manner $\mathbf{e}_x$ corresponds to a displacement purely in the x direction, etc. The magnitudes of the basis vectors are taken so that the spatial basis vectors are of unit length

$$\mathbf{e}_x \cdot \mathbf{e}_x = \mathbf{e}_y \cdot \mathbf{e}_y = \mathbf{e}_z \cdot \mathbf{e}_z = 1$$

and the timelike basis vector satisfies

$$\mathbf{e}_0 \cdot \mathbf{e}_0 = -1 .$$

Since these basis vectors are fairly obviously mutually orthogonal all the inner products can be summarized by

$$\mathbf{e}_\mu \cdot \mathbf{e}_\nu = \eta_{\mu\nu} \tag{3.5}$$

and the set of basis vectors is called an "orthonormal tetrad." With these basis vectors a general displacement $\mathbf{ds}$ corresponding to coordinate changes dt,dx,dy,dz is

$$\mathbf{ds} = c\,dt\,\mathbf{e}_0 + dx\,\mathbf{e}_x + dy\,\mathbf{e}_y + dz\,\mathbf{e}_z = dx^\mu\,\mathbf{e}_\mu \tag{3.6}$$

and the components of the four-velocity are

$$U^\mu = \frac{dx^\mu}{d\tau} \tag{3.7}$$

and so forth.

For our simple example let us introduce new basis vectors $\mathbf{e}_{0'},\mathbf{e}_{x'},\mathbf{e}_{y'},\mathbf{e}_{z'}$ at some point (or at every point) according to

$$\mathbf{e}_{x'} = \gamma(\mathbf{e}_x + \beta\mathbf{e}_0), \quad \mathbf{e}_{y'} = \mathbf{e}_y ,$$
$$\mathbf{e}_{0'} = \gamma(\mathbf{e}_0 + \beta\mathbf{e}_x), \quad \mathbf{e}_{z'} = \mathbf{e}_z , \tag{3.8}$$

where β is some number smaller than unity, and $\gamma \equiv (1-\beta^2)^{-1/2}$. A nice feature of the new basis system is that it too is an orthonormal tetrad since (as is easily checked)

$$\mathbf{e}_{\mu'} \cdot \mathbf{e}_{\nu'} = \eta_{\mu\nu} .$$

To find the new components $U^{\mu'}$ of the four-velocity, as an example, we must find components for which

$$U^{\mu'}\mathbf{e}_{\mu'} = U^\mu \mathbf{e}_\mu = \mathbf{U} .$$

It is easily verified that the unique solution is

$$U^{x'} = \gamma(U^x - \beta U^0), \quad U^{y'} = U^y ,$$
$$U^{0'} = \gamma(U^0 - \beta U^x), \quad U^{z'} = U^z , \tag{3.9}$$

which is immediately recognized as a Lorentz transformation. The transformation in Eq. (3.8) must therefore be the corresponding Lorentz transformation of basis vectors. In general, any transformation from one orthonormal tetrad, in Minkowski space-time, to another is a Lorentz transformation of basis vectors. Note also that any orthonormal tetrad in Minkowski space-time can be associated with a Minkowski coordinate system.

Lorentz transformations of orthonormal tetrads and vector components are just one type of transformation. More generally at a point in N-dimensional space(time) any transformation is allowed in which N linearly dependent new basis vectors $\mathbf{e}_{\mu'}$ are given as linear combinations of the old basis vectors $\mathbf{e}_\mu$. The technical details of transformation of basis vectors and the reciprocal transformations of components are mostly bookkeeping problems, based on the single guiding principle $v^{\mu'}\mathbf{e}_{\mu'} = v^\mu \mathbf{e}_\mu$ for any vector. We shall manage to avoid such bookkeeping in this article and shall not dwell on the details here. One point, though obvious, is important enough to deserve a remark: If a component equation, e.g., $v^\mu = v_0^\mu + \frac{1}{2}a^\mu t^2$, is true in one basis then if the components are transformed to a new basis, the resulting component equation, e.g., $v^{\mu'} = v_0^{\mu'} + \frac{1}{2}a^{\mu'}t^2$, will also be true.

There are two special and important types of basis systems in curved space-time that are used for very different purposes. The first type of system helps to answer a question like "What energy, momentum, etc. does some observer measure?" There is no problem with such a question in SRT even if the observer is accelerating. At any point on the observer's world line we take his measurements of vector components to be the same as those in the inertial frame with which he is momentarily comoving. His basis vectors then are simply the orthonormal tetrad of that inertial frame. In curved space-time gravity prohibits the existence of inertial frames, but gravity should have no influence on the local laboratory measurements an observer would make of momentum, energy, etc. We take the observer's orthonormal tetrad, therefore, to be that of the *locally* inertial frame (i.e., the freely falling frame) with which he is momentarily comoving.

We denote these basis vectors as $\mathbf{e}_{\hat{0}},\mathbf{e}_{\hat{x}},\mathbf{e}_{\hat{y}},\mathbf{e}_{\hat{z}}$ (with the carets reminding us that the basis is orthonormal). For an observer with four-velocity $\mathbf{U}_{\text{obs}}$ the basis vectors are easily constructed when it is noted that $\mathbf{U}_{\text{obs}}$ can have no spatial components ($U^{\hat{x}}_{\text{obs}}$, etc.) in this system; such components would indicate that the observer has a nonzero velocity with respect to the locally inertial frame with which he is momentarily comoving! This implies that $\mathbf{U}_{\text{obs}} \propto \mathbf{e}_{\hat{0}}$ and, in fact that

$$\mathbf{e}_{\hat{0}} = c^{-1}\mathbf{U}_{\text{obs}} . \tag{3.10}$$

The remaining vectors $\mathbf{e}_{\hat{x}},\mathbf{e}_{\hat{y}},\mathbf{e}_{\hat{z}}$ can be chosen to be any three convenient mutually orthogonal spatial unit vectors, each of which is orthogonal to $\mathbf{e}_{\hat{0}}$. They are not unique; any rotation of one set produces another equally good set.

As an example of the application of this system we come back to the question "What energy does the observer mea-

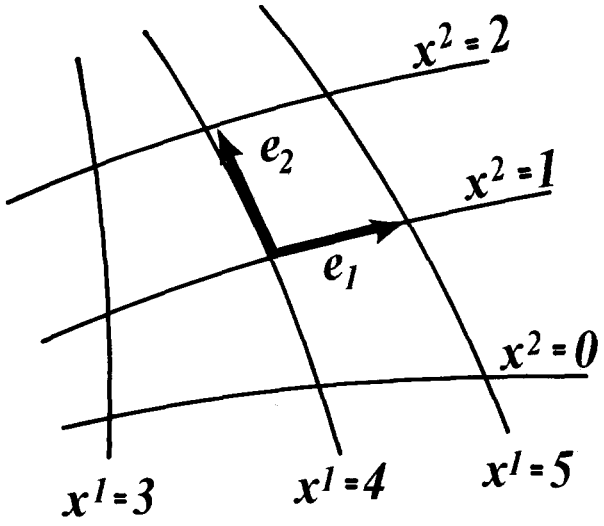

Fig. 5. Coordinate basis vectors for two-dimensional curvilinear coordinates. At the point $x^1 = 4$ and $x^2 = 1$ the $\mathbf{e}_1$ basis vector is tangent to the x^1 coordinate curve (the curve on which x^2 is constant). The arrows representing $\mathbf{e}_1$ and $\mathbf{e}_2$ should be considered suggestive only; finite vectors cannot be drawn as arrows except in a flat space.

sure for a particle, or photon?" Let the particle four-momentum be $\mathbf{p}$ so that in the observer's system it can be written

$$\mathbf{p} = p^{\hat{0}}\mathbf{e}_{\hat{0}} + p^{\hat{x}}\mathbf{e}_{\hat{x}} + p^{\hat{y}}\mathbf{e}_{\hat{y}} + p^{\hat{z}}\mathbf{e}_{\hat{z}} .$$

Paying no heed to gravity or to his own acceleration the observer, as in SRT, takes the energy to be $p^{\hat{0}}c$ or

$$\text{energy} = -c\mathbf{e}_{\hat{0}}\cdot\mathbf{p} ,$$

which is often most conveniently expressed as

$$\text{energy} = -\mathbf{p}\cdot\mathbf{U}_{\text{obs}} . \tag{3.11}$$

It has been emphasized that there is no *necessary* relationship between the choice of basis vectors at a point and the coordinate system for the space(time). The second type of special basis system, however, defines at every point in the space(time) a set of basis vectors, "coordinate basis vectors", uniquely *defined by the coordinate system*. Very loosely speaking the coordinate basis vector $\mathbf{e}_2$, for example, is the displacement generated by taking x^2 to increase by $\Delta x^2 = 1$ along the x^2 coordinate line. (See Fig. 5.) More precisely, the coordinate basis system $\mathbf{e}_\mu$ is the system in which a set of infinitesimal coordinate increments dx^μ gives a displacement

$$\mathbf{ds} = dx^\mu \mathbf{e}_\mu . \tag{3.12}$$

In words: the components, in a coordinate basis system, of the displacement vector are the coordinate changes. The simplest example is the orthonormal system in Minkowski space [see Eq. (3.6)].

The distance, or interval, assigned to the displacement $\mathbf{ds}$ must agree with the geometric structure of space(time) as dictated by the metric formula. This, and the linearity of the inner product, leads to an important fact about coordinate basis systems:

$$\mathbf{ds}\cdot\mathbf{ds} = (dx^\mu \mathbf{e}_\mu)\cdot(dx^\nu \mathbf{e}_\nu) = (\mathbf{e}_\mu\cdot\mathbf{e}_\nu)dx^\mu\, dx^\nu . \tag{3.13}$$

But

$$\mathbf{ds}\cdot\mathbf{ds} = (ds)^2 = g_{\mu\nu}\, dx^\mu\, dx^\nu ,$$

so the inner products of the coordinate basis vectors are related to the metric coefficients by

$$\mathbf{e}_\mu\cdot\mathbf{e}_\nu = g_{\mu\nu} . \tag{3.14}$$

This shows, incidentally, that the coordinate basis vectors are orthonormal if and only if the coordinates are flat.

We end this section by introducing a very useful notational convenience for coordinate basis systems. We characterize a vector $\mathbf{A}$ not only with its usual components A^μ ("contravariant components") but also with a second set A_μ ("covariant components") defined by

$$A_\mu \equiv g_{\mu\alpha}A^\alpha . \tag{3.15}$$

To find the contravariant components in terms of the covariant ones, we can turn Eq. (3.15) around by multiplying by $g^{\beta\mu}$ [see Eq. (2.18)] and summing on μ:

$$g^{\beta\mu}A_\mu = g^{\beta\mu}g_{\mu\alpha}A^\alpha = \delta^\beta{}_\alpha A_\alpha = A^\beta . \tag{3.16}$$

(The raising and lowering of indices with $g_{\alpha\beta}$ and $g^{\alpha\beta}$ will presently be generalized to noncoordinate basis systems.)

Coordinate basis vectors so greatly simplify most computations that they will be used almost exclusively hereafter.[10]

B. Tensors

The necessity of dealing with linear relationships among vectors at a point leads us unavoidably to mathematical objects beyond vectors. As an example, suppose that the vector $\mathbf{V}$ is a linear function of another vector $\mathbf{W}$. Then in any basis system the components V^μ must be linear in the components W^μ and we can write

$$V^\mu = T^\mu{}_\nu W^\nu . \tag{3.17}$$

The indexed symbol $T^\mu{}_\nu$ represents N^2 numbers [at a point in an N-dimensional space(time)] and these numbers should be regarded as defined by the relationship of $\mathbf{V}$ and $\mathbf{W}$. As another example take a scalar f to be a linear function of three vectors $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$. In any basis system we can write the relationship

$$f = S_{\alpha\beta\gamma}A^\alpha B^\beta C^\gamma , \tag{3.18}$$

thereby defining the symbol $S_{\alpha\beta\gamma}$, which represents N^3 numbers in general.

The above equations are called tensor equations as are all equations expressing linear relationships among vectors. The indexed symbols $T^\mu{}_\nu$ and $S_{\alpha\beta\gamma}$ are called tensor components and are summarized in an index free notation as the tensors $\mathbf{T}$ and $\mathbf{S}$. The rank of a tensor refers to the number of indices on its components so that, for example, $\mathbf{T}$ is said to be rank 2 and extended to include vectors as rank 1 tensors and scalars as rank 0. As a further example of tensor, rank, and notation we can have a fifth-rank tensor $\mathbf{M}$ defined by a linear relationship between a second-rank tensor $\mathbf{L}$ and three vectors

$$L_{\mu\nu} = M_{\alpha\mu\beta\nu\gamma}A^\alpha B^\beta C^\gamma . \tag{3.19}$$

A simple example of a tensor relationship in three-dimensional physics is the dependence, in classical electrostatics, of the electric polarization vector $\mathbf{P}$ in a dielectric medium on the electric field $\mathbf{E}$. This dependence is linear (approximately, for weak fields) and hence we can write

$$P^\mu = \chi^\mu{}_\nu E^\nu , \tag{3.20}$$

thereby defining the second-rank "susceptibility tensor" χ. In an isotropic dielectric the tensor nature of susceptibility can be avoided. In this case the susceptibility tensor takes

the form

$$\chi^{\mu}{}_{\nu} = \chi\delta^{\mu}{}_{\nu} \tag{3.21}$$

in any basis. Here δ as always is the Kronecker delta, and the number χ is called, simply, the susceptibility. With this type of susceptibility Eq. (3.20) becomes

$$P^{\mu} = \chi\delta^{\mu}{}_{\nu}E^{\nu} = \chi E^{\mu}, \tag{3.22}$$

showing that **P** and **E** are parallel, as must be the case in an isotropic dielectric. A mathematically similar example from classical mechanics is the second-rank inertia tensor, defined by the linear dependence of the angular momentum vector on the angular velocity vector.

The stress tensor is yet another, though less familiar, second-rank tensor of classical physics. Its importance to GRT justifies a few remarks about it here. Let **dA** be a vector representing a differential area element in a stressed medium. The vector is associated with the area in the usual way, i.e., $|\mathbf{dA}|$ is the magnitude of the area and the direction of **dA** is normal to the area pointing from (arbitrarily chosen) inside to outside. The force **dF** transmitted through the area, due to the material inside, acting on the material outside, can be shown to be not only proportional to the size of the area but in fact linearly dependent on the vector **dA**. A second-rank tensor **T**, the stress tensor, is then defined by

$$(dF)^{\mu} = T^{\mu}{}_{\nu}(dA)^{\nu}. \tag{3.23}$$

In an isotropic medium **dF** must be parallel to **dA** and [cf. Eq. (3.21)] the components of the stress tensor must be

$$T^{\mu}{}_{\nu} = p\delta^{\mu}{}_{\nu}, \tag{3.24}$$

where p is the pressure in the medium.

As with vector components, indices of tensor components can be raised or lowered. In a coordinate basis this is done with the metric coefficients $g_{\mu\nu}$ and $g^{\mu\nu}$, as in

$$T^{\mu\nu}{}_{\alpha} = g^{\mu\sigma}T_{\sigma}{}^{\nu}{}_{\alpha} = g_{\alpha\beta}T^{\mu\nu\beta}$$

and so forth. It is easily verified [with Eqs. (2.18), (3.15), and (3.16)] that, for example,

$$V^{\mu} = T^{\mu}{}_{\nu}W^{\nu}, \quad V^{\mu} = T^{\mu\nu}W_{\nu},$$

$$V_{\mu} = T_{\mu\nu}W^{\nu}, \quad V_{\mu} = T_{\mu}{}^{\nu}W_{\nu},$$

all represent the same tensor relationship. As with vectors a subscript is called a covariant index and a superscript a contravariant index. The numerical value of a tensor component will in general depend on whether indices are covariant or contravariant (e.g., $T_1{}^2$ need not equal T_{12}).

A tensor is said to be symmetric on two of its indices if the interchange of those two indices does not change the value of the component. For example, a third-rank tensor **Q** is symmetric on its first and third indices if

$$Q_{\alpha\beta\gamma} = Q_{\gamma\beta\alpha}$$

for all choices of α, β, and γ. The interchange of indices, of course, is considered with both indices covariant or both contravariant. For the three examples given of second-rank tensors in classical physics, the susceptibility tensor, the inertia tensor, and the stress tensor, it can be shown that each is symmetric. For each then, there are only six (rather than $N^2 = 3^2 = 9$) independent components in general.

A very important tensor in geometry is the "metric tensor" **g**, which can be thought of as defined by the dot product. If **A** and **B** are any two vectors then the dot product is a scalar, call is S, given by

$$S = \mathbf{A}\cdot\mathbf{B}.$$

Since this is a linear relationship a second-rank tensor **g** is defined by

$$S = \mathbf{A}\cdot\mathbf{B} = g_{\mu\nu}A^{\mu}B^{\nu}. \tag{3.25}$$

The components of **g** are easily found in any basis since

$$\mathbf{A}\cdot\mathbf{B} = (A^{\mu}\mathbf{e}_{\mu})\cdot(B^{\mu}\mathbf{e}_{\nu}) = \mathbf{e}_{\mu}\cdot\mathbf{e}_{\nu}A^{\mu}B^{\nu},$$

and hence

$$g_{\mu\nu} = \mathbf{e}_{\mu}\cdot\mathbf{e}_{\nu}. \tag{3.26}$$

We have already used the symbol $g_{\mu\nu}$ to represent the metric coefficients, i.e., the coefficients in the metric formula [Eq. (2.16)]. The present notation for **g** would be intolerable if it were not for the fact that in a coordinate basis the tensor components $g_{\mu\nu}$ are precisely the metric coefficients $g_{\mu\nu}$. [Compare Eqs. (3.14) and (3.26).] The metric coefficients can thus be considered a special case (coordinate basis) of the metric tensor components or the metric tensor can be considered to generalize the metric coefficients to noncoordinate basis systems. In any case the raising and lowering of component indices in a noncoordinate basis system is done with $g_{\mu\nu}$ in precisely the same manner as in coordinate basis systems.

So far tensors have been introduced in terms of indexed quantities that arise in the description of linear relations among vectors. This glib prescription underemphasizes the crucial point that the numbers represented by these indexed quantities depend on basis system. They relate, after all, vector components that depend on basis system. Tensor components must change from one basis to another in such a manner that they represent the same linear relationships. Any indexed quantity whose values do not transform in this way cannot represent a linear relationship of vectors; it cannot be a tensor. The importance of this point warrants an example: In two-dimensional flat space let $\mathbf{e}_x$ and $\mathbf{e}_y$ be orthonormal vectors and let them have their usual relationship that $\mathbf{e}_y$ is 90° counterclockwise from $\mathbf{e}_x$. Define the *symbol* $J^{\mu}{}_{\nu}$ by

$$J^{\mu}{}_{\nu} \equiv \begin{cases} 1 & \text{if } \mu \neq \nu \\ 0 & \text{if } \mu = \nu \end{cases} \tag{3.27}$$

in any basis, and ask whether this indexed quantity defines a tensor. If it does define a tensor then

$$V^{\mu} = J^{\mu}{}_{\nu}W^{\nu} \tag{3.28}$$

defines a relationship between vectors **V** and **W**. In the $\mathbf{e}_x$, $\mathbf{e}_y$ basis let $W^x = 1$, $W^y = 0$ so that, by Eq. (3.28), $V^y = 1$ and $V^x = 0$. In words: **V** has the same magnitude as **W** and is 90° counterclockwise from it. Consider $\mathbf{e}'_x = \mathbf{e}_y$ and $\mathbf{e}'_y = -\mathbf{e}_x$. This of course is simply a basis system rotated 90° from the original system. In this system **W** has components $W^{x'} = 0$, $W^{y'} = -1$ (so that $W^{\mu'}\mathbf{e}_{\mu'} = W^{\mu}\mathbf{e}_{\mu}$). But according to

$$V^{\mu'} = J^{\mu}{}_{\nu}W^{\nu'}$$

and Eq. (3.27), $V^{x'} = -1$ and $V^{y'} = 0$. In this case **V** is 90° *clockwise* of **W** so a different relation of **V** and **W** results in two basis systems and hence Eq. (3.27) does not define the components of a tensor.

We have avoided, in Sec. III A, giving the general scheme for transforming vector components with the excuse that the procedures were merely an issue of accounting. The details of the transformation of tensor components are no more profound and are even more tedious. The

details, therefore, will not be given here but the general principles will be repeated: The transformation of basis vectors forces a transformation of components of a vector in such a manner as to ensure that the same vector is represented in the old system and the new; the transformation of vector components forces a transformation of the components of a tensor in such a manner as to ensure that the component equations represent the same relationships of vectors in the old system and the new.

Without the explicit component transformation laws we shall not be able to give explicit proofs of a few needed facts about tensors. These facts, ranging from obvious to plausible, are simply stated here:

(i) If a tensor is "zero" (i.e., all its components vanish) in one basis, then it is zero in any basis.

(ii) If tensor indices are symmetric in one frame (e.g., $T_{\mu\nu} = T_{\nu\mu}$) then they are symmetric in any frame.

(iii) If **A** and **B** are vectors then $A_\mu B_\nu$ are the (covariant) components of a second-rank tensor. (Similarly $A^\mu B^\nu$ are the tensor's contravariant components and $A^\mu B_\nu$ and $A_\mu B^\nu$ are mixed components of that tensor.) More generally a tensor of rank $r_1 + r_2$ can be formed from tensors of rank r_1 and r_2, as in the fifth-rank tensor $T_\mu{}^\nu S_{\alpha\beta\gamma}$.

(iv) Summing ("contracting") over a covariant and contravariant index of a tensor of rank $r \geqslant 2$ produces a tensor of rank $r-2$. From a fifth-rank tensor **F**, for example, we can form a third-rank tensor **T** by contracting on the first and third indices. The covariant components of **T** would be

$$T_{\alpha\beta\gamma} = F_{\mu\alpha}{}^{\mu}{}_{\beta\gamma} .$$

We could also, for example, contract over the second and fifth indices

$$S_{\alpha\beta\gamma} = F_{\alpha\nu\beta\gamma}{}^{\nu} .$$

Here **S** and **T** will in general be different third-rank tensors.

Since it is far from obvious that contraction produces tensors, an example is appropriate. Let **D** be a second-rank tensor with components

$$D^{\mu\nu} = A^\mu B^\nu .$$

The claim now is that

$$D^\mu{}_\mu = A^\mu B_\mu$$

is a scalar (tensor of rank 0). But

$$A^\mu B_\mu = A^\mu g_{\mu\alpha} B^\alpha = \mathbf{A}\cdot\mathbf{B}$$

[cf. Eqs. (3.15) and (3.25)] so $D^\mu{}_\mu$ is indeed a scalar.

C. Covariant differentiation

So far we have focussed on tensors as mathematical objects at a single point. Now we shall consider also tensor fields, tensors defined at every point in some region of space(time). It will be worthwhile starting with the simplest example, a scalar field.

The mathematical description of a scalar is not particularly difficult. If x^μ is some coordinate system in N-dimensional space(time) then a scalar field Ψ is defined simply by specifying a single function $\Psi(x^\mu)$ of the N coordinate variables x^μ. Information about the rate of change of Ψ is contained in the partial derivatives

$$\Psi_{,\mu} \equiv \partial\Psi/\partial x^\mu . \tag{3.29}$$

The derivatives $\Psi_{,\mu}$ turn out to be the covariant components, in the coordinate basis system, of a vector which we call the gradient of Ψ and denote by **grad** Ψ. The claim then is that

$$(\mathbf{grad}\,\Psi)_\mu \equiv \Psi_{,\mu} \tag{3.30}$$

defines a vector. To prove this let **ds** be the displacement between two nearby points; in the coordinate basis the components of **ds**, are the coordinate differentials. The change $d\Psi$ in Ψ for this displacement is just

$$d\Psi = \frac{\partial\Psi}{\partial x^\mu} dx^\mu \equiv (\mathbf{grad}\,\Psi)_\mu\, dx^\mu . \tag{3.31}$$

Since $d\Psi$ is a scalar and dx^μ are vector components, **grad** Ψ must be a tensor (i.e., a vector).

Note the pattern: From a zero-rank tensor Ψ we have computed a first-rank tensor **grad** Ψ that contains the information about the rate of change. The same pattern applies to all tensors. Let **V** be a vector field, a vector defined at every point in some region. Across a displacement **ds** the change—call it **dV**—in the vector will itself be a vector. From the relationship of **dV** and **ds** we can define a second-rank tensor **grad V** by

$$(\mathbf{dV})^\mu = (\mathbf{grad}\,\mathbf{V})^\mu{}_\nu\, dx^\nu . \tag{3.32}$$

More generally the rate of change of a tensor of rank r is a new tensor, the gradient, of rank $r+1$. The process of finding tensors to represent rates of change of tensors is called "covariant differentiation."

If the components, in a coordinate basis, of the vector field **V** are the functions $V^\mu(x^\nu)$ it is natural to guess that

$$(\mathbf{grad}\,\mathbf{V})^\mu{}_\nu = \partial V^\mu/\partial x^\nu \equiv V^\mu{}_{,\nu} . \tag{3.33}$$

A simple example will show that this guess cannot be true in general. In two-dimensional flat space let **V** be a vector field with components in polar coordinates [Eq. (2.17b)] given by

$$V^\theta = 1, \quad V^r = 0 \tag{3.34}$$

at every point. The indexed symbol $V^\mu{}_{,\nu}$ then is zero for all values of μ and ν. But clearly the vector field is not constant since the "θ direction" is not a fixed direction and since the magnitude of **V**, found from

$$\mathbf{V}\cdot\mathbf{V} = g_{\mu\nu} V^\mu V^\nu = g_{\theta\theta}(1)^2 = r^2 ,$$

increases with distance from the origin. [See Eqs. (2.17b) and (3.25).]

How then are the components of **grad V** to be found? The above example suggests that difficulty arises because basis vectors are different from point to point. If we allow for this we have

$$\mathbf{dV} = d(V^\mu \mathbf{e}_\mu) = (dV^\mu)\mathbf{e}_\mu + V^\mu\, \mathbf{de}_\mu . \tag{3.35}$$

The first term on the right involves the change in the value of the component V^μ. For a displacement **ds** with (coordinate basis) components dx^μ this change is just

$$dV^\mu = V^\mu{}_{,\nu}\, dx^\nu . \tag{3.36}$$

The last term on the right involves the change in the coordinate basis vector $\mathbf{e}_\mu$, and is a bit more subtle. The change $\mathbf{de}_\mu$ is itself a vector and hence can be written as a sum of the basis vectors. Since $\mathbf{de}_\mu$ will be linear in **ds** we can write it as

$$\mathbf{de}_\mu = \Gamma^\alpha_{\mu\nu}\, dx^\nu\, \mathbf{e}_\alpha . \tag{3.37}$$

The coefficient symbol $\Gamma^\alpha_{\mu\nu}$ is called the Christoffel symbol[11]; the $\Gamma^\alpha_{\mu\nu}$ are *not* components of a third-rank tensor.[12] Combining the above results and changing dummy indices we find

$$\mathbf{dV} = V^{\mu}{}_{,\nu}\, dx^{\nu}\, \mathbf{e}_{\mu} + V^{\mu}\Gamma^{\alpha}_{\mu\nu}\, dx^{\nu}\, \mathbf{e}_{\alpha}$$
$$= (V^{\mu}{}_{,\nu} + V^{\sigma}\Gamma^{\mu}_{\sigma\nu}) dx^{\nu}\, \mathbf{e}_{\mu} .$$

The comparison with Eq. (3.32) gives us

$$(\mathbf{grad\ V})^{\mu}{}_{\nu} = V^{\mu}{}_{,\nu} + V^{\sigma}\Gamma^{\mu}_{\sigma\nu} . \tag{3.38}$$

Neither term on the right gives tensor components by itself, but together the two terms do give the components of the second-rank tensor (**grad V**).

The combination of symbols on the right of Eq. (3.38) is commonly[13] and conveniently abbreviated as $V^{\mu}{}_{;\nu}$ so that

$$(\mathbf{grad\ V})^{\mu}{}_{\nu} = V^{\mu}{}_{;\nu} . \tag{3.39}$$

The semicolon notation is used similarly for the components of the gradients of tensors of all types. For example, the covariant components of the gradient of a second-rank tensor **T** would be

$$(\mathbf{grad\ T})_{\alpha\beta\gamma} = T_{\alpha\beta;\gamma} . \tag{3.40}$$

The mixed components (some covariant indices, some contravariant) of the gradient a third-rank tensor **S** are written

$$(\mathbf{grad\ S})_{\alpha}{}^{\beta\gamma}{}_{\mu} = S_{\alpha}{}^{\beta\gamma}{}_{;\mu} \tag{3.41}$$

and so forth.

The prescription [Eq. (3.38)] has been given so far only for the computation of these gradient components in one case. In other cases (covariant components of **grad V**, components of higher rank gradients) there is no new difficulty, only the old difficulty of basis vectors which change from point to point. It should not be surprising then that the same Christoffel symbol appears in all cases to handle the rate of change of the basis vectors. The general prescription will be inherent in the following examples.

For a vector field **V** the covariant components $(\mathbf{grad\ V})_{\mu\nu} = V_{\mu;\nu}$ are calculated from

$$V_{\mu;\nu} = V_{\mu,\nu} - V_{\lambda}\Gamma^{\lambda}_{\mu\nu} \tag{3.42}$$

(see Appendix). Since $V^{\mu}{}_{;\nu}$ and $V_{\mu;\nu}$ are tensor components they could equally well be computed from

$$V_{\mu;\nu} = g_{\mu\alpha} V^{\alpha}{}_{;\nu} . \tag{3.43}$$

The components in Eq. (3.41) are given by

$$S_{\alpha}{}^{\beta\gamma}{}_{;\mu} = S_{\alpha}{}^{\beta\gamma}{}_{,\mu} - S_{\lambda}{}^{\beta\gamma}\Gamma^{\lambda}_{\alpha\mu} + S_{\alpha}{}^{\lambda\gamma}\Gamma^{\beta}_{\lambda\mu} + S_{\alpha}{}^{\beta\lambda}\Gamma^{\gamma}_{\lambda\mu} . \tag{3.44}$$

The pattern is a partial derivative term followed by one Γ term for each index of the tensor being differentiated. If the index is covariant the Γ is added with a negative sign; if the index is contravariant, with a plus sign.

We have up to now avoided a direct confrontation with the question: What is the meaning of a change in a vector (or tensor) from one point to another? How can we compare vectors at different points? This question of course seems superfluous in flat space(time). A vector there is the same at two points if it has the same direction and magnitude, i.e., if it has the same components in a flat coordinate basis, a system in which the basis vectors do not change. This gives the key to the answer in curved space(time): To find the change in **V** from point P_1 to the nearby point P_2 we set up a locally flat coordinate system at P_1 (or, equally well, at P_2). In this system we simply find the changes in the components V^{μ} and that tells us the vector **dV**. We can then transform the components of the vector **dV** to a useful coordinate basis. The conceptual meaning is then: Curvature does not enter into the differential change of a vector. The change is a local question and a local answer can be given.

This is a nice conceptual background for covariant differentiation, but not a practical algorithm for computing components of gradients. For that we need a way of computing the Christoffel symbols. It is shown in the Appendix that the needed formula is

$$\Gamma^{\alpha}_{\beta\gamma} = \tfrac{1}{2} g^{\alpha\lambda}(g_{\lambda\beta,\gamma} + g_{\lambda\gamma,\beta} - g_{\beta\gamma,\lambda}) . \tag{3.45}$$

It is well worth emphasizing that this formula, as well as the pattern of covariant differentiation given in Eq. (3.44), apply *only in a coordinate basis system*. Covariant differentiation can certainly be done in other basis systems, but the formulas are more complicated.

From Eq. (3.45) it follows that the Γ's are symmetric on the lower indices

$$\Gamma^{\alpha}_{\beta\gamma} = \Gamma^{\alpha}_{\gamma\beta} ,$$

so that in an N-dimensional space there are only $\frac{1}{2}N^2(N+1)$ independent Γ's. Perhaps the simplest example is given by polar coordinates in two-dimensional flat space [Eq. (2.17b)]. In the polar coordinate basis the Γ's are

$$\Gamma^{r}_{\theta\theta} = -r , \quad \Gamma^{r}_{rr} = 0 , \quad \Gamma^{\theta}_{\theta\theta} = 0 , \quad \Gamma^{\theta}_{rr} = 0 ,$$
$$\Gamma^{\theta}_{r\theta} = \Gamma^{\theta}_{\theta r} = r^{-1} , \quad \Gamma^{r}_{r\theta} = \Gamma^{r}_{\theta r} = 0 .$$

We are now in a position to find the gradient of the vector field example of Eq. (3.34). With Eq. (3.38) and the Γ's above we find that the only nonvanishing components of **grad V** are

$$V^{r}{}_{;\theta} = -r , \quad V^{\theta}{}_{;r} = r^{-1} .$$

The "locally flat viewpoint" facilitates the proof of several important features of covariant differentiation. For example, for any metric tensor

$$g_{\alpha\beta;\gamma} = 0 . \tag{3.46}$$

That is, the gradient of the metric tensor is always zero. The proof is simple when LF coordinates are introduced at a point. At this point $g_{\alpha\beta,\gamma} = 0$ [Eq. (2.22)]. But clearly there are no first-order changes in the basis vectors in a locally flat system, hence all Γ's are zero. We conclude that **grad g** is zero in an LF coordinate basis and hence in any basis. With similar arguments it is easy to show that in most instances the semicolon operates like the comma of partial differentiation:

$$(A^{\alpha}B^{\beta}{}_{\gamma})_{;\mu} = A^{\alpha}{}_{;\mu}B^{\beta}{}_{\gamma} + A^{\alpha}B^{\beta}{}_{\gamma;\mu} , \tag{3.47a}$$

$$(A_{\alpha\nu}B^{\nu\beta})_{;\mu} = A_{\alpha\nu;\mu}B^{\nu\beta} + A_{\alpha\nu}B^{\nu\beta}{}_{;\mu} , \tag{3.47b}$$

$$g^{\alpha\beta}{}_{;\mu} = 0 , \tag{3.47c}$$

$$V_{\alpha;\beta} = (g_{\alpha\gamma}V^{\gamma})_{;\beta} = g_{\alpha\gamma;\beta}V^{\gamma} + g_{\alpha\gamma}V^{\gamma}{}_{;\beta} = g_{\alpha\gamma}V^{\gamma}{}_{;\beta} . \tag{3.47d}$$

The second, third, etc., covariant derivatives of a tensor field are defined in the obvious way, e.g., for the vector field **V**,

$$[\mathbf{grad}(\mathbf{grad\ V})]_{\mu\alpha\beta} = (\mathbf{grad\ V})_{\mu\alpha;\beta} \equiv V_{\mu;\alpha;\beta} . \tag{3.48}$$

Arguments based on locally flat coordinates cannot be used to prove that

$$V_{\mu;\alpha;\beta} = V_{\mu;\beta;\alpha} \tag{3.49}$$

since it is not true in general! We will find in Sec. IV that the failure of this equation is closely related to a quantitative measure of curvature.

IV. GEODESICS, CURVATURE, AND GENERAL RELATIVITY

A. Geodesic equation

In Sec. II C gravity and geometry were united by the principle that free particles move on geodesics. Geodesics were defined as "locally straight" curves, but the equation for geodesics was derived in Sec. II B from a variational principle, not a "local" principle. The mathematics of Sec. III will now be exploited to do things right.

If a force **F** is acting on a particle, the particle's four-momentum will not be constant but will change according to

$$\frac{\mathbf{dp}}{d\tau} = \mathbf{F} . \tag{4.1}$$

Whether or not a force is acting is a local question. At a point P on a particle's worldline Eq. (4.1) can be viewed in a locally Minkowskian system. The rate of change of four-momentum prescribed by this equation is just the rate that would be observed in the corresponding freely falling frame. The weak equivalence principle tells us that in the freely falling frame no acceleration of the particle will be observed at P if only gravity is acting, i.e., **F** has no gravitational contributions. A free particle must therefore obey

$$\frac{\mathbf{dp}}{d\tau} = 0 . \tag{4.2}$$

The dynamical meaning of this equation is given to us by the weak equivalence principle. Now let us consider the geometric meaning. The four-momentum **p** is a tangent to the particle's worldline since, for displacement **ds** along the worldline,

$$\mathbf{p} = m\frac{\mathbf{ds}}{d\tau}$$

[see Eqs. (3.1) and (3.2)]. Equation (4.2) tells us that this tangent to the worldline does not change along the worldline. But this is precisely what we mean by a locally straight line: a curve along which the tangent to one segment is in the same direction as the tangent to the next! The geodesic equation for a timelike curve is then precisely Eq. (4.2).

To show the equivalence with the result in Sec. II B we analyze Eq. (4.2) in a coordinate basis system. For a displacement **ds** (components dx^ν) along the worldline, we know from Sec. III C that

$$(\mathbf{dp})^\mu = p^\mu{}_{;\nu}\, dx^\nu ,$$

so that the μ component of Eq. (4.2) is

$$\left(\frac{\mathbf{dp}}{d\tau}\right)^\mu = \frac{(\mathbf{dp})^\mu}{d\tau} = p^\mu{}_{;\nu}\frac{dx^\nu}{d\tau} .$$

But

$$\frac{dx^\nu}{d\tau} \equiv U^\nu = \frac{p^\nu}{m}$$

and the equation becomes

$$p^\mu{}_{;\nu}\, p^\nu = 0 . \tag{4.3}$$

We can equally well write this equation as

$$U^\mu{}_{;\nu}\, U^\nu = 0 \tag{4.4}$$

after dividing by m^2. (Our definition of "particle" tacitly includes the requirement that rest mass is constant.) From Eq. (4.4) and $U^\nu \equiv dx^\nu/d\tau$ we have

$$0 = (U^\mu{}_{,\nu} + U^\beta \Gamma^\mu_{\beta\nu})U^\nu = \left(\frac{\partial U^\mu}{\partial x^\nu} + U^\beta \Gamma^\mu_{\beta\nu}\right)\frac{dx^\nu}{d\tau}$$
$$= \frac{dU^\mu}{d\tau} + \frac{dx^\nu}{d\tau}\frac{dx^\beta}{d\tau}\Gamma^\mu_{\beta\nu} ,$$

and finally

$$\frac{d^2x^\mu}{d\tau^2} + \frac{1}{2}\frac{dx^\nu}{d\tau}\frac{dx^\beta}{d\tau} g^{\mu\gamma}(g_{\gamma\beta,\nu} + g_{\nu\gamma,\beta} - g_{\beta\nu,\gamma}) = 0 . \tag{4.5}$$

When this equation is multiplied by $g_{\mu\alpha}$ (and summed over μ) the result is identical to Eq. (2.31).

Equations (4.4) and (4.5) cannot apply to photon worldlines since neither τ nor **U** is defined for such worldlines. The basic principle that $\mathbf{dp} = 0$ along the worldlines still applies, of course. To use this for a null line $x^\mu(\lambda)$ we can choose a special parameter, an "affine parameter" λ such that[14]

$$p^\mu = \frac{dx^\mu}{d\lambda} . \tag{4.6}$$

The equation

$$\frac{d\mathbf{p}}{d\lambda} = 0 \tag{4.7}$$

has as its components Eq. (4.3). When Eq. (4.6) is used in Eq. (4.3) and the details worked out the result is just Eq. (4.5) with τ replaced by λ.

Geodesics in Minkowski space are, of course, straight lines. For a more interesting example let us turn to the (as yet unexplained) Schwarzschild geometry in Eq. (2.17e) and ask whether there are "circular" (constant r) geodesics. Such geodesics correspond to circular particle orbits in this geometry. We will furthermore intuit that such circular orbits can lie in the "equatorial plane" $\theta = \pi/2$. The real test of this assumption of course will be whether the geodesic equation is satisfied. We seek then geodesics characterized by $\theta = \pi/2$, $d\theta = 0$, $r = \text{const}$, $dr = 0$. For such geodesics we immediately have $p^\theta = 0$ and $p^r = 0$. Only p^0 and p^ϕ are nonzero. We now use these in Eq. (4.3), or

$$p^\mu{}_{,\nu} p^\nu + p^\alpha p^\nu \Gamma^\mu_{\alpha\nu} = 0 , \tag{4.8}$$

which applies equally well to (massive) particles or to photons. The four differential equations represented by Eq. (4.8) can be evaluated with the Christoffel symbols given in the Appendix. The $\mu = 0$ and $\mu = \phi$ equations tell us that p^0 and p^ϕ are constants along the geodesics, and the $\mu = \theta$ equation is automatically satisfied. The critical equation is the radial ($\mu = r$) equation that reduces to

$$(p^0)^2\Gamma^r_{00} + (p^\phi)^2\Gamma^r_{\phi\phi} = 0 .$$

With the Γ's in the Appendix this gives

$$(p^\phi/p^0)^2 = r_g/2r^3$$

From $p^\mu \propto dx^\mu$ it follows that $p^\phi/p^0 = c^{-1}\, d\phi/dt$. This tells us that a geodesic is given by

$$r = \text{const} , \quad \theta = \pi/2 , \quad \phi = \pm\omega t , \tag{4.9}$$

where

$$\omega^2 \equiv r_g c^2/2r^3 . \tag{4.10}$$

We now can ask what kind of geodesic this is, timelike or null. From the metric coefficients of Eq. (2.17e) and the formula in Eq. (3.25) we calculate

$$\mathbf{p}\cdot\mathbf{p} = g_{\mu\nu}p^{\mu}p^{\nu} = g_{00}(p^0)^2 + g_{\phi\phi}(p^{\phi})^2$$
$$= (p^0)^2\left[-\left(1-\frac{r_g}{r}\right)+\frac{r^2\omega^2}{c^2}\right] = (p^0)^2\left(-1+\frac{3r_g}{2r}\right). \tag{4.11}$$

For $r > 3r_g/2$ the geodesic is timelike and represents a possible orbit for a free (massive) particle. For $r = 3r_g/2$ the geodesic is null and represents a possible photon orbit. Circular orbits are impossible for $r < 3r_g/2$.

B. Geodesic deviation and curvature

As discussed in Sec. II the real measure of gravity is the relative acceleration of nearby free particles. We are now ready to relate this to space-time geometry. Imagine a family of timelike geodesics such that in some finite region of space-time there is a geodesic through each point (i.e., the geodesics fill the region and do not cross). Adjust the zero point of proper time (the time the clock on each geodesic starts) so that for nearby points on nearby geodesics the proper time difference is small. Since there is a geodesic through every point we have a value of $\mathbf{U} = \mathbf{ds}/d\tau$ at every point and we can treat $\mathbf{U}$ as a vector field. Since every curve is a geodesic Eq. (4.4) is satisfied at every point in the region.

We shall now pay particular attention to one arbitrarily chosen geodesic C_1 in the neighborhood of one of its points P. (See Fig. 6.) Let C_2 be a nearby geodesic and define ξ as the differential vector connecting nearby points of equal proper time τ on C_1 and C_2. We now calculate the relative four-velocity $\mathbf{V}_{\rm rel}$ of the two geodesics. By taking the difference of $\mathbf{U}$ at the ends of the differential vector ξ we find $\mathbf{V}_{\rm rel} = \mathbf{dU}$ or, in components,

$$V^{\beta}_{\rm rel} = U^{\beta}{}_{;\gamma}\xi^{\gamma}. \tag{4.12}$$

Since the relative velocity is also the rate at which ξ changes, $\mathbf{V}_{\rm rel}$ must equal $\mathbf{d\xi}/d\tau$ so that

$$V^{\beta}_{\rm rel} = \left(\frac{\mathbf{d\xi}}{d\tau}\right)^{\beta} = \xi^{\beta}{}_{;\gamma}\frac{dx^{\gamma}}{d\tau} = \xi^{\beta}{}_{;\gamma}U^{\gamma}. \tag{4.13}$$

The result

$$U^{\beta}{}_{;\gamma}\xi^{\gamma} = \xi^{\beta}{}_{;\gamma}U^{\gamma} \tag{4.14}$$

follows from a comparison of Eqs. (4.12) and (4.13). We shall need also a result that follows from Eq. (4.4):

$$0 = (U^{\alpha}{}_{;\gamma}U^{\gamma})_{;\beta}\xi^{\beta} = U^{\alpha}{}_{;\gamma;\beta}U^{\gamma}\xi^{\beta} + U^{\alpha}{}_{;\gamma}U^{\gamma}{}_{;\beta}\xi^{\beta}. \tag{4.15}$$

With the aid of Eq. (4.14) this can be written as

$$U^{\alpha}{}_{;\gamma;\beta}U^{\gamma}\xi^{\beta} = -U^{\alpha}{}_{;\gamma}\xi^{\gamma}{}_{;\beta}U^{\beta}. \tag{4.16}$$

The relative velocity $\mathbf{V}_{\rm rel}$ between the geodesics tells us nothing about space-time geometry, just as the relative velocity of two particles tells us nothing about the effects of gravity. Rather we must consider the relative acceleration $\mathbf{a}_{\rm rel}$

$$\frac{\mathbf{dV}_{\rm rel}}{d\tau} = \frac{d}{d\tau}\left(\frac{\mathbf{d\xi}}{d\tau}\right) \equiv \mathbf{a}_{\rm rel}. \tag{4.17}$$

This will tell us—in terms of Newtonian jargon—the relative gravitational acceleration of nearby free particles and thereby, in a coordinate independent manner, will reveal the presence of gravity. From the viewpoint of geometry, it will measure the relative bending of "the best straight lines," and hence will reveal curvature. We need only persevere in the rather tedious task of evaluating, with Eqs. (4.12) and (4.16), the components of Eq. (4.17) in a coordinate basis system:

$$a^{\alpha}_{\rm rel} = \left(\frac{d\mathbf{V}_{\rm rel}}{d\tau}\right)^{\alpha} = V^{\alpha}_{{\rm rel};\beta}\frac{dx^{\beta}}{d\tau} = V^{\alpha}_{{\rm rel};\beta}U^{\beta}$$
$$= [U^{\alpha}{}_{;\gamma}\xi^{\gamma}]_{;\beta}U^{\beta} = U^{\alpha}{}_{;\gamma;\beta}\xi^{\gamma}U^{\beta} + U^{\alpha}{}_{;\gamma}\xi^{\gamma}{}_{;\beta}U^{\beta}$$
$$= [U^{\alpha}{}_{;\gamma;\beta} - U^{\alpha}{}_{;\beta;\gamma}]\xi^{\gamma}U^{\beta}. \tag{4.18}$$

The term in square brackets here is interesting since it involves the commutation of second covariant derivatives [cf. Eq. (3.49)]. In flat space-time we could introduce flat (not just locally flat!) coordinates and the term in brackets would read

$$U^{\alpha}{}_{,\gamma,\beta} - U^{\alpha}{}_{,\beta,\gamma}\{\text{flat space-time} \tag{4.19}$$

and would vanish since partial derivatives commute. Since the term in square brackets is tensorial it follows that it must vanish in flat space-time in any basis; it is a zero tensor in flat space-time. Its value then will indicate something of the nature of space-time curvature.

Equation (4.18) still awaits final evaluation which we accomplish using

$$U^{\alpha}{}_{;\gamma;\beta} = [U^{\alpha}{}_{;\gamma}]_{;\beta} = [U^{\alpha}{}_{;\gamma}]_{,\beta} + U^{\lambda}{}_{;\gamma}\Gamma^{\alpha}_{\lambda\beta} - U^{\alpha}_{;\lambda}\Gamma^{\lambda}_{\gamma\beta}, \tag{4.20b}$$

and so forth. After much manipulation we are rewarded with a surprising result

$$a^{\alpha}_{\rm rel} = R^{\alpha}{}_{\lambda\beta\gamma}U^{\lambda}U^{\beta}\xi^{\gamma}, \tag{4.21}$$

$$R^{\alpha}{}_{\lambda\beta\gamma} \equiv \frac{\partial}{\partial x^{\beta}}(\Gamma^{\alpha}_{\lambda\gamma}) - \frac{\partial}{\partial x^{\gamma}}(\Gamma^{\alpha}_{\lambda\beta}) + \Gamma^{\alpha}_{\sigma\beta}\Gamma^{\sigma}_{\lambda\gamma} - \Gamma^{\alpha}_{\sigma\gamma}\Gamma^{\sigma}_{\lambda\beta}. \tag{4.22}$$

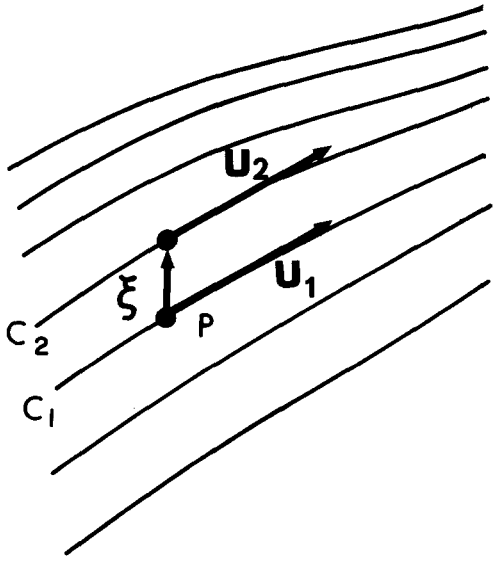

Fig. 6. Family of geodesics used for the discussion of geodesic deviation.

C. Curvature

The big surprise in Eq. (4.21) is that $a^{\alpha}_{\rm rel}$ at a point does not depend on the derivatives of $\mathbf{U}$ or ξ but only on the values of $\mathbf{U}$ and ξ at the point. Equation (4.21) is thus a tensorial equation relating $\mathbf{U}$, ξ, and $\mathbf{a}_{\rm rel}$ thereby defining a fourth-rank tensor $\mathbf{R}$ the "Riemann curvature tensor." The components of this tensor can be evaluated with the unlikely seeming combination of symbols on the right-hand side of Eq. (4.22). It is worthwhile to note that in LF coordinates

the Γ's vanish but not their derivatives, nor can we in general find coordinates [except in flat space(time)] in which the derivatives of the Γ's vanish. If we could find such coordinates then we would calculate the components of **R** to be zero in this coordinate system and hence in any coordinate system, since they are tensor components. For the Schwarzschild geometry and coordinates of Eq. (2.17e) we have, for example, that

$$R^{0}{}_{r0r} = r_g/r^2(r - r_g)\,, \tag{4.23}$$

so we can be sure for this geometry that coordinates do not exist in which the derivatives of the Γ's vanish; we can be sure that the Schwarzschild space-time is not flat.

The Riemann tensor clearly tells us something about curvature. Although it is certainly not obvious, the Riemann tensor in the following sense tells us *everything* about curvature: If the Riemann tensor vanishes in some finite region of space-time, coordinates can be found for which the metric formula takes on the flat form. *A region of space-time (or the whole of space-time) is flat if and only if the Riemann tensor vanishes in that region.* A complete examination of space-time curvature then requires the computation, in some basis system of all the components of **R**. This task is formidable but eased somewhat by the fact that the $4^4 = 256$ components of **R** are not all independent. The component symmetries of R reduce the number of independent components to a "mere" 20. [In an N-dimensional metric space or space-time the general result is $N^2(N^2-1)/12$.] These 20 components carry the information about curvature that we heuristically summarized as the "radius of curvature" in Sec. II B.

There is no measure of spacetime curvature at a point with more information than that contained in the Riemann tensor, but there are measures containing less information. Most importantly, by "contracting" (see Sec. III B) **R** on two of its indices we define a second-rank tensor, the "Ricci curvature tensor" a second-rank tensor with components

$$R_{\lambda\gamma} \equiv R^{\alpha}{}_{\lambda\alpha\gamma}\,. \tag{4.24}$$

The tensor is symmetric on its indices and hence only 10 of its 16 components are independent. By contracting once more we arrive at the Ricci scalar

$$R \equiv R^{\lambda}{}_{\lambda} = g^{\lambda\gamma}R_{\lambda\gamma}\,, \tag{4.25}$$

which gives only a single numerical measure of curvature at each point of spacetime.

D. GRT field equations

We have already discussed, toward the end of Sec. II C, the need for at least two mathematical measures of curvature. The Riemann tensor certainly seems correct for Eq. (2.35) since it describes the relative acceleration of free particles and since it describes everything about spacetime curvature, just as $\nabla\Phi$ in Newtonian theory describes everything about gravity. According to the discussion at the end of Sec. II C the Riemann curvature tensor itself must not be set equal to an expression involving gravitational sources. The Ricci tensor and Ricci scalar seem more appropriate to this task since they can vanish without the Riemann tensor vanishing. A plausible seeming source equation for gravity might, for example, be

$$R = kGc^{-4}\rho\,, \tag{4.26}$$

where R is the Ricci scalar, ρ is the mass-energy density, and k is a dimensionless constant. This source equation would have the Ricci scalar vanish outside gravity sources (i.e., where $\rho = 0$) but the Riemann tensor could still be nonzero. This theory then would allow space-time curvature, tidal gravitational effects, etc. outside of sources.

Equation (4.26) cannot of course be the correct source equation. The left-hand side of this equation is a scalar but the right-hand side—as was pointed out in Sec. II C—is a component of a second-rank tensor. This tensor, the "stress-energy tensor" **T**, is a symmetric ($T_{\mu\nu} = T_{\nu\mu}$) tensor that contains all the information we need about the nature of a gravity source. The physical meaning of **T** can be seen in its components as evaluated in an orthonormal basis corresponding to some observer. The components are related to the mass-energy density, etc. measured by that observer in the following way:

$$\left.\begin{aligned} T^{00} &= \text{mass-energy density} \\ T^{0i} &= i\text{th component of energy flux} \times c^{-1} \\ T^{ij} &= \text{components of the stress tensor} \end{aligned}\right\}\ i,j = x,y,z\,. \tag{4.27}$$

(For the meaning of the stress tensor see Sec. III B.) It should be understood that **T** can contain contributions from particles (massive or massless) and from fields (electromagnetic, nuclear, . . .) but the *stress-energy tensor contains no contributions identifiable as gravitational energy, gravitational energy flux, or gravitational stress.* This is inherent in the prescription above that the components are measured in a local system, in which gravity does not make an appearance. These few sketchy remarks are all that will appear here about the stress-energy tensor. A deeper discussion of **T** and (most regrettably) examples of how it is computed will be victims of the brevity of this tutorial article.

From the discussion in Sec. II it is by now clear that the source equation for gravity should relate the Ricci tensor to the stress-energy tensor. In GRT the precise form of this relation (the Einstein field equations) between curvature and stress-energy is[15]

$$R_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R = 8\pi Gc^{-4}T_{\mu\nu}\,. \tag{4.28}$$

This equation seems plausible enough. It sets a symmetric tensor equal to a symmetric tensor, it is dimensionally correct, and it has the right aura of having mass-energy density, etc. place a condition on space-time curvature. The real justification of Eq. (4.28) is rather more solid and lies in its self consistency. Suppose that the source of gravity is a set of massive particles that interact only via their mutual gravity. The stress energy then is that calculated (in a way not to be found in this article) for the particles. The particle motion and the time-changing geometry are interrelated according to Eq. (4.28). At the same time, we know, the particle motion and the geometry are related in that particle world lines must follow geodesics. For the specific form of Eq. (4.28) these requirements turn out to be compatible. This would not be the case, e.g., if the $-\tfrac{1}{2}g_{\mu\nu}R$ term were dropped in the equation.

For regions of space-time with $T_{\mu\nu} = 0$, Eq. (4.28) simplifies to

$$R_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R = 0\,. \tag{4.29}$$

If we contract the left-hand side on its two indices and use the fact that

$$g_{\mu}{}^{\mu} = g_{\mu\nu}g^{\nu\mu} = \delta^{\mu}{}_{\mu} = 4\,, \tag{4.30}$$

we have

$$R - \tfrac{1}{2}(4)R = -R = 0\,, \tag{4.31}$$

so that Eq. (4.29) may be replaced by

$$R_{\mu\nu} = 0 \ \{\text{Einstein field equations for vacuum.} \tag{4.32}$$

This is the form of the field equations we shall use in Secs. V and VI to examine the gravitational field outside of sources.

E. Calculations in GRT

We now claim to have a theory of gravity. Equation (4.28) tells us how mass-energy, etc. generates curvature, and geodesic motion—more generally Eq. (4.1)—tells us how particle dynamics[16] are influenced by curvature, but how is this theory used? It's helpful to review how a much more familiar theory, classical electrodynamics, is used. In principle in electrodynamics we could specify at any initial time the position and velocity of all charge carrying particles, and the instantaneous values of the **E** and **B** fields. The Lorentz force equation could then be used to evolve the position and velocity of the charged particle and the Maxwell equations could be used to evolve the **E** and **B** fields. In principle also, gravitational problems could be attacked in this way, but there are some important new subtleties. What do we evolve forward in time analogously to the evolution of **E** and **B**? The answer seems at first to be the metric components $g_{\mu\nu}$. The field equations (4.28) after all are, in essence, second-order differential equations for the $g_{\mu\nu}$'s. [The Riemann tensor, and hence the Ricci tensor and scalar contain derivatives of $g_{\mu\nu}$ only up to second order; see Eqs. (3.108) and (4.31).] The analogy to electrodynamics would then be to specify the $g_{\mu\nu}$ and their first time derivatives, at some initial time and to evolve the $g_{\mu\nu}$ via these differential equations. But this is not and cannot be all there is to it. If we could truly solve for a specific set of functions $g_{\mu\nu}$ we would be finding not only the geometry but also the coordinates in which we are describing the geometry subsequent to the initial time! The field equations cannot presume to force us to use a particular coordinate system, so contained in this set of equations there must not be (and there are not) enough independent equations to evolve the $g_{\mu\nu}$. What must in fact be done is that an *a priori* choice must be made of some characteristics of the coordinates, so that the coordinate system is in some sense fixed at the outset before the $g_{\mu\nu}$ are ever calculated.[17] The manner in which this is done is to put some constraints on the functional form of the $g_{\mu\nu}$ (four constraints to eliminate the four degrees of coordinate freedom). The constraints must be of such a nature that they impose no *a priori* restrictions on the geometry. Constraints involving the first derivatives of the $g_{\mu\nu}$ are used for this purpose since the information about the geometry is contained in the second derivatives of the $g_{\mu\nu}$'s. The technical details of this general approach are considerably beyond the scope of this tutorial article.

This all appears terribly complicated and it is. Only in the past few years, with very large computers fed by very persevering researchers have computations of this type ("numerical general relativity") become at all feasible. It is important to realize though that this type of "evolving forward in time" calculation is not the sort of problem scientists usually study in classical electrodynamics. Neither is it the type of problem usually investigated in GRT. Most problems in electrodynamics can be categorized as one of two types: (i) From considerations of the physical problem the nature of the sources is known at the outset (examples: static distributions of charge density, steady currents in wires, etc.); the problem is to find the fields due to these sources. (ii) The fields are known and the problem is to find the resulting motion of charged particles, magnetic dipoles, etc. In GRT tractable problems are also usually of these types. (Example: For a spherical static distribution of mass-energy find the exterior gravitational field—i.e., the space-time geometry.) In Sec. VI we shall see an example of such tractable problems but first, in Sec. V, we shall try to make a bridge to more familiar territory by asking what happens to the relativistic theory of gravity when gravitational fields are weak.

V. LINEARIZED THEORY

In GRT, gravity is described not as a field in space-time but rather as the curvature of space-time itself. We have, on the other hand, a venerable theory of gravity that for 300 years and for most present day science and engineering, gives perfectly adequate results, yet that contains no hint of the concept of space-time curvature. We investigate now the appearance of GRT in the limit of weak gravitational fields—gentle space-time curvature—in order to understand its correspondence to Newtonian theory.

If space-time curvature is zero in some region or the whole of space-time it is possible to introduce coordinates in which the metric takes on the flat Minkowski form $g_{\mu\nu} = \eta_{\mu\nu}$. If space-time is nearly flat it must be possible to introduce nearly flat coordinates, in which

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} \tag{5.1}$$

with small $h_{\mu\nu}$,

$$h_{\mu\nu} \ll 1\,. \tag{5.2}$$

"Nearly flat" here means the effects of gravity are weak. Somewhat more precisely it means the "radius of curvature" of space-time as measured by the Riemann tensor is large compared with other length scales of the problem. Clearly, the small "metric perturbations" $h_{\mu\nu}$ contain all the information about gravity. The essence of our technique in studying weak field gravity will be to compute needed expressions only to lowest order in the $h_{\mu\nu}$. The appearance of the resulting "linearized GRT" will be that of a theory couched in terms of Minkowski space-time, containing a dynamical tensor field $h_{\mu\nu}$, which governs gravitational phenomena. We *define*, for example, $h^{\mu}{}_{\nu}$ and $h^{\mu\nu}$ by

$$\begin{aligned} h^{\mu}{}_{\nu} &\equiv \eta^{\mu\alpha} h_{\alpha\nu}\,, \\ h^{\mu\nu} &\equiv \eta^{\mu\alpha}\eta^{\nu\beta} h_{\alpha\beta}\,, \\ h_{\mu\nu} &= \eta_{\mu\alpha}\eta_{\nu\beta} h^{\alpha\beta}\,, \end{aligned} \tag{5.3}$$

as if the metric and coordinates were Minkowskian. It should be noted that contravariant h's are not the perturbations of the contravariant metric components. Rather to first order in the perturbations,

$$g^{\mu\nu} = \eta^{\mu\nu} - h^{\mu\nu}\,. \tag{5.4}$$

The proof can be seen from

$$\begin{aligned} g^{\mu\nu} g_{\nu\alpha} &= (\eta^{\mu\nu} - h^{\mu\nu})(\eta_{\nu\alpha} + h_{\nu\alpha}) \\ &= \delta^{\mu}{}_{\alpha} - h^{\mu}{}_{\alpha} + h^{\mu}{}_{\alpha} + O(h^2) = \delta^{\mu}{}_{\alpha} + O(h^2)\,. \end{aligned} \tag{5.5}$$

A. Newtonian limit

Before developing the linearized theory in a little detail let us make the connection between GRT and Newtonian

theory. The weak field assumption is not enough for this; two more assumptions are needed. First, we recall that Newtonian theory and SRT are not compatible, so in our comparison we will require that particle velocities be very small compared to light velocity. Second, we know that in Newtonian theory changes in the gravitational field are propagated instantaneously. To avoid that possibility we shall require that the gravitational field be static. For our space-time then we will require that there exist a choice of time coordinate for which

$$h_{\mu\nu,0} \equiv \frac{1}{c}\frac{\partial h_{\mu\nu}}{\partial t} = 0 . \tag{5.6}$$

The geodesic equation, describing the motion of particles, has spatial ($i = x, y, z$) components

$$\frac{d^2x^i}{d\tau^2} = -U^\mu U^\nu \Gamma^i_{\mu\nu} . \tag{5.7}$$

The requirement of low velocity gives us

$$U^0 \gg U^i , \quad U^0 \approx c , \quad d\tau \approx dt , \tag{5.8}$$

so that Eq. (5.7) becomes

$$\frac{d^2x^i}{dt^2} \approx -c^2\Gamma^i_{00} \approx \tfrac{1}{2}c^2 h_{00,j}\eta^{ij} , \tag{5.9}$$

where approximately equal signs ($\approx$) indicate that corrections of order v/c and h have been ignored. In ordinary three-vector notation Eq. (5.9) reads

$$\frac{d^2\mathbf{x}}{dt^2} = \nabla(\tfrac{1}{2}h_{00}c^2) . \tag{5.10}$$

This has precisely the form of the Newtonian equation for gravitational acceleration if we make the identification to the Newtonian gravitational potential Φ by

$$h_{00} = -2\Phi/c^2 . \tag{5.11}$$

In linearized theory the field equations simplify greatly. The terribly nonlinear relationship between the metric and the Riemann tensor given in Eq. (4.22) simplifies to

$$R_{\alpha\mu\beta\nu} = \tfrac{1}{2}(h_{\alpha\nu,\mu\beta} + h_{\mu\beta,\nu\alpha} - h_{\mu\nu,\alpha\beta} - h_{\alpha\beta,\mu\nu}) , \tag{5.12}$$

if terms of order h^2 are ignored. The commas in the subscripts here indicate, of course, partial derivatives (equal to covariant derivatives for the Minkowski metric). With Eq. (5.12) we can fairly easily compute the Ricci tensor to find that

$$R_{00} = -\tfrac{1}{2}h_{00,\mu\nu}\eta^{\mu\nu} . \tag{5.13}$$

The vacuum field equation [see Eq. (4.32)] $R_{00} = 0$ then reads

$$R_{00} = -\tfrac{1}{2}\sum_{i=1}^{3}\frac{\partial^2 h_{00}}{\partial(x^i)^2} = \nabla^2(-\tfrac{1}{2}h_{00}) = 0 , \tag{5.14}$$

which is compatible with the identification in Eq. (5.11) and the Newtonian equation $\nabla^2\Phi = 0$. In fact the Newtonian limit[18] of the nonvacuum equation (4.28) is precisely the Newtonian equation

$$\nabla^2(-\tfrac{1}{2}h_{00}c^2) = 4\pi G\rho , \tag{5.15}$$

(where ρ is mass density) if the identification in Eq. (5.11) is made.

The above Newtonian limit of GRT clarifies an issue raised in Sec. II. It was noted that Newtonian theory cannot be modified into a scalar theory of gravity because the Newtonian potential Φ is really the component of a second-rank space-time tensor, just as the electric potential Φ_e is really the component of a space-time vector **A**. We have in the above discussion the mathematical meaning of that claim.

B. Lorentz gauge

We now return to the question of linearized theory without the restriction of the Newtonian limit. To do this we must confront an important but subtle point: There is not a unique choice of coordinates in which the metric takes on the form [see Eqs. (5.1) and (5.2)] "metric = Minkowski metric plus small perturbation." We have of course the freedom to make Lorentz transformations of our nearly Minkowskian coordinates. Under such transformations $\eta_{\mu\nu}$ and $h_{\mu\nu}$ transform, of course, just as the tensors of SRT do. In addition, however, we have the possibility of transforming from one set of nearly flat coordinates to another that wiggle in a slightly different way. The metric perturbation $h_{\mu\nu}$ in other words contains information not only about gravitational fields but also about just which "nearly flat coordinates" we have chosen. This in fact is a miniature version of the coordinate choice ambiguities sketched out in Sec. IV E. [This caused no difficulty in the above discussion of the Newtonian limit chiefly because the condition in Eq. (5.6) sufficiently fixed the coordinates.]

Rather than decry this situation we shall exploit it and shall choose the coordinates in such a way that the $h_{\mu\nu}$ have a conveniently simple form. This type of exploitation should not be totally unfamiliar. The electromagnetic potentials, for example, have ambiguity somewhat analogous to that of the metric perturbations. This ambiguity is expressed in the fact that the four-vector potential **A** can be replaced by

$$\mathbf{A}_{\text{new}} = \mathbf{A} + \mathbf{grad}\,\Lambda , \tag{5.16}$$

where Λ is any scalar function, without changing the measureable quantities (the **E** and **B** fields) that can be derived from **A**. The general freedom to make a change in a potential-like field (i.e., one that is part of the mathematical structure of the theory but is not directly measureable) is called gauge freedom and the allowable change in the field, a gauge transformation. By changing the nearly flat coordinates in our nearly flat space-time we effect therefore a gauge transformation of the $h_{\mu\nu}$. In electrodynamics this freedom is used to demand that **A** satisfy some simple condition, e.g., the "Lorentz gauge"

$$\nabla\cdot\mathbf{A} \equiv A^\mu{}_{,\mu} = 0 . \tag{5.17}$$

In linearized theory we make a similarly convenient gauge choice. This choice is expressed most easily if we first define

$$\begin{aligned} h &\equiv h_\alpha{}^\alpha \equiv \eta^{\alpha\beta}h_{\alpha\beta} , \\ \bar{h}_{\alpha\beta} &\equiv h_{\alpha\beta} - \tfrac{1}{2}\eta_{\alpha\beta}h , \\ \bar{h}^{\alpha\beta} &\equiv \eta^{\alpha\mu}\eta^{\beta\nu}\bar{h}_{\mu\nu} , \\ \bar{h} &\equiv \bar{h}_\alpha{}^\alpha = \eta^{\alpha\beta}\bar{h}_{\alpha\beta} = -h . \end{aligned} \tag{5.18}$$

The gauge choice, also called the "Lorentz gauge," for the metric perturbations is

$$\bar{h}^{\alpha\beta}{}_{,\beta} = 0 . \tag{5.19}$$

The convenience of this choice shows up in the linearized field equations. If the Riemann tensor is computed from Eq. (5.12) the first-order field equation in terms of $\bar{h}$ are found to be

$$\Box^2\bar{h}_{\mu\nu} \equiv \bar{h}_{\mu\nu,\alpha\beta}\eta^{\alpha\beta} = -16\pi G c^{-4} T_{\mu\nu} \quad (5.20a)$$

$$= 0 \text{ (vacuum)} . \quad (5.20b)$$

In this gauge, linearized GRT does indeed have the appearance of an "ordinary" special relativistic field theory.

C. Gravitational waves

Equation (5.20b) has the form of a tensor wave equation and we can immediately find solutions to it of the form

$$\bar{h}_{\mu\nu} = A_{\mu\nu} \exp[ik(z - ct)] . \quad (5.21)$$

This is the solution for a monochromatic gravitational wave (a wave of the gravitational field, a wave of space-time geometry) propagating in the positive z direction with frequency $\omega = kc$. The $A_{\mu\nu}$ are constant, are symmetric $A_{\mu\nu} = A_{\nu\mu}$, and are required by Eq. (5.19) to satisfy the four gauge conditions

$$A^{00} - A^{0z} = A^{x0} - A^{xz} = A^{y0} - A^{yz} = A^{z0} - A^{zz} = 0 . \quad (5.22)$$

The physical meaning of the gravitational wave is contained in the effect of the wave on the relative acceleration of nearby particles.[19] Let us consider therefore an experiment in which two nearby particles are nearly at rest in the laboratory, so that in the laboratory frame the four-velocity $\mathbf{U}$ of either satisfies

$$U^0 \approx c , \quad U^i \ll U^0 . \quad (5.23)$$

The relative acceleration of the two particles is then given by the equation of geodesic deviation (4.21) in this slow-motion approximation:

$$a^{\alpha}_{\text{rel}} = c^2 R^{\alpha}_{00\gamma} \xi^{\gamma} , \quad (5.24)$$

where t is laboratory time and $\boldsymbol{\xi}$ is the displacement vector between the two particles.

With Eq. (5.12) the Riemann components for the gravitational wave of Eqs. (5.21) and (5.22) can be explicitly computed. The only nonvanishing (to first order in $h_{\mu\nu}$) components of the type that is needed for Eq. (5.24) are

$$c^2 R^{x}{}_{00x} = -c^2 R^{y}{}_{00y} = \tfrac{1}{4}\omega^2 (A_{yy} - A_{xx}) \exp[ik(z - ct)] , \quad (5.25)$$

$$c^2 R^{x}{}_{00y} = c^2 R^{y}{}_{00x} = -\tfrac{1}{2}\omega^2 A_{xy} \exp[ik(x - ct)] .$$

This result shows that the effects of the passage of the wave in the z direction are purely transverse to the z direction; the accelerations are only in the x and y direction and involve only the x and y components of the particle separation. The wave, furthermore, has only two degrees of freedom, contained in the numbers A_+ and $A_\times$:

$$A_+ \equiv \tfrac{1}{2}(A_{xx} - A_{yy}) , \quad A_\times \equiv A_{xy} . \quad (5.26)$$

In Fig. 7 is shown the effect, on a circular ring of free particles, of the passage of a "pure A_+" (i.e., $A_\times = 0$) wave and of a pure $A_\times$ (i.e., $A_+ = 0$) wave. It is clear in the figure that the effect of the wave is to force the ring to deform into an elliptical appearance. For the pure A_+ wave the ring of particles oscillates between an ellipse with the x axis as the major axis and an ellipse with the y axis as the major axis. The pure $A_\times$ wave has a similar effect but the principal axes are rotated by $\frac{1}{4}\pi$. This is reminiscent of the two linear polarizations of a plane electromagnetic wave for which the two independent linear polarizations (e.g., E^x and E^y) have the same physical effect rotated by $\pi/2$. Much of the familiar terminology of electromagnetic radiation is used also for gravitational waves. The two independent gravitational wave modes A_+ and $A_\times$ are called linear polarization modes and have many mathematical properties similar to those of the linear polarization modes of the electromagnetic wave. A superposition of A_+ and $A_\times$ waves with the same phase (or 180° phase difference) gives a wave whose effect is to deform a ring of particles into an oscillating ellipse with principal axes not necessarily aligned with, or at $\frac{1}{4}\pi$ to, the xy axes. A superposition of A_+ and $A_\times$ of equal magnitude but 90° phase difference gives a circularly polarized wave, a wave whose effect is to produce a rotating elliptical deformation on the ring of particles.

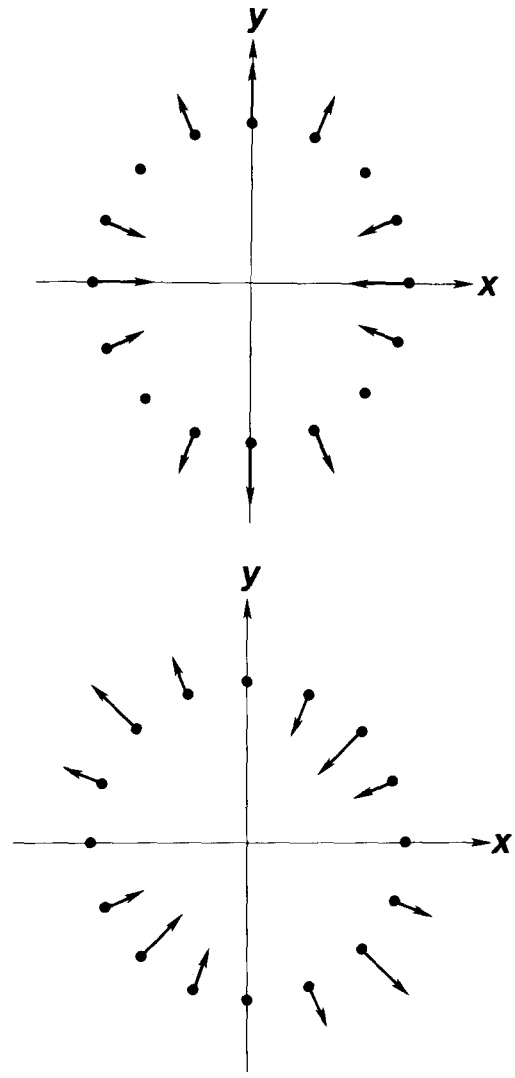

Fig. 7. Effect on a circle of particles of the passage in the z direction of a gravitational wave of (a) pure type A_+ and (b) pure type $A_\times$. The arrows indicate the magnitude of the acceleration relative to the center of the circle, at phase zero of the wave.

The difficulty of experimentally detecting gravitational waves can be found roughly from the above discussion of the geodesic deviation equation. Equations (5.24) and (5.25) tell us that the gravitationally induced acceleration is of order:

$$\text{acceleration} \sim \omega^2 (\text{displacement})(\text{size of } h_{\mu\nu}) . \quad (5.27)$$

Attempts are now being made to detect gravitational waves of typical frequency $\omega \approx 10^3$ sec^{-1} with detectors of size ~ 1 m, so that effective accelerations are

$\sim 10^6$ m/sec^2(size of $h_{\mu\nu}$). (5.28)

The expected size of metric perturbations in the neighborhood of the Earth due to explosive astrophysical events in the Universe, e.g., supernovae in distant galaxies, is of order 10^{-21} or 10^{-22} so that to detect such events gravitational detectors must measure acceleration on the order of 10^{-15} m/sec^2.

Do gravitational waves carry energy? Imagine the ring of particles to contain weak springs connecting the particles. We can be sure, even without going through the details, that under the combined action of the springs and the gravitational wave, the ring of particles must still deform somewhat. This will cause the springs to stretch so that elastic energy appears in the springs. In this sense the gravitational wave certainly must contain energy if we are to maintain that energy is conserved.

There is a good reason that the word "energy" is being handled with such delicacy here. With GRT we have the usual kind of energy conservation on a small scale; in a laboratory small compared to the radius of curvature of space-time, energy is conserved. "Gravitational energy" of course doesn't enter into the energy budget since there are no gravitational effects on such a small space-time scale. There is no large scale equivalent of this very basic principle of physics. In GRT there is no *general* principle of large-scale energy conservation. It only requires some musing about such concepts as the "gravitational potential energy of the whole Universe" to make it plausible—or at least acceptable—that in GRT there is no such thing. Common sense, on the other hand, tells us there are situations (e.g., gravitational binding energy of a star) in which there is such a thing as gravitational energy and common sense, even in curved space-time, is a powerful principle of physics. For this reason it is good that in situations where gravitational energy makes sense we can find mathematical things to say about it. One such situation is the energy stored in gravitational waves. This subject was not really clarified[20] until the 1960s and we cannot discuss it in detail here except to say that gravitational wave energy is meaningful only when averaged over several wavelengths. Unlike the case for classical electromagnetic waves, energy cannot be sharply localized in the waves. In any case with this conceptual foundation fairly simple formulas, not very different in appearance from the analogous electromagnetic formulas, can be worked out for gravitational wave energy flux, and with these formulas gravitational wave astronomy can be done in blissful ignorance of the slippery nature of gravitational energy.

VI. SCHWARZSCHILD GEOMETRY

A. Derivation of the metric

The most important solution to the GRT field equations is fortunately one of the simplest, the spherically symmetric vacuum solution. This is the solution for the gravitational field outside a spherical star (or spherical whatever). It is the analog of the solution $\Phi = -GM/r$ to the Newtonian field equation $\nabla^2\Phi = 0$. The geometry is called the Schwarzschild geometry after Karl Schwarzschild[21] who first found this solution to Einstein's vacuum field equations in 1916.

As is always the case in GRT the metric will contain information both about the geometry and about the coordinates. Even in Minkowski space-time the metric can be made to look arbitrarily complicated with a sufficiently foolhardy choice of coordinates. To arrive at a form of the metric that makes the simplicity of the Schwarzschild geometry manifest let us then first carefully choose coordinates appropriate to the spherical nature of the geometry.[22] By "spherical symmetry" we mean that space-time can be filled with closed two-dimensional surfaces ("spherical surfaces") and that on each of these surfaces there is no geometric distinction between one point and any other. The familiar coordinates for labeling points on such a surface are the usual angles θ and ϕ. The other two coordinates of space-time, the coordinates that distinguish one spherical surface from another, will be named T and r. The requirement of spherical symmetry is then the requirement that for $dT = 0$, $dr = 0$ (i.e., at fixed T,r) distances are determined by

$$ds^2 = f(r,T)(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{6.1}$$

So far we have asked nothing special of our coordinates r and T. By being more demanding we can make Eq. (6.1) look more familiar. We require that r and T are labels on the surfaces such that $f = r^2$, or

$$ds^2 = r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{6.2}$$

We are then geometrically fixing the meaning of r: it is the surface label for which surface area $= 4\pi r^2$. This coordinate, the "Schwarzschild radial coordinate," is geometrically defined and is convenient for many purposes but it is by no means forced on us by spherical symmetry.[23]

The question now is how to extend Eq. (6.2) for displacements in which dT and dr are nonzero. The answer is

$$ds^2 = -A(r,T)dT^2 + B(r,T)dr^2 + C(r,T)dr\,dT + r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{6.3}$$

Many terms are missing in this formula. It is simple to argue, for example, that there can be no $dr\,d\phi$ term; such a term would constitute a geometric distinction between increasing and decreasing ϕ (positive or negative $d\phi$) and would be incompatible with spherical symmetry. Variations of this argument account for the omission of $dr\,d\theta$ terms, $dT\,d\phi$ terms, and $dT\,d\theta$ terms.

Spherical symmetry does not require the absence of the $dr\,dT$ term. Rather, we can eliminate it by making yet another demand on our coordinates. So far θ and ϕ have been geometrically fixed (except for trivial rotations) and r has been fixed but nothing has been asked of T. We now specialize T in such a way that the $dT\,dr$ term is banished.[24] With this term gone we can rewrite the metric formula in the very simple form

$$ds^2 = -e^{2\Phi}c^2\,dT^2 + e^{2\Lambda}\,dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2), \tag{6.4}$$

where Φ and Λ are functions of r and T. But even now we have not completely specialized the coordinates. We can still make a transformation to a new time coordinate t by

$$T = F(t) \tag{6.5}$$

without ruining the pleasingly simple form of Eq. (6.4). (We shall presently take advantage of this freedom.) It is appropriate to pause in the middle of these mathematical manipulations and admire what a wondrous thing we have accomplished. Merely by invoking spherical symmetry—and by imposing constraints on the coordinates—we have reduced the problem of finding a space-time geometry to that of finding two functions, $\Phi(r,t)$ and $\Lambda(r,t)$.

This is as far as we can go with symmetry arguments alone. To find Φ and Λ we need physics in the form of the (vacuum) GRT field equations $R_{\mu\nu}=0$. To find explicitly the differential equations that result from $R_{\mu\nu}=0$ is straightforward with Eqs. (4.22), (4.24), and (4.32), but very tedious. The resulting equations, however, are rather simple. From $\mu\nu=0r$ we get

$$R_{0r}=\frac{2}{r}\frac{\partial\Lambda}{\partial T}=0\,, \tag{6.6}$$

showing, surprisingly, that Λ is a function of r only. From the other components of $R_{\mu\nu}=0$ we find

$$-2e^{-2\Lambda}\frac{\partial\Lambda}{\partial r}=\frac{1}{r}(1-e^{-2\Lambda})\,, \tag{6.7}$$

$$2\Phi_{,r}=-\frac{1}{r}(1-e^{2\Lambda})\,. \tag{6.8}$$

Equation (6.7) can immediately be solved to give

$$e^{-2\Lambda}=1-r_g/r\,, \tag{6.9}$$

where r_g, the "gravitational radius," is an integration constant. With this result Eq. (6.8) then gives

$$2\Phi=\ln(1-r_g/r)+2K(T)\,, \tag{6.10}$$

where $K(T)$ is an integration constant of Eq. (6.8). The Schwarzschild geometry then is

$$ds^2=-(1-r_g/r)c^2e^{2K}\,dT^2+(1-r_g/r)^{-1}\,dr^2 +r^2(d\theta^2+\sin^2\theta\,d\phi^2)\,. \tag{6.11}$$

There is an arbitrary function $K(T)$ because we still have not "fixed" the meaning of the time coordinate. The best choice of course is that of the time coordinate corresponding to Minkowski time in the asymptotically flat geometry at $r\to\infty$. To arrange this we use the freedom inherent in Eq. (6.5) and choose

$$\frac{dF}{dt}=e^{-K(T)}\,, \tag{6.12}$$

so that the metric takes the form [cf. Eq. (2.17e)]

$$ds^2=-(1-r_g/r)c^2\,dt^2+(1+r_g/r)^{-1}\,dr^2 +r^2(d\theta^2+\sin^2\theta\,d\phi^2)\,. \tag{6.13}$$

A rather remarkable feature of this result is worth emphasizing. There is no t dependence in the metric functions of Eq. (6.13). It is, in this sense, static—the same at all t. The exterior geometry of a star then pays no attention to whether the star itself is static, oscillating, or collapsing. No gravitational signal carries information about the dynamical status of a gravitating *spherical* source. To make this seem more plausible we should remember that in Sec. V we saw that the effects of gravitational waves are transverse. Any radially outgoing disturbance in the gravitational field would therefore have to choose directions perpendicular to the radial direction and this would violate the spherical symmetry of the geometry. (Similar arguments explain why time varying spherically symmetric charge distributions produce static exterior electric fields.)

B. Interpretation of r_g

The only characteristic of the exterior geometry that tells anything about the source is the single parameter r_g, the "gravitational radius." We can take two approaches to investigating its meaning. One possibility is to revert to the nonvacuum field equations (4.28) and to relate r_g to the stress-energy of the source. We take the second approach and determine r_g operationally. To do this we need only ask what the behavior of distant test bodies will be in the Schwarzschild geometry. In Sec. IV A we saw that in this geometry circular timelike geodesics, e.g., planetary orbits, have an angular frequency given by [Eq. (4.10)]

$$\omega^2=c^2r_g/2r^3\,. \tag{6.14}$$

An astronomer studying a star (or whatever) infers its mass by measuring its gravitational effect on distant bodies. If he observes planets, in orbit about a star, with angular frequency ω at radius r, he will *measure* the mass to be $r^3\omega^2/G$. In a Schwarzschild geometry he then measures a mass

$$M=c^2r_g/2G\,. \tag{6.15}$$

This is what we shall *define* as the mass of the source in the Schwarzschild geometry. It corresponds to the familiar Newtonian mass

$$M_{\text{Newt}}=\int(\text{mass density})4\pi r^2\,dr\,, \tag{6.16}$$

only in the limit that space-time curvature is weak and the material of the star is nonrelativistic in the sense of Sec. V A. One further opportunity to check Eq. (6.15) and the consistency of our arguments is offered by the weak field limit ($r\gg r_g$) of the metric functions. According to Eq. (5.11), and the discussion that precedes it, in this limit we should have

$$g_{00}\approx-1-2\Phi/c^2=-1+r_g/r\,, \tag{6.17}$$

so that the Newtonian potential $\Phi=-GM/r$ must be given by

$$\Phi=-r_gc^2/2r\,, \tag{6.18}$$

and r_g must equal $2GM/c^2$, in agreement with Eq. (6.15).

C. Perihelion shift

A notorious non-Newtonian aspect of the Schwarzschild geometry is the explanation of the anomalous perihelion shift of the planet Mercury. We can derive this result by first developing a very useful lemma. Suppose that in some coordinate system the metric functions $g_{\mu\nu}$ are not functions of one of the coordinates,[25] say x^α. For a timelike geodesic describing free-particle motion, the covariant component U_α of the particle's four-velocity is a constant. (For spacelike and null geodesics the analogous result holds.) The proof is just a direct application of the geodesic equation (4.4) written as

$$U_{\alpha;\nu}U^\nu=0\,,$$

from which we have

$$\begin{aligned}U_{\alpha,\nu}U^\nu&=U_{\alpha,\nu}\frac{dx^\nu}{d\tau}=\frac{dU_\alpha}{d\tau}=U_\beta\Gamma^\beta_{\alpha\nu}U^\nu\\&=U^\sigma U^\nu g_{\sigma\beta}\Gamma^\beta_{\alpha\nu}\\&=\tfrac{1}{2}U^\sigma U^\nu(g_{\sigma\alpha,\nu}+g_{\sigma\nu,\alpha}-g_{\alpha\nu,\sigma})\,.\end{aligned}$$

Therefore for geodesic motion $U_\alpha(\tau)$ obeys, in general,

$$\frac{dU_\alpha}{d\tau}=\tfrac{1}{2}U^\sigma U^\nu g_{\sigma\nu,\alpha}\,. \tag{6.19}$$

If the metric functions are independent of x_α then $dU_\alpha/d\tau=0$ and U_α is a constant along the geodesic. This also implies that p_α is a constant and in this form the lemma applies also to photons.

Let us now use this result in the study of nearly circular orbits. The mathematical details will be much simpler if we choose the orbit to be in the equatorial plane ($\theta = \pi/2$, $U^\theta = 0$). We can of course check that this is consistent with the geodesic equation. It also involves no loss of generality since we can always choose the $\theta = \pi/2$ surface to coincide with the plane of the orbit. (These are the same assumptions we used for circular orbits in Sec. IV A.) We can use our lemma by noticing that the metric coefficients of (6.13) are independent of both t and ϕ, and hence that

$$U_0 = g_{00}U^0 = g_{00}c\frac{dt}{d\tau} = -\left(1-\frac{r_g}{r}\right)c\frac{dt}{d\tau} \equiv -E/c\,, \tag{6.20a}$$

$$U_\phi = g_{\phi\phi}U^\phi = g_{\phi\phi}\left(\frac{d\phi}{d\tau}\right) = r^2\frac{d\phi}{d\tau} \equiv J\,, \tag{6.20b}$$

where E and J are constants, called, respectively, the energy parameter and the angular momentum parameter, of the orbit. We have furthermore $\mathbf{U}\cdot\mathbf{U} = -c^2$ so that

$$\begin{aligned} c^2 &= -g_{00}(U^0)^2 - g_{rr}(U^r)^2 - g_{\phi\phi}(U^\phi)^2 \\ &= -g^{00}(U_0)^2 - g_{rr}(U^r)^2 - g^{\phi\phi}(U_\phi)^2 \\ &= \left(1-\frac{r_g}{r}\right)^{-1}\frac{E^2}{c^2} - \left(1-\frac{r_g}{r}\right)^{-1}\left(\frac{dr}{d\tau}\right)^2 - \frac{1}{r^2}J^2\,. \end{aligned} \tag{6.21}$$

Equations (6.20) and (6.21) give us three equations for the unknown functions $t(\tau)$, $\phi(\tau)$, $r(\tau)$ so we need never write down the geodesic equation. We have already used the geodesic equation, as much as we need to, by finding that U_0 and U_ϕ are constants. The lemma has indeed been useful. It has eliminated the need to deal with the second-order differential equations given by the geodesic equation, and has allowed us immediately to find three first integrals of those equations.

We are not really interested here in $t(\tau)$, $\phi(\tau)$, $r(\tau)$ but rather in the shape of the orbit, so we can ignore Eq. (6.20a) and use (6.20b) to write

$$\frac{dr}{d\tau} = \frac{dr}{d\phi}\frac{d\phi}{d\tau} = \frac{1}{r^2}\frac{dr}{d\phi}J$$

in Eq. (6.21). We next follow the standard trick of Newtonian orbital calculations by using the variable

$$u \equiv 1/r$$

rather than r itself. Equation (6.21) then takes the form

$$\left(\frac{du}{d\phi}\right)^2 + u^2 = \frac{E^2-c^4}{J^2c^2} + \frac{c^2}{J^2}r_g u + r_g u^3\,.$$

The result becomes even more familiar if we differentiate by ϕ and replace r_g according to Eq. (6.15)

$$\frac{d^2u}{d\phi^2} + u = \frac{GM}{J^2} + \frac{3GMu^2}{c^2}\,. \tag{6.22}$$

If we identify J as angular momentum per unit mass then Eq. (6.22) is just the Newtonian orbit equation except for the last term on the right-hand side, the GRT correction. [Equation (6.22) is treated in several undergraduate mechanics books.[26]] For nearly circular orbit the radius of the orbit is almost constant

$$u \approx u_0 \equiv 1/r_0\,,$$

and we write

$$u = u_0 + u'\,,$$

where $u' \ll u_0$. The orbital equation (6.22), to first order in u', is

$$\frac{d^2u'}{d\phi^2} + u' = \frac{GM}{J^2} + \frac{3GM}{c^2r_0^2} + \frac{6GM}{r_0c^2}u'\,. \tag{6.23}$$

If we define

$$\psi = (1 - 6GM/r_0c^2)^{1/2}\phi\,, \tag{6.24}$$

the equation simplifies to

$$\frac{d^2u'}{d\psi^2} + u' = \left(\frac{GM}{J^2} + \frac{3GM}{c^2r_0^2}\right)\Big/\left(1 - \frac{6GM}{r_0c^2}\right),$$

which has solutions of the form

$$u' = A + B\cos\psi + C\sin\psi\,.$$

The angular distance $\Delta\phi$ from one periastron to the next as shown in Fig. 8 is given by

$$\Delta\psi = 2\pi = (1 - 6GM/r_0c^2)^{1/2}\Delta\phi\,. \tag{6.25}$$

For a nearly Newtonian orbit with $r_0 \gg r_g$ we have

$$GM/r_0c^2 \ll 1\,,$$

so that

$$\Delta\phi \approx 2\pi(1 + 3GM/r_0c^2)\,. \tag{6.26}$$

For a nearly circular nearly Newtonian orbit, like that of Mercury, the periastron advance is therefore

$$6\pi GM/r_0c^2 \text{ rad/ orbit}\,.$$

For Mercury's orbital radius ($r_0 = 2.0\times10^{12}$ cm) and the sun's mass ($M = 2.0\times10^{33}$ g) this predicts an advance of Mercury's perihelion by 4.8×10^{-7} rad per orbit or 40 arcsec per century.

D. Gravitational red shift

Our intuition tells us that as a particle rises against the pull of gravity it loses (nongravitational) energy. How is this described in the mathematics of GRT? The answer starts with Eq. (3.11), which tells us that an observer with four-velocity $\mathbf{U}_{\text{obs}}$ will measure a particle (or photon) with four-momentum $\mathbf{p}$ to have observed energy $-\mathbf{p}\cdot\mathbf{U}_{\text{obs}}$. This results is based on considerations of local measurements and local frames but it is a scalar result and therefore gives

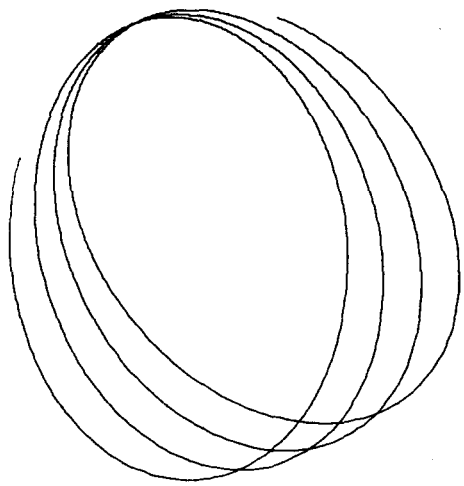

Fig. 8. General relativistic advance of the periastron of a planet. The periastron advance per orbit δ shown here is 0.19 rad per orbit, much larger than that of Mercury (4.8×10^{-7} rad per orbit).

the same numerical answer for evaluation in any frame. Let us now consider an observer "sitting still" in the Schwarzschild geometry at some fixed value of r,θ,ϕ. Since $dr = d\theta = d\phi = 0$ the only nonzero component (in the Schwarzschild coordinate basis) of his four-velocity is $\mathbf{U}^0_{\rm obs}$. To satisfy this requirement that $\mathbf{U}\cdot\mathbf{U} = -c^2$ this component must be

$$U^0_{\rm obs} = c(1 - r_g/r)^{-1/2} . \tag{6.27}$$

This observer, of course, is not following a geodesic; Eq. (6.27) does not satisfy the geodesic equation (4.4). He needs a rocket, or something, to remain "sitting still" against the "pull" of gravity. If a particle (free or accelerating) with four-momentum $\mathbf{p}$ passes this observer he will measure its energy to be

$$E = -\mathbf{p}\cdot\mathbf{U}_{\rm obs} = -p_\mu U^\mu_{\rm obs} = -c(1 - r_g/r)^{-1/2} p_0 . \tag{6.28}$$

If the particle is free (no influence except gravity) then according to our lemma following Eq. (6.19) p_0 is a constant. Equation (6.28) then gives us a simple equation for locally measured energy as a function of r. In particular, two observers, one at r_a and one at r_b will measure energies related by

$$\frac{E_b}{E_a} = \frac{(1 - r_g/r_a)^{1/2}}{(1 - r_g/r_b)^{1/2}} . \tag{6.29}$$

Clearly if $r_b > r_a$ the measurement at r_b reveals less energy than at r_a. It is in this sense that the free-particle loses energy as it moves outward, or gains energy as it falls inward. If the outer observer is at ∞ (e.g., an astronomer at $r_b \gg r_g$, and $r_b \gg r_a$) then Eq. (6.29) simplifies to

$$E_\infty / E_a = (1 - r_g/r_a)^{1/2} . \tag{6.30}$$

If a particle fights its way out against gravity from r_a to ∞, Eq. (6.30) gives the fraction of its energy it loses. In the nonrelativistic limit (slow particle motion; $r \gg r_g$) the particle energy is $\sim mc^2$, where m is the particle rest mass, so that the energy lost is roughly

$$(1 - r_g/r_a)^{1/2} mc^2 - mc^2 \simeq -\tfrac{1}{2}(r_g/r_a)mc^2 \simeq -(GM/r_a)m ,$$

in accordance with Newtonian theory.

A more interesting application of Eq. (6.30) is to consider a photon of frequency ν_a produced at radius r_a. An observer sitting still at r_a could, for example, produce such a photon as the result of an atomic transition. The frequency ν_a is the frequency he measures and the frequency that we normally associate with that transition neglecting gravity. From Eq. (6.30) the frequency observed at ∞ will be given by

$$\nu_\infty = \nu_a(1 - r_g/r_a)^{1/2} . \tag{6.31}$$

The "red shift" z of a photon is defined in terms of emitted wavelength $\lambda_a (\equiv c/\nu_a)$ and observed wavelength $\lambda_\infty (\equiv c/\nu_\infty)$. The gravitational red shift in the Schwarzschild geometry is

$$z \equiv \frac{\lambda_\infty - \lambda_a}{\lambda_a} = \frac{1}{(1 - r_g/r_a)^{1/2}} - 1 . \tag{6.32}$$

It is appropriate that the Planck constant h does not appear in any of these results since they have nothing to do with the quantum nature of light. The idea of photons with energy $h\nu$ is expedient, but not necessary.

The above derivation of the red shift is concise but it does not address a disturbing question: Suppose radiation is emitted at a constant rate outward from r_a. Since the frequency at which waves are received at $r_b > r_a$ is less than the frequency of emission at r_a, don't waves have to "pile up" in the region between the two radii? The answer to this question clarifies the meaning of the measurement process and the several different meanings of "time." Let two wave fronts of a light beam be emitted from radius r_a at coordinate times t_1 and t_2. Since the geometry itself is independent of t it must require the same "travel time," call it $t_{\rm trav}$, for each wave front to arrive at r_b. It follows that the difference

$$\Delta t_a \equiv t_2 - t_1$$

between emission times for the wave fronts at r_a is the same as the difference

$$\Delta t_b = (t_2 + t_{\rm trav}) - (t_1 + t_{\rm trav}) = t_2 - t_1$$

between reception times at r_b. The period of the wave *measured in "t time"* is the same at r_a and r_b. If this were not true waves would pile up in the intermediate region.

The coordinate time t, however, is *not* the time measured by the clocks of the observers measuring the waves' parameters. Their clocks measure *proper time* of the observers, related to coordinate time by

$$\frac{dt}{d\tau} = c^{-1} U^0_{\rm obs} . \tag{6.33}$$

The observers at different radii will see the same Δt for the wave fronts, hence they must observe different proper time intervals $\Delta\tau_a, \Delta\tau_b$. The ratio of *measured* frequencies will then be

$$\frac{\nu_b}{\nu_a} = \frac{\Delta\tau_a}{\Delta\tau_b} = \frac{(U^0_{\rm obs})_b}{(U^0_{\rm obs})_a} . \tag{6.34}$$

In the case of an observer "sitting still" in the Schwarzschild geometry Eq. (6.27) applies and Eq. (6.34) agrees with our previous result, Eq. (6.29).

D. Black holes

There is an expected and an unexpected pathology in the Schwarzschild metric, Eq. (6.13). At $r = 0$, presumably the location of the "point mass" curving space-time, the metric coefficients are ill behaved. We of course expect pathological effects there and we are not disappointed. These pathologies are real; an observer falling into the neighborhood of $r = 0$ experiences unboundedly large tidal forces. The pathology at $r = r_g$ is rather different. Whatever it is, we do not have to worry about it in common sense astronomy since r_g is too small. For the sun as an example $r_g = (2G/c^2) \times$ solar mass $= 3$ km, whereas the radius of the sun itself is 7×10^5 km. The Schwarzschild metric applies only in the source-free region *outside* the sun so the strangeness of the metric at $r = r_g$ is irrelevant. Nevertheless it is important, at least in principle, to understand this feature of the geometry. It is interesting that Laplace[27] in 1796 noticed something strange about $r = r_g$. The Newtonian escape velocity from radius r_a is

$$v^2_{\rm escape} = 2GM/r_a . \tag{6.35}$$

At a sufficiently small radius, Laplace argued, the escape velocity will be the velocity of light. That radius is given according to Eq. (6.35) by $2GM/c^2$, the gravitational radius r_g. Laplace must have been correct in some sense since Eq.

(6.30) tells us that a photon escaping from $r_a = r_g$ reaches ∞ with zero frequency, which is to say in some sense it doesn't reach ∞ at all. Equation (6.29) in fact tells us that a photon from $r_a = r_g$ can never reach *any* larger radius with finite energy.[28]

It is far from clear what all this means. There is something decidedly strange about $r = r_g$, but how strange? Our musings so far suggest that the geometry is singular at $r = r_g$; the metric coefficients, after all, are singular there. There is, however, an alternative to this drastic conclusion: the pathology of the metric coefficients may be a coordinate effect.[29] A simple example of such a coordinate pathology occurs for the two-dimensional spatial geometry

$$(ds)^2 = X^{-3}(dX)^2 + Y(dY)^2 . \tag{6.36}$$

This would seem to describe a very exotic geometry with pathological behavior at $X = 0$ and at $Y = 0$. Such a conclusion is an embarrassment once we make the coordinate transformation

$$X = 4x^{-2}, \quad Y = (\tfrac{3}{2}y)^{2/3}$$

with which Eq. (6.35) reduces to

$$(ds)^2 = (dx)^2 + (dy)^2 ,$$

the familiar, completely nonpathological, two-dimensional Euclidean geometry. The problem with the metric of (6.36) is the deceptive choice of coordinates. We cannot *a priori* be certain that the pathology in the metric of Eq. (6.13) is not also, at least partially, due to a deceptive coordinate choice. There must of course be *something* strange about $r = r_g$ since the infinite red shift predicted by Eq. (6.29) is a coordinate-independent prediction.

The idea that the Schwarzschild coordinates could be deceptive is disturbing since these coordinates are so solidly based on geometrical and physical considerations. The r, θ, ϕ coordinates are directly defined via spherical symmetry; the t coordinate is the choice of time coordinate that expresses the static nature of the geometry. Can such coordinates be deceptive? Our confusion at this point reflects to a small degree the confusion and controversy that still existed until relatively recently (the 1960s) about the nature of $r = r_g$.

Let us take a look at an argument that suggests that there *is* something misrepresentative about the Schwarzschild coordinates at $r = r_g$. We will consider a particle in radial free fall. ("Radial" means $d\theta = d\phi = 0$ so that $U^\theta = U^\phi = 0$ for the particle worldline.) From the r component of the geodesic equation (4.4) we have

$$\begin{aligned} U^r{}_{;\alpha}U^\alpha &= U^r{}_{,\alpha}U^\alpha + U^\alpha U^\lambda \Gamma^r_{\alpha\lambda} \\ &= \frac{dU^r}{d\tau} + (U^0)^2\Gamma^r_{00} + 2U^0U^r\Gamma^r_{0r} \\ &\quad + (U^r)^2\Gamma^r_{rr} . \end{aligned}$$

With the Γ's listed in the Appendix and $U^r \equiv dr/d\tau$ this becomes

$$\frac{d^2r}{d\tau^2} + \frac{r_g}{2r^2}\left[\left(1 - \frac{r_g}{r}\right)(U^0)^2 - \left(1 - \frac{r_g}{r}\right)^{-1}(U^r)^2\right].$$

The terms in square brackets [ ] are just

$$-g_{00}(U^0)^2 - g_{rr}(U^r)^2 = -\mathbf{U}\cdot\mathbf{U} = c^2 ,$$

so that the equation simplifies to

$$\frac{d^2r}{d\tau^2} = -\frac{r_g c^2}{2r^2} = -\frac{GM}{r^2}, \tag{6.37}$$

which has precisely the appearance of the analogous Newtonian result. Most important, it shows no strange behavior at $r = r_g$. According to this equation a free particle starting at rest will fall from any finite radius to $r = r_g$ in a finite proper time and will continue on inward to smaller radii. The time component, $U^0{}_{;\alpha}U^\alpha = 0$, of the geodesic equation, on the other hand, tells us

$$\frac{d^2t}{d\tau^2} = -\left(\frac{dr}{d\tau}\right)\left(\frac{dt}{d\tau}\right)\frac{r_g}{r^2}\left(1 - \frac{r_g}{r}\right)^{-1}, \tag{6.38}$$

which again shows something peculiar about $r = r_g$. A simultaneous solution of Eqs. (6.37) and (6.38) (simplest case: $dr/d\tau = 0$ at $r = \infty$) reveals in fact that $t \to \infty$ as $r \to r_g$. The falling particle reaches $r = r_g$ at finite *proper* time τ but infinite *coordinate* time t. An interesting analogy can be drawn to the $Y =$ constant geodesic in the metric of Eq. (6.36). From $X = X_0$ to $X = \infty$ is an infinite X coordinate distance, but the real distance, i.e., $\int ds$, is $2(X_0)^{-1/2}$. Reasoning by analogy we are tempted to conclude that the Schwarzschild coordinates are indeed poorly suited to the description of the geometry at the Schwarzschild radius.

The real proof that the Schwarzschild coordinates are the problem is the presentation of a coordinate system in which the metric coefficients are well behaved at $r = r_g$. Several such coordinate systems have indeed been found. These coordinate systems all retain the θ,ϕ coordinates that are natural coordinates for spherical symmetry, but replace r and t. The coordinate system in which the real nature of $r = r_g$ is clearest is the Kruskal–Szekeres system[30] introduced in 1960. In this system r and t are replaced by u,v defined through the coordinate transformation

$$(r/r_g - 1)e^{r/r_g} = u^2 - v^2 , \tag{6.39a}$$

$$t = (r_g/c)\ln\left|\frac{u+v}{u-v}\right| . \tag{6.39b}$$

In these coordinates the metric for the Schwarzschild geometry takes on the form

$$ds^2 = \frac{4r_g^3}{r}e^{-r/r_g}(du^2 - dv^2) + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) , \tag{6.40}$$

where r is regarded not as a coordinate but rather as the function of u and v defined in Eq. (6.39a). This is straightforward to verify: When Eqs. (6.39) are used in Eq. (6.40) the metric of Eq. (6.13) is recovered, thus Eqs. (6.40) and (6.13) must represent the same space-time in different coordinates. The metric of Eq. (6.40) has no pathologies at $u = v$, i.e., at $r = r_g$, but only at $u^2 - v^2 = -1$, i.e., at $r = 0$. This demonstrates conclusively that the space-time *geometry* is nonpathological at $r = r_g$. The pathological appearance of the metric coefficients of Eq. (6.13) at $r = r_g$ is due to the pathological nature of r,t coordinates there.

There is a rather surprising feature of Eq. (6.40): for every value of r,t there are two values of u,v. For example, $r/r_g = 1.603\,545\,7\ldots$ and $ct/r_g = \ln 3 = 1.0986\ldots$ corresponds both to $u = 2$, $v = 1$ and to $u = -2$, $v = -1$. The relationship between u,v and r,t coordinates is shown in Fig. 9. The u,v coordinates are plotted horizontally and vertically and the curves for constant r and constant t are indicated; the dark hyperbolas at $v^2 - u^2 = 1$ represent $r = 0$, the boundary of the geometry.

A great convenience of Kruskal–Szekeres coordinates is that it is easy to identify graphically the nature of a radial

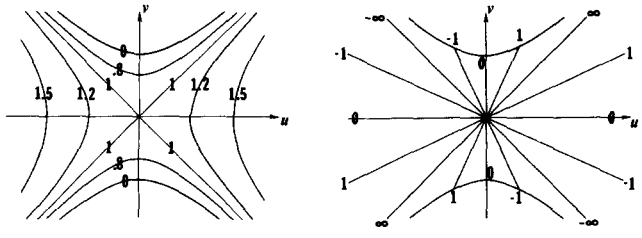

Fig. 9. Comparison of Schwarzschild and Kruskal–Szekeres coordinates. In (a) the hyperbolas representing constant radius curves are shown and are labeled with the value of r/r_g. In (b) straight lines representing constant t are shown and are labeled with the value of ct/r_g.

($d\theta = d\phi = 0$) displacement. It is clear from Eq. (6.40) that a radial displacement is timelike, lightlike, or spacelike if the ratio $|dv|/|du|$ is greater than, equal to, or less than unity. Figure 9 then shows why the original r,t coordinates are deceptive. For the $r < r_g$ region, t is a spacelike coordinate and r is a timelike coordinate! From Eq. (6.40) it is also clear that a displacement, radial or not, is spacelike if $|dv| < |du|$. Since particle and photon worldlines must be nowhere spacelike, *particle or photon worldlines drawn on the Kruskal–Szekeres graph must always make an angle less than 45° with the vertical.*

In Fig. 10 a worldline is drawn representing a particle falling inward from large radius in region I ($r \geqslant r_g$, $u \geqslant 0$) of the geometry, across $r = r_g$ into region II ($r \leqslant r_g$, $v \geqslant 0$). From this figure and the necessity for worldlines to be "more vertical than horizontal" we are forced to some fascinating conclusions. A particle once having moved inward across $r = r_g$ must inevitably reach $r = 0$. Regardless of its

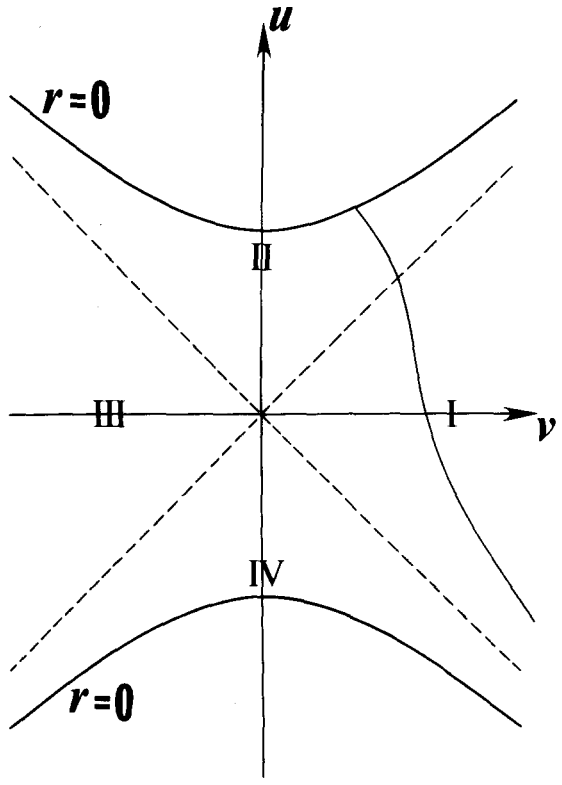

Fig. 10. Timelike worldline shown in a Kruskal–Szekeres graph. If the worldline represents the surface of a collapsing spherical body only the region of the graph to the right of the worldline applies.

acceleration it can never return to the $r > r_g$ region of space-time. Furthermore, any "information" (particles, photons, etc.) it emits also unavoidably proceeds to smaller and smaller radii and cannot ever reach the $r > r_g$ region of space-time. Since any event in region II will remain forever unknown to observers in the asymptotically flat region I, the $r = r_g$ boundary of region II is called an "event horizon," a horizon across which distant observers cannot see. The region interior to $r = r_g$ from which no particle, no information, and no light can emerge is called, with obvious justification, a "black hole."

It remains to explain the strange topology of spacetime suggested by the Kruskal–Szekeres graphs. Why are there *two* asymptotically flat regions, I and III, and why are there two $r < r_g$ regions, II and IV? First, we can take a no-nonsense common-sense view of the question and consider how a black hole forms. Ordinary stars, star clusters, galaxies, etc. are characterized by a surface radius R_S much bigger than $2GM/c^2$. For such an astronomical body we can draw the hyperbola in region I of Fig. 9 representing $r = R_S$ and ignore everything to the left. For $r < R_S$, points are inside the star, etc. and the Schwarzschild geometry (a *vacuum* solution of the GRT equations) does not apply. For a black hole to form, a star (or whatever) must collapse. Let us suppose that the worldline drawn in Fig. 10 is the worldline of the surface of a collapsing spherical star. Once the worldline passes $r = r_g$ (i.e., once the star has collapsed beyond the point that its surface radius equals $2GM/c^2$) it is clear from Fig. 10 that complete gravitational collapse to $r = 0$ is absolutely unavoidable. Every particle of the surface, after all moves on a worldline of the exterior Schwarzschild geometry. No matter what internal pressure forces are generated in the star it must completely collapse. But the worldline of the surface of the star is the boundary between matter and vacuum, so only the region of Fig. 10 to the right of the worldline is valid. Regions III and IV do not exist.

For astrophysical formation of black holes the issue of regions III and IV is evaded but the geometry of Eq. (6.40), for all values of u,v corresponding to $r > 0$, is after all a solution to the spherically symmetric field equations. Can it have any physical meaning? In principle, it can. We can imagine a universe in which the geometry with all four of its regions doesn't form but simply *is*. The spherically symmetric curved geometry must be built in from the beginning. The black hole must exist at all times in the universe (a "primordial black hole"). Such a universe would have two different asymptotically flat regions, I and III connected in some sense by the geometry in the region of small u and v. This connection, sometimes called the "Einstein–Rosen bridge," is related to a subject that is a frequent plot necessity in science fiction[31]: beating the speed-of-light limit on travel to distant stars.

The Einstein–Rosen bridge is not quite what a science fiction writer would want. It connects two essentially different asymptotically flat "universes"—or two region of the same universe, since regions I and III can be imagined to be connected also at large radius. It does not, however, provide a useful plot device. It is clear in the Kruskal–Szekeres graph that timelike lines cannot extend between regions I and III. Astronauts therefore cannot rocket across the Einstein–Rosen bridge to invade distant star systems.

We have been exclusively discussing so far only one par-

ticular black hole solution, the Schwarzschild black hole. A black hole can more generally be defined as a space-time geometry with a nonsingular event horizon, i.e., a geometrically smooth boundary across which information cannot pass outward to an asymptotically flat region. Solutions can also be found representing black holes with electric charge and with spin (angular momentum). These black hole solutions have space-time topologies even more fantastic than the four region structure of the Schwarzschild solution. As in the case of Schwarzschild black holes these exotic topologies are irrelevant if the black hole forms by collapse of an astronomical body. If the more general black holes are primordial features of a universe, however, there are bridges across which timelike world lines can pass from one asymptotically flat region to another. It is perhaps a sign of the good taste built into the mathematical structure of GRT that recent research[32] indicates that such bridges are unstable and cannot exist even for a primordial black hole.

ACKNOWLEDGMENTS

I would like to thank James Hartle for suggestions on the pedagogy of GRT. Helpful comments of "typical" readers were given by Steven Giddings and William Schwalm. Portions of this article were written with the support of National Science Foundation Grant PHY81-06909.

APPENDIX

Here we supply a derivation of Eq. (3.45), omitted in Sec. III C, and in the process justify a few other useful results. The pattern of proof, to be used several times is this: (i) A tensor equation is written down. Since the equation is tensorial its truth or falsity can be tested in any basis system. (ii) The most convenient system for such a test is the LF coordinate basis at some point. In this system the Christoffel symbols vanish (since $\mathbf{de}_\mu = 0$). (iii) With the truth of the equation established in one basis, it is established in all basis systems. Its implications in a general (non-LF) coordinate basis then lead to results about Christoffel symbols.

Our first task will be to verify Eq. (3.42). To this end we write

$$(\mathbf{grad\ V})_{\mu\nu} \equiv V_{\mu;\nu} = V_{\mu,\nu} - V_\lambda M^\lambda_{\mu\nu}$$

and seek to prove that M and Γ are the same. This can be done by considering the equation:

$$(\mathbf{grad\ A{\cdot}B})_\nu = (A_\nu B^\mu)_{;\nu} = A_{\mu;\nu}B^\mu + A_\mu B^\mu_{\ ;\nu} . \qquad \text{(A1)}$$

Now $(\mathbf{A{\cdot}B})$ is a scalar so the components of its gradient in *any* coordinate basis are

$$(\mathbf{grad\ A{\cdot}B})_\nu = (\mathbf{A{\cdot}B})_{,\nu} = (A_\mu B^\mu)_{,\nu} = A_{\mu,\nu}B^\mu + A_\mu B^\mu_{\ ,\nu} . \qquad \text{(A2)}$$

We now note that in an LF basis at a point, Eqs. (A1) and (A2) agree, hence Eq. (A1) must be true since it is tensorial. (True in one basis means true in all.) But Eq. (A1), expanded, tells us

$$(\mathbf{grad\ A{\cdot}B})_\nu = A_{\mu,\nu}B^\mu + A_\mu B^\mu{}_{,\nu} - A_\lambda M^\lambda_{\mu\nu} B^\mu + A_\mu B^\lambda \Gamma^\mu_{\lambda\nu} .$$

This can only be compatible with Eq. (A2), for all $\mathbf{A}$ and $\mathbf{B}$, if $M^\lambda_{\mu\nu} = \Gamma^\lambda_{\mu\nu}$ and, hence, if Eq. (3.42) is correct.

The next step is to use this result to examine

$$V_{\mu;\nu} - V_{\nu;\mu} = 0 \qquad \text{(A3)}$$

in the special case that $\mathbf{V}$ is the gradient of a scalar field Ψ:

$$V_\mu = (\mathbf{grad}\,\Psi)_\mu = \Psi_{,\mu} .$$

Equation (A3) then reads

$$\begin{aligned} V_{\mu;\nu} - V_{\nu;\mu} &= V_{\mu,\nu} - V_{\nu,\mu} - V_\lambda \Gamma^\lambda_{\mu\nu} + V_\lambda \Gamma^\lambda_{\nu\mu} \\ &= \Phi_{,\mu,\nu} - \Phi_{,\nu,\mu} + \Phi_{,\lambda}(\Gamma^\lambda_{\mu\nu} - \Gamma^\lambda_{\nu\mu}) \\ &= \Phi_{,\lambda}(\Gamma^\lambda_{\mu\nu} - \Gamma^\lambda_{\nu\mu}) . \end{aligned} \qquad \text{(A4)}$$

The left-hand side of Eq. (A3) is a tensor, hence, the final line of Eq. (A4) must be a tensor. This tensor vanishes in an LF coordinate basis and, hence, in any coordinate basis, and thus

$$\Gamma^\lambda_{\mu\nu} = \Gamma^\lambda_{\nu\mu} . \qquad \text{(A5)}$$

[This was already noted in Sec. III C, but as a *consequence* of Eq. (3.45), which still awaits proof.]

In Sec. III C it was shown [without the use of Eq. (3.45)] with an argument based on LF coordinates that the gradient of the metric tensor is always zero [Eq. (3.46)]. It follows that

$$\tfrac{1}{2}(g_{\beta\mu;\lambda} + g_{\mu\gamma;\beta} - g_{\gamma\beta;\mu}) = 0 .$$

When this is expanded in an arbitrary coordinate basis four of the six Γ's involved cancel [with the application of Eq. (A5)] and we are left with

$$\tfrac{1}{2}(g_{\beta\mu,\gamma} + g_{\mu\gamma,\beta} - g_{\gamma\beta,\mu}) = \tfrac{1}{2}g_{\sigma\mu}(\Gamma^\sigma_{\beta\gamma} + \Gamma^\sigma_{\gamma\beta}) .$$

We next multiply this equation by $g^{\alpha\mu}$ (and sum on μ) to find

$$\begin{aligned} \tfrac{1}{2}g^{\alpha\mu}(g_{\beta\mu,\gamma} + g_{\mu\gamma,\beta} - g_{\gamma\beta,\mu}) = g^{\alpha\mu}g_{\sigma\mu}\Gamma^\sigma_{\beta\gamma} &= \delta^\alpha_\sigma \Gamma^\sigma_{\beta\gamma} \\ &= \Gamma^\alpha_{\beta\gamma} , \end{aligned}$$

thereby proving Eq. (3.45).

For the Schwarzschild geometry [Eqs. (2.17e) and (6.13)] with coordinates r,θ,ϕ and $x^0 \equiv ct$, it is straightforward to compute the Christoffel symbols with Eq. (3.45). The only nonvanishing Γ's turn out to be

$$\Gamma^0_{0r} = \Gamma^0_{r0} = -\Gamma^r_{rr} = \frac{r_g}{2r^2}\left(1 - \frac{r_g}{r}\right)^{-1} ,$$

$$\Gamma^r_{00} = \frac{r_g}{2r^2}\left(1 - \frac{r_g}{r}\right) ,$$

$$\Gamma^\theta_{\theta r} = \Gamma^\theta_{r\theta} = \Gamma^\phi_{\phi r} = \Gamma^\phi_{r\phi} = r^{-1} ,$$

$$\Gamma^r_{\phi\phi} = -r\left(1 - \frac{r_g}{r}\right)\sin^2\theta ,$$

$$\Gamma^r_{\theta\theta} = -r\left(1 - \frac{r_g}{r}\right) , \quad \Gamma^\phi_{\phi\theta} = \Gamma^\phi_{\theta\phi} = \cot\theta ,$$

$$\Gamma^\theta_{\phi\phi} = -\sin\theta\cos\theta .$$

[1] An incomplete list of the more recent texts is Charles W. Misner, Kip S. Thorne, and John Archibald Wheeler, *Gravitation* (Freeman, San Francisco, 1973); Steven Weinberg, *Gravitation and Cosmology* (Wiley, New York, 1972); Ronald Adler, Maurice Bazin, and Menahem Schiffer, *Introduction to General Relativity* (McGraw-Hill, New York, 1965); Hans C. Ohanian, *Gravitation and Spacetime* (Norton, New York, 1976). Among the shorter or less detailed texts, Wolfgang Rindler, *Essential Relativity: Special, General, and Cosmological* (Van Nostrand, New York, 1969) is notable for the emphasis on physical ideas; J. Foster and J. D. Nightingale, *A Short Course in General Relativity* (Longman, London, 1979) gives a compact overview; the second half of L. D. Landau

and E. M. Lifschitz, *The Classical Theory of Fields*, 4th ed. (Pergamon, New York, 1975) contains a concise development of GRT.

[2]A somewhat similar approach is taken in Sec. 3.4 of Melvin Schwartz, *Principles of Electrodynamics* (McGraw-Hill, New York, 1972).

[3]Here the equations are given in the Lorentz gauge and we gloss over the possibility that the equations could have any other appearance. The issue of gauge transformations, not very relevant here, is discussed in Sec. V.

[4]Recent measurements have confirmed the weak equivalence principle to better than one part in 10^{12}. For a summary and discussion of experimental evidence see, e.g., Sec. 1.5 of Ohanian or Box 1.2 of Misner, Thorne, and Wheeler (Ref. 1). The "weak" equivalence principle tells us that the paths of nonspinning point particles cannot be used to distinguish locally between gravitational fields and accelerated frames. The "strong" equivalence principle states that no local physical measurement whatever can be used to make this distinction. Whether or not the strong equivalence principle is embodied in GRT is a somewhat semantic but controversial question. Contrast, for example, Ohanian Sec. 1.7 with Misner, Thorne, and Wheeler, Sec. 16.2.

[5]See, for example, R. P. Feynman, *Lectures on Gravitation*, unpublished lecture notes prepared by Fernando B. Morinigo and William G. Wagner, California Institute of Technology, 1963. For other references see Sec. 7.1 or Sec. 18.1 of Misner, Thorne, and Wheeler (Ref. 1).

[6]Clearly in the metric formula any antisymmetric part $(g_{\mu\nu} - g_{\nu\mu})$ of the metric coefficients is unimportant. In some unified field theories—theories that combine electromagnetism and gravity—the antisymmetric part of the metric coefficients is used to carry information about the electromagnetic field. Einstein first proposed such a theory thirty years ago. See, e.g., A. Einstein, *The Meaning of Relativity* (Princeton, New Jersey, 1955), Appendix II. There have been several modified version of such a theory proposed in the ensuing years but no wholly successful one.

[7]The reader expecting "minimum" rather than "extremal" deserves an explanation. In flat space-time straight worldlines are in fact the curves of *maximum* length (i.e., proper time) between two events with a timelike separation; an accelerated observer traveling between these two events always measures a smaller clock time. Straight lines between events with spacelike separation are the curves of minimum length. In any case, in curved space(time) it is not even true that geodesics are always curves of extremal length. On the surface of the Earth (idealized as a perfect sphere) any segment of a great circle is a geodesic. The ~40 000-km great circle to New York, starting northward from Montreal, is therefore a geodesic, though it is neither a minimum nor a maximum with respect to small deviations of the path. The geodesic equation is a necessary condition for an extremal path, but not a sufficient one.

[8]In view of this it is interesting that some physics teachers consider the expression "centrifugal force" to be obscene, but have no reluctance to talk about weight.

[9]It is difficult to give simple examples of the difference of the Ricci and the Riemann measures of curvature. Curvature is simply visualizable only for two-dimensional spaces. But in two- and three-dimensional space(time)s it turns out that Riemann curvature must vanish if Ricci curvature vanishes. We live in the smallest number (4) of dimensions in which GRT can work.

[10]The almost exclusive use of coordinate basis systems is a concession to the length of this article. Many results are more easily achieved or understood with noncoordinate basis systems. It should be noted that the usual basis systems of physicists, for spherical or cylindrical coordinates in flat three-dimensional space, are orthonormal systems, rather than coordinate basis systems.

[11]Some authors, especially of older works, use $\{^{\mu}_{\alpha\beta}\}$ in place of $\Gamma^{\mu}_{\alpha\beta}$.

[12]An example suffices to prove this: In two-dimensional flat space the basis vectors $\mathbf{e}_\theta$, $\mathbf{e}_r$ are not constant and so the values $\Gamma^{r}_{\theta\theta}$, etc. cannot all be zero. If these were tensor components they could not then all be zero in *any* basis. But the Cartesian basis vectors $\mathbf{e}_x$, $\mathbf{e}_y$ for two-dimensional flat space are constant and hence Γ^{x}_{xy}, etc., are all zero.

[13]Some authors use a vertical bar rather than semicolon ($V^{\nu}{}_{|\alpha}$).

[14]More generally, an affine parameter is one that puts the geodesic equation in the form of Eq. (4.5). If a nonaffine parameter, e.g., $\lambda = \tau^2$, is used extra terms are generated as in the right-hand side of Eq. (2.30). Proper time is always an affine parameter for timelike curves.

[15]The left-hand side of this equation is often called the components of the Einstein tensor $\mathbf{G}$, that is $G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R$.

[16]The influence of curvature on particle dynamics may be summarized in the statement that there is no local influence on the dynamics of (nonspinning) particles. Particle dynamics, however, is not all of dynamics. We need also to know how fields (e.g., the electromagnetic field) are influenced by curvature. The answer is the same: no local influence of curvature. This general lack of gravitational effects on the dynamics of all (nongravitational) forms of stress-energy turns out to be mathematically summarizes by $T^{\mu\nu}{}_{;\nu} = 0$.

[17]This is somewhat similar to the situation in electromagnetism where a gauge choice for $\mathbf{A}$ must be made before a solution for $\mathbf{A}$ can be found. The constraints imposed on the $g_{\mu\nu}$ are also called a gauge choice. A particular gauge choice in the case of weak fields is discussed in Sec. V.

[18]The Newtonian limit in this case requires an additional assumption, that mass-energy density is much larger than the stresses in the source. In a microscopic picture this turns out to mean among other things that the particles constituting the material of the source are moving at nonrelativistic velocities, and the sound speed in the material is small compared to c.

[19]At this point the traditional approach would be first to make a further simplification in Eq. (5.21). The gauge freedom is not completely exhausted by Eq. (5.19), nor is it exhausted in electrodynamics by Eq. (5.17). The remaining gauge freedom can be used to simplify the mathematical description of the waves by fixing all the components of $\bar{h}_{\mu\nu}$ to vanish except $\bar{h}_{xy}$ and $\bar{h}_{xx} = -\bar{h}_{yy}$. [For details see Landau and Lifschitz, Sec. 108, or Misner, Thorne, and Wheeler, Sec. 35.4 (Ref. 1).] We shall not need this simplification.

[20]Richard A. Isaacson, Phys. Rev. **166**, 1272 (1968). See also the discussion in Secs. (35.13)–(35.15) of Misner, Thorne, and Wheeler (Ref. 1).

[21]Karl Schwarzschild, Sitzber. Deut. Akad. Wiss. Berlin Kl. Math.-Phys. 189 (1916).

[22]The symmetry argument here is meant to be plausible and persuasive; omission of detail and rigor here are concessions to conciseness. A more complete discussion can be found in Misner, Thorne, and Wheeler, Box 23.3 (Ref. 1).

[23]For example, any function of r could serve just as well as a radial coordinate. One such change in radius results in "isotropic coordinates," which simplify several calculations [see, e.g., pp. 284–285 of Ohanian (Ref. 1)]. A more general "mixing" of time and radius is involved in the Kruskal–Szekeres coordinates, introduced in Sec. VI D.

[24]This is accomplished with a coordinate transformation $T = F(T', r)$. If F is chosen to satisfy $F_{,r} = C/2A$ no $dT'\,dr$ appears in the metric formula and T' is our "special" time coordinate.

[25]The nonappearance of x^{α} in the metric functions is an indication that the geometry is symmetric in some way. A formalism, that of "Killing vectors," exists for describing such symmetries without the need of a special coordinate system. See, e.g., Sec. 13.1 of Weinberg (Ref. 1).

[26]For example, see Jerry B. Marion, *Classical Dynamics of Particles and Systems* (Academic, New York, 1965), Sec. 10.10.

[27]Pierre-Simon Laplace, *Le Systéme du monde* (Paris, 1796), Vol. II.

[28]Laplace's argument cannot be taken too seriously since it is a prediction about light in a strong gravitational field, made in ignorance of SRT restrictions on light and of GRT modifications of strong fields. Laplace's result for the "no-escape radius" seems less impressive when it is realized that GM/c^2 is the only combination of parameters that is dimensionally a radius. His "precisely correct" prediction $2GM/c^2$ for the no-escape radius agrees with GRT only when *radius* is interpreted as "Schwarzschild radial coordinate." With another useful radial coordinate (that of isotropic coordinates) the no-escape radius is $\frac{1}{2}GM/c^2$.

[29]It might seem that a coordinate pathology could be distinguished from a geometric pathology with an examination of the Riemann tensor, a coordinate independent quantity. The difficulty with this viewpoint is that we calculate *components* of the Riemann tensor. If the coordinate system itself is pathological we might expect the coordinate basis vectors and, hence, tensor components to be pathological, even if the tensor itself is unrelated to a geometric or physical pathology. The singular behavior of the Riemann component in Eq. (4.23) is, therefore, not strong evidence for a coordinate singularity at r_g. There is a way around this difficulty, in principle. A "foolproof" recipe can be given for con-

structing locally flat coordinates at a point. If the recipe succeeds it produces coordinates guaranteed to be nonpathological near the point. If it fails it unambiguously signals the presence of a geometric singularity at the point. Such a method has been used to show that $r = 0$ is truly a curvature singularity. [Lawrence Mysak and George Szekeres, Can. J. Phys. **44**, 617 (1965)]. To deal with $r = r_g$ there are easier ways.

[30]Discovered independently by M. D. Kruskal, Phys. Rev. **119**, 1743 (1960) and G. Szekeres, Publ. Mat. Debrecen **7**, 285 (1960).

[31]Sometimes difficult to distinguish from "science" in black hole physics.

[32]Yekta Gürsel, Igor D. Novikov, Vernon D. Sandberg, and A. A. Starobinsky, Phys. Rev. D **20**, 1260 (1979); J. M. McNamara, Proc. R. Soc. London A **358**, 499 (1978).

Energy waste in a university building

Neil J. Numark and Albert A. Bartlett
Department of Physics, Box 390, University of Colorado, Boulder, Colorado 80309

(Received 19 March 1981; accepted for publication 16 June 1981)

Interesting physics problems that can be used as examples in introductory physics courses relating to the waste of thermal energy can be found in the mechanical systems of campus buildings. The design of these wasteful systems may represent the "state of the art" as it existed just a few years ago, so such examples are probably abundant. Our Student Recreation Center was opened in 1973. It has an ice skating rink with the associated large refrigeration system. Simple calculations using elementary thermodynamics applied to this system show that the heat rejected by the system is roughly a quarter of a megawatt, which is approximately the average thermal power needed to heat water for the showers in the building. An outcome of this student project is the recommendation that the rejected heat be used to heat (or preheat) the shower water at an estimated annual saving of $40 000 in current energy costs.

Mechanical systems for the heating and cooling of buildings consume large quantities of energy. Until recently it was thought to be less expensive to build systems that waste energy than it was to invest in systems that conserve and reuse energy. The rapid escalation of energy prices has changed all this and the earlier systems that waste energy are now prime candidates for energy saving redesign and retrofit. The physics involved in understanding the systems, how they waste energy, and what is needed to improve them, is elementary and can be used very effectively in beginning physics classes.

An example is the Student Recreation Building of the University of Colorado in Boulder (Fig. 1). This new building contains all manner of recreation facilities including swimming pools and an ice skating rink (26×56 m). We wish to focus our attention on the large refrigeration system that maintains the ice in the skating rink. Three compressors, each driven by a 56-kW electric motor (75 hp) are the heart of the refrigeration system. Typically only two are running at any one time at an estimated 90% of peak capacity and the third is on standby. Thus the steady-state electrical power consumption of this system is approximately $2 \times 0.9 \times 75 \times 0.746 = 1.0 \times 10^2$ kW. The energy flow diagram is shown in Fig. 2. The power extracted from cold reservoir is P_c, the power delivered by motors to operate the refrigeration cycle is $P_e = 1.0 \times 10^2$ kW, and P_h is the power delivered to the warm reservoir. The numerical calculations that follow are very rough ($\pm 20\%$) so we can ignore the power loss in motors. From the first law of thermodynamics,

$$P_h = P_c + P_e. \tag{1}$$

The coefficient of performance C_p of a refrigeration system is

$$C_p = P_c/P_e = P_c/(P_h - P_c). \tag{2}$$

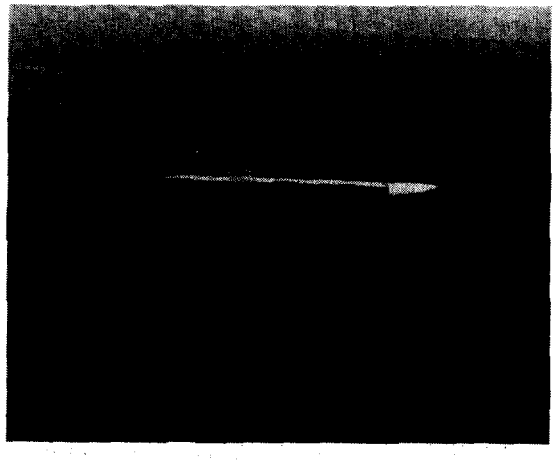

Fig. 1. Student Recreation Center of the University of Colorado at Boulder. This building was completed in 1973 at a cost of $4.9 million. The location of the snow on the roof of the building is an interesting thermal pattern. Even though the roof has some insulation, snow has melted everywhere except on that part of the roof that is directly over the ice skating rink. The faint cloud of steam that is drifting to the left is from the waste heat that is discussed in this article.

