Contents

1 Physics in Euclidean Space and Flat Spacetime: Geometric Viewpoint 1

1.1 [N & R] Overview . . . 1
1.2 Foundational Concepts . . . 4
1.2.1 [N] Newtonian Concepts . . . 4
1.2.2 [R] Special Relativistic Concepts: Inertial frames, inertial coordinates, events, vectors, and spacetime diagrams . . . 5
1.2.3 [R] Special Relativistic Concepts: Principle of Relativity; the Interval and its Invariance . . . 8
1.3 [N & R] Tensor Algebra Without a Coordinate System . . . 14
1.4 Particle Kinetics and Lorentz Force Without a Reference Frame . . . 16
1.4.1 [N] Newtonian Particle Kinetics . . . 16
1.4.2 [R] Relativistic Particle Kinetics: World Lines, 4-Velocity, 4-Momentum and its Conservation, 4-Force . . . 17
1.4.3 [R] Geometric Derivation of the Lorentz Force Law . . . 20
1.5 Component Representation of Tensor Algebra . . . 21
1.5.1 [N] Euclidean 3-space . . . 21
1.5.2 [R] Minkowski Spacetime . . . 23
1.5.3 [N & R] Slot-Naming Index Notation . . . 25
1.6 [R] Particle Kinetics in Index Notation and in a Lorentz Frame . . . 27
1.7 Orthogonal and Lorentz Transformations of Bases, and Spacetime Diagrams 32
1.7.1 [N] Euclidean 3-space: Orthogonal Transformations . . . 32
1.7.2 [R] Minkowski Spacetime: Lorentz Transformations . . . 34
1.7.3 [R] Spacetime Diagrams for Boosts . . . 35
1.8 [R] Time Travel . . . 38
1.9 [N & R] Directional Derivatives, Gradients, Levi-Civita Tensor, Cross Product and Curl . . . 41
1.10 [R] Nature of Electric and Magnetic Fields; Maxwell's Equations . . . 44
1.11 Volumes, Integration, and Integral Conservation Laws . . . 48
1.11.1 [N] Newtonian Volumes and Integration . . . 48
1.11.2 [R] Spacetime Volumes and Integration . . . 49
1.11.3 [R] Conservation of Charge in Spacetime . . . 51
1.11.4 [R] Conservation of Particles, Baryons and Rest Mass . . . 53
1.12 The Stress-Energy Tensor and Conservation of 4-Momentum . . . 56
1.12.1 [N] Newtonian Stress Tensor and Momentum Conservation . . . 56

1.12.2 [R] Relativistic Stress-Energy Tensor 58
1.12.3 [R] 4-Momentum Conservation 60
1.12.4 [R] Stress-Energy Tensors for Perfect Fluid and Electromagnetic Field 61

Chapter 1

Physics in Euclidean Space and Flat Spacetime: Geometric Viewpoint

Version 0801.1.K by Kip, 1 October 2008

Please send comments, suggestions, and errata via email to kip@tapir.caltech.edu, or on paper to Kip Thorne, 130-33 Caltech, Pasadena CA 91125

Box 1.1
Reader's Guide

- Sections and exercises labeled [N] are Newtonian, those labeled [R] are relativistic. The N material can be read without the R material, but the R material requires the N material as a foundation.
- Readers who plan to study only the non-relativistic portions of this book should learn this book's geometric viewpoint on Newtonian physics and some mathematical tools we shall use by reading or browsing the [N] sections. They will also need to know two items of relativity; see Box 1.3.
- The R sections are a self-contained introduction to special relativity, though it will help if the reader has already had an elementary introduction.
- Readers who already know special relativity well should browse this chapter, especially Secs. 1.1–1.4, 1.5.3, 1.9–1.12, to learn this book's geometric viewpoint and a few concepts (such as the stress-energy tensor) that they might not have met.

1.1 [N & R] Overview

In this book, we shall adopt a different viewpoint on the laws of physics than that found in most elementary texts. In elementary textbooks, the laws are expressed in terms of

quantities (locations in space or spacetime, momenta of particles, etc.) that are measured in some coordinate system or reference frame. For example, Newtonian vectorial quantities (momenta, electric fields, etc.) are triplets of numbers [e.g., $(1, 9, -4)$] representing the vectors' components on the axes of a spatial coordinate system, and relativistic 4-vectors are quadruplets of numbers representing components on the spacetime axes of some reference frame.

By contrast, in this book, we shall express all physical quantities and laws in a **geometric form**, i.e. a form that is **independent of any coordinate system or reference frame**. For example, in Newtonian physics, momenta and electric fields will be vectors described as arrows that live in the 3-dimensional, flat Euclidean space of everyday experience. They require no coordinate system at all for their existence or description—though sometimes coordinates will be useful. We shall state physical laws, e.g. the Lorentz force law, as geometric (i.e. coordinate-free) relationships between these geometric (i.e. coordinate-independent) quantities.

By adopting this geometric viewpoint, we shall gain great conceptual power and often also computational power. For example, when we ignore experiment and simply ask what forms the laws of physics can possibly take (what forms are allowed by the requirement that the laws be geometric), we shall find remarkably little freedom. Coordinate independence strongly constrains the laws (see, e.g., Sec. 1.4 below). This power, together with the elegance of the geometric formulation, suggests that in some deep (ill-understood) sense, Nature's physical laws *are* geometric and have nothing whatsoever to do with coordinates or reference frames.

The mathematical foundation for our geometric viewpoint is *differential geometry* (also called "tensor analysis" by physicists). This differential geometry can be thought of as an extension of the vector analysis with which all readers should be familiar.

There are three different frameworks for the classical physical laws that scientists use, and correspondingly three different geometric arenas for the laws; cf. Fig. 1.1. *General relativity* is the most accurate classical framework; it formulates the laws as geometric relationships in the arena of *curved 4-dimensional spacetime*. *Special relativity* is the limit of general relativity in the complete absence of gravity; its arena is *flat, 4-dimensional Minkowski spacetime*[1]. *Newtonian physics* is the limit of general relativity when (i) gravity is weak but not necessarily absent, (ii) relative speeds of particles and materials are small compared to the speed of light c, and (iii) all stresses (pressures) are small compared to the total density of mass-energy; its arena is *flat, 3-dimensional Euclidean space* with time separated off and made universal (by contrast with the frame-dependent time of relativity).

In Parts I–V of this book (statistical physics, optics, elasticity theory, fluid mechanics, plasma physics) we shall confine ourselves to the Newtonian and special relativistic formulations of the laws, and accordingly our arenas will be flat Euclidean space and flat Minkowski spacetime. In Part VI we shall extend many of the laws we have studied into the domain of strong gravity (general relativity), i.e., the arena of curved spacetime.

In Parts I and II (statistical physics and optics), in addition to confining ourselves to flat space or flat spacetime, we shall avoid any sophisticated use of curvilinear coordinates; i.e., when using coordinates in nontrivial ways, we shall confine ourselves to Cartesian coordinates

[1] so-called because it was Hermann Minkowski (1908) who identified the special relativistic invariant interval as defining a metric in spacetime, and who elucidated the resulting geometry of flat spacetime.

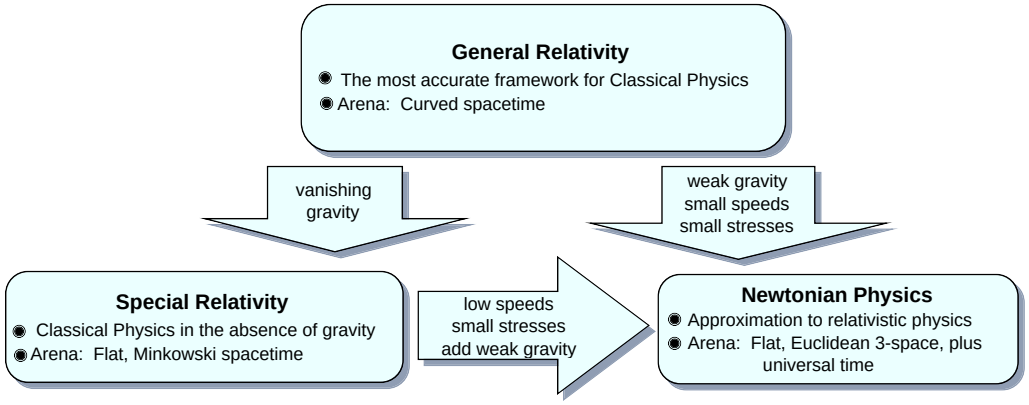


Fig. 1.1: The three frameworks and arenas for the classical laws of physics, and their relationship to each other.

in Euclidean space, and Lorentz coordinates in Minkowski spacetime. This chapter is an introduction to all the differential geometric tools that we shall need in these limited arenas.

In Parts III, IV, and V, when studying elasticity theory, fluid mechanics, and plasma physics, we will use curvilinear coordinates in nontrivial ways. As a foundation for them, at the beginning of Part III we will extend our flat-space differential geometric tools to curvilinear coordinate systems (e.g. cylindrical and spherical coordinates). Finally, at the beginning of Part VI, we shall extend our geometric tools to the arena of curved spacetime.

In this chapter we shall alternate back and forth, one section after another, between the laws of physics and flat-space differential geometry, using each to illustrate and illuminate the other. We begin in Sec. 1.2 by recalling the foundational concepts of Newtonian physics and of special relativity. Then in Sec. 1.3 we develop our first set of differential geometric tools: the tools of coordinate-free tensor algebra. In Sec. 1.4 we illustrate our tensor-algebra tools by using them to describe—without any coordinate system or reference frame whatsoever—the kinematics of point particles that move through the Euclidean space of Newtonian physics and through relativity's Minkowski spacetime; the particles are allowed to collide with each other and be accelerated by an electromagnetic field. In Sec. 1.5, we extend the tools of tensor algebra to the domain of Cartesian and Lorentz coordinate systems, and then in Sec. 1.6 we use these extended tensorial tools to restudy the motions, collisions, and electromagnetic accelerations of particles. In Sec. 1.7 we discuss rotations in Euclidean space and Lorentz transformations in Minkowski spacetime, and we develop relativistic spacetime diagrams in some depth and use them to study such relativistic phenomena as length contraction, time dilation, and simultaneity breakdown. In Sec. 1.8 we illustrate the tools we have developed by asking whether the laws of relativity permit a highly advanced civilization to build time machines for traveling backward in time as well as forward. In Sec. 1.9 we develop additional differential geometric tools: directional derivatives, gradients, and the Levi-Civita tensor, and in Sec. 1.10 we use these tools to discuss Maxwell's equations and the geometric nature of electric and magnetic fields. In Sec. 1.11 we develop our final set of geometric tools: volume elements and the integration of tensors over spacetime, and in Sec. 1.12 we use these tools to define the stress tensor of Newtonian physics and relativity's stress-energy tensor, and to formulate very general versions of the conservation of 4-momentum.

1.2 Foundational Concepts

1.2.1 [N] Newtonian Concepts

The arena for the Newtonian laws is a spacetime composed of the familiar 3-dimensional Euclidean space of everyday experience (which we shall call *3-space*), and a universal time t. We shall denote points (locations) in 3-space by capital script letters such as $\mathcal{P}$ and $\mathcal{Q}$. These points and the 3-space in which they live require no coordinates for their definition.

A *scalar* is a single number that we associate with a point, $\mathcal{P}$, in 3-space. We are interested in scalars that represent physical quantities, e.g., temperature T. When a scalar is a function of location $\mathcal{P}$ in space, e.g. $T(\mathcal{P})$, we call it a *scalar field*.

A *vector* in Euclidean 3-space can be thought of as a straight arrow that reaches from one point, $\mathcal{P}$, to another, $\mathcal{Q}$ (e.g., the arrow $\Delta\mathbf{x}$ of Fig. 1.2a). Equivalently, $\Delta\mathbf{x}$ can be thought of as a direction at $\mathcal{P}$ and a number, the vector's length. Sometimes we shall select one point $\mathcal{O}$ in 3-space as an "origin" and identify all other points, say $\mathcal{Q}$ and $\mathcal{P}$, by their vectorial separations $\mathbf{x}_\mathcal{Q}$ and $\mathbf{x}_\mathcal{P}$ from that origin.

The Euclidean distance $\Delta\sigma$ between two points $\mathcal{P}$ and $\mathcal{Q}$ in 3-space can be measured with a ruler and so, of course, requires no coordinate system for its definition. (If one does have a Cartesian coordinate system, it can be computed by the Pythagorean formula, a precursor to the "invariant interval" of flat spacetime, Sec. 1.2.3.) This distance $\Delta\sigma$ is also the length $|\Delta\mathbf{x}|$ of the vector $\Delta\mathbf{x}$ that reaches from $\mathcal{P}$ to $\mathcal{Q}$, and the square of that length is denoted

$$|\Delta\mathbf{x}|^2 \equiv (\Delta\mathbf{x})^2 \equiv (\Delta\sigma)^2 \ . \tag{1.1}$$

Of particular importance is the case when $\mathcal{P}$ and $\mathcal{Q}$ are neighboring points and $\Delta\mathbf{x}$ is a differential (infinitesimal) quantity $d\mathbf{x}$. By traveling along a sequence of such $d\mathbf{x}$'s, laying them down tail-at-tip, one after another, we can map out a curve to which these $d\mathbf{x}$'s are tangent (Fig. 1.2b). The curve is $\mathcal{P}(\lambda)$, with λ a parameter along the curve; and the infinitesimal vectors that map it out are $d\mathbf{x} = (d\mathcal{P}/d\lambda)d\lambda$.

The product of a scalar with a vector is still a vector; so if we take the change of location $d\mathbf{x}$ of a particular element of a fluid during a (universal) time interval dt, and multiply it by $1/dt$, we obtain a new vector, the fluid element's velocity $\mathbf{v} = d\mathbf{x}/dt$, at the fluid element's location $\mathcal{P}$. Performing this operation at every point $\mathcal{P}$ in the fluid defines the velocity field $\mathbf{v}(\mathcal{P})$. Similarly, the sum (or difference) of two vectors is also a vector and so taking the

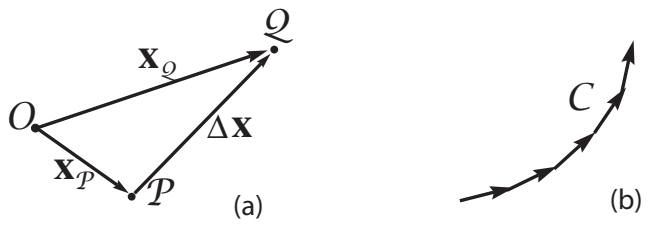


Fig. 1.2: (a) A Euclidean 3-space diagram depicting two points $\mathcal{P}$ and $\mathcal{Q}$, their vectorial separations $\vec{x}_\mathcal{P}$ and $\vec{x}_\mathcal{Q}$ from the (arbitrarily chosen) origin $\mathcal{O}$, and the vector $\Delta\mathbf{x} = \mathbf{x}_\mathcal{Q} - \mathbf{x}_\mathcal{P}$ connecting them. (b) A curve $\mathcal{C}$ generated by laying out a sequence of infinitesimal vectors, tail-to-tip.

difference of two velocity measurements at times separated by dt and multiplying by $1/dt$ generates the acceleration $\mathbf{a} = d\mathbf{v}/dt$. Multiplying by the fluid element's (scalar) mass m gives the force $\mathbf{F} = m\mathbf{a}$ that produced the acceleration; dividing an electrically produced force by the fluid element's charge q gives another vector, the electric field $\mathbf{E} = \mathbf{F}/q$, and so on. We can define inner products [Eq. (1.9a) below] of pairs of vectors at a point (e.g., force and displacement) to obtain a new scalar (e.g., work), and cross products [Eq. (1.60a)] of vectors to obtain a new vector (e.g., torque). By examining how a differentiable scalar field changes from point to point, we can define its gradient [Eq. (1.54b)]. In this fashion, which should be familiar to the reader and will be elucidated and generalized below, we can construct all of the standard scalars and vectors of Newtonian physics. What is important is that *these physical quantities require no coordinate system for their definition.* They are geometric (coordinate-independent) objects residing in Euclidean 3-space at a particular time.

It is a fundamental (though often ignored) principle of physics that *the Newtonian physical laws are all expressible as geometric relationships between these types of geometric objects, and these relationships do not depend upon any coordinate system or orientation of axes, nor on any reference frame* (on any purported velocity of the Euclidean space in which the measurements are made).[2] We shall return to this principle throughout this book. It is the Newtonian analog of Einstein's Principle of Relativity (Sec. 1.2.3 below).

1.2.2 [R] Special Relativistic Concepts: Inertial frames, inertial coordinates, events, vectors, and spacetime diagrams

Because the nature and geometry of Minkowski spacetime are far less obvious intuitively than those of Euclidean 3-space, we shall need a crutch in our development of the Minkowski foundational concepts. That crutch will be inertial reference frames. We shall use them to develop in turn the following frame-independent Minkowski-spacetime concepts: events, 4-vectors, the principle of relativity, geometrized units, the interval and its invariance, and spacetime diagrams.

An *inertial reference frame* is a (conceptual) three-dimensional latticework of measuring rods and clocks (Fig. 1.3) with the following properties: (*i*) The latticework moves freely through spacetime (i.e., no forces act on it), and is attached to gyroscopes so it does not rotate with respect to distant, celestial objects. (*ii*) The measuring rods form an orthogonal lattice and the length intervals marked on them are uniform when compared to, e.g., the wavelength of light emitted by some standard type of atom or molecule; and therefore the rods form an orthonormal, Cartesian coordinate system with the coordinate x measured along one axis, y along another, and z along the third. (*iii*) The clocks are densely packed throughout the latticework so that, ideally, there is a separate clock at every lattice point. (*iv*) The clocks tick uniformly when compared, e.g., to the period of the light emitted by some standard type of atom or molecule; i.e., they are *ideal clocks*. (*v*) The clocks are synchronized by the Einstein synchronization process: If a pulse of light, emitted by one of the clocks, bounces off a mirror attached to another and then returns, the time of bounce t_b

[2] By changing the velocity of Euclidean space, one adds a constant velocity to all particles, but this leaves the laws, e.g. Newton's $\mathbf{F} = m\mathbf{a}$, unchanged.

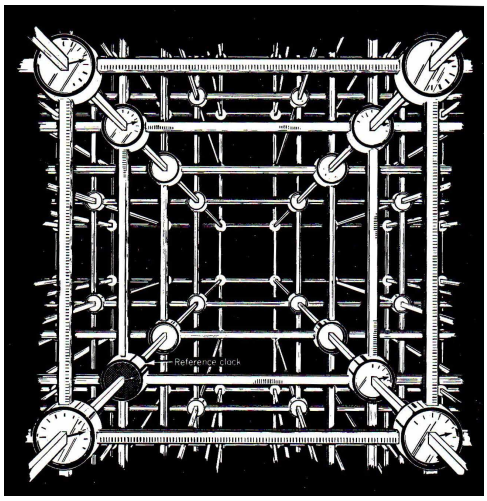


Fig. 1.3: An inertial reference frame. From Taylor and Wheeler (1992).

as measured by the clock that does the bouncing is the average of the times of emission and reception as measured by the emitting and receiving clock: $t_b = \frac{1}{2}(t_e + t_r)$.[3]

Our second fundamental relativistic concept is the *event*. An event is a precise location in space at a precise moment of time; i.e., a precise location (or "point") in 4-dimensional spacetime. We sometimes will denote events by capital script letters such as $\mathcal{P}$ and $\mathcal{Q}$ — the same notation as for points in Euclidean 3-space; there need be no confusion, since we will avoid dealing with 3-space points and Minkowski-spacetime points simultaneously.

A *4-vector* (also often referred to as a *vector in spacetime*) is a straight arrow $\Delta\vec{x}$ reaching from one event $\mathcal{P}$ to another $\mathcal{Q}$; equivalently, $\Delta\vec{x}$ is a direction in spacetime at the event $\mathcal{P}$ where it lives, together with a number that tell us its length. We often will deal with 4-vectors and ordinary (3-space) vectors simultaneously, so we shall need different notations for them: bold-face Roman font for 3-vectors, $\Delta\mathbf{x}$, and arrowed italic font for 4-vectors, $\Delta\vec{x}$. Sometimes we shall identify an event $\mathcal{P}$ in spacetime by its vectorial separation $\vec{x}_{\mathcal{P}}$ from some arbitrarily chosen event in spacetime, the "origin" $\mathcal{O}$.

An inertial reference frame provides us with a coordinate system for spacetime. The coordinates $(x^0, x^1, x^2, x^3) = (t, x, y, z)$ which it associates with an event $\mathcal{P}$ are $\mathcal{P}$'s location (x, y, z) in the frame's latticework of measuring rods, and the time t of $\mathcal{P}$ *as measured by the clock that sits in the lattice at the event's location.* (Many apparent paradoxes in special relativity result from failing to remember that the time t of an event is always measured by a clock that resides at the event, and never by clocks that reside elsewhere in spacetime.)

It is useful to depict events on *spacetime diagrams*, in which the time coordinate $t = x^0$ of some inertial frame is plotted upward, and two of the frame's three spatial coordinates, $x = x^1$ and $y = x^2$, are plotted horizontally. Figure 1.4 is an example. Two events $\mathcal{P}$ and $\mathcal{Q}$ are shown there, along with their vectorial separations $\vec{x}_{\mathcal{P}}$ and $\vec{x}_{\mathcal{Q}}$ from the origin and the vector $\Delta\vec{x} = \vec{x}_{\mathcal{Q}} - \vec{x}_{\mathcal{P}}$ that separates them from each other. The coordinates of $\mathcal{P}$ and $\mathcal{Q}$,

[3]For a deeper discussion of the nature of ideal clocks and ideal measuring rods see, e.g., pp. 23–29 and 395–399 of Misner, Thorne, and Wheeler (1973).

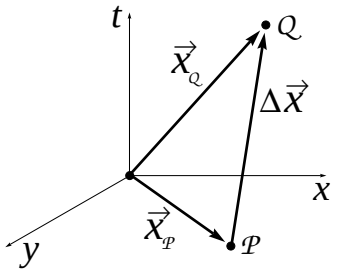


Fig. 1.4: A spacetime diagram depicting two events $\mathcal{P}$ and $\mathcal{Q}$, their vectorial separations $\vec{x}_{\mathcal{P}}$ and $\vec{x}_{\mathcal{Q}}$ from an (arbitrarily chosen) origin, and the vector $\Delta\vec{x} = \vec{x}_{\mathcal{Q}} - \vec{x}_{\mathcal{P}}$ connecting them. The laws of physics cannot involve the arbitrary origin $\mathcal{O}$; we introduce it only as a conceptual aid.

which are the same as the components of $\vec{x}_{\mathcal{P}}$ and $\vec{x}_{\mathcal{Q}}$ in this coordinate system, are ($t_{\mathcal{P}}$, $x_{\mathcal{P}}$, $y_{\mathcal{P}}$, $z_{\mathcal{P}}$) and ($t_{\mathcal{Q}}$, $x_{\mathcal{Q}}$, $y_{\mathcal{Q}}$, $z_{\mathcal{Q}}$); and correspondingly, the components of $\Delta\vec{x}$ are

$$\begin{aligned} \Delta x^0 = \Delta t = t_{\mathcal{Q}} - t_{\mathcal{P}}, \quad & \Delta x^1 = \Delta x = x_{\mathcal{Q}} - x_{\mathcal{P}}, \\ \Delta x^2 = \Delta y = y_{\mathcal{Q}} - y_{\mathcal{P}}, \quad & \Delta x^3 = \Delta z = z_{\mathcal{Q}} - z_{\mathcal{P}} \ . \end{aligned} \tag{1.2}$$

We shall denote these components of $\Delta\vec{x}$ more compactly by Δx^α, where the α index (and every other lower case Greek index that we shall encounter) takes on values $t = 0$, $x = 1$, $y = 2$, and $z = 3$. Similarly, in 3-dimensional Euclidean space, we shall denote the Cartesian components $\Delta\mathbf{x}$ of a vector separating two events by Δx^j, where the j (and every other lower case Latin index) takes on the values $x = 1$, $y = 2$, and $z = 3$.

When the physics or geometry of a situation being studied suggests some preferred inertial frame (e.g., the frame in which some piece of experimental apparatus is at rest), then we typically will use as axes for our spacetime diagrams the coordinates of that preferred frame. On the other hand, when our situation provides *no* preferred inertial frame, or when we wish to emphasize a frame-independent viewpoint, we shall use as axes the coordinates of a completely arbitrary inertial frame and we shall think of the spacetime diagram as depicting spacetime in a coordinate-independent, frame-independent way.

The coordinate system (t, x, y, z) provided by an inertial frame is sometimes called an *inertial coordinate system*, and sometimes a *Minkowski coordinate system* (a term we shall not use), and sometimes a *Lorentz coordinate system* [because it was Lorentz (1904) who first studied the relationship of one such coordinate system to another, the Lorentz transformation]. We shall use the terms "Lorentz coordinate system" and "inertial coordinate system" interchangeably, and we shall also use the term *Lorentz frame* interchangeably with *inertial frame*. A physicist or other intelligent being who resides in a Lorentz frame and makes measurements using its latticework of rods and clocks will be called an *observer*.

Although events are often described by their coordinates in a Lorentz reference frame, and vectors by their components (coordinate differences), it should be obvious that the concepts of an event and a vector need not rely on any coordinate system whatsoever for their definition. For example, the event $\mathcal{P}$ of the birth of Isaac Newton, and the event $\mathcal{Q}$ of the birth of Albert Einstein are readily identified without coordinates. They can be regarded as *points* in spacetime, and their separation vector is the straight arrow reaching through

spacetime from $\mathcal{P}$ to $\mathcal{Q}$. Different observers in different inertial frames will attribute different coordinates to each birth and different components to the births' vectorial separation; but all observers can agree that they are talking about the same events $\mathcal{P}$ and $\mathcal{Q}$ in spacetime and the same separation vector $\Delta\vec{x}$. In this sense, $\mathcal{P}$, $\mathcal{Q}$, and $\Delta\vec{x}$ are *frame-independent, geometric objects* (points and arrows) that reside in spacetime.

1.2.3 [R] Special Relativistic Concepts: Principle of Relativity; the Interval and its Invariance

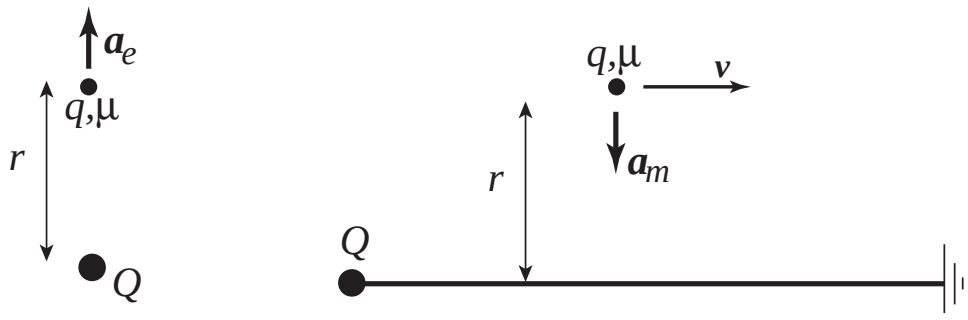
The *principle of relativity* states that *Every (special relativistic) law of physics must be expressible as a geometric, frame-independent relationship between geometric, frame-independent objects*, i.e. objects such as points in spacetime and vectors and tensors, which represent physical quantities such as events and particle momenta and the electromagnetic field.

Since the laws are all geometric (i.e., unrelated to any reference frame or coordinate system), there is no way that they can distinguish one inertial reference frame from any other. This leads to an alternative form of the principle of relativity (one commonly used in elementary textbooks and equivalent to the above): *All the (special relativistic) laws of physics are the same in every inertial reference frame, everywhere in spacetime.* A more operational version of this principle is the following: Give identical instructions for a specific physics experiment to two different observers in two different inertial reference frames at the same or different locations in Minkowski (i.e., gravity-free) spacetime. The experiment must be self-contained, i.e., it must not involve observations of the external universe's properties (the "environment"), though it might utilize carefully calibrated tools derived from the external universe. For example, an *unacceptable* experiment would be a measurement of the anisotropy of the Universe's cosmic microwave radiation and a computation therefrom of the observer's velocity relative to the radiation's mean rest frame; such an experiment studies the Universal environment, not the fundamental laws of physics. An *acceptable* experiment would be a measurement of the speed of light using the rods and clocks of the observer's own frame, or a measurement of cross sections for elementary particle reactions using cosmic-ray particles whose incoming energies and compositions are measured as initial conditions for the experiment. The principle of relativity says that in these or any other similarly self-contained experiments, the two observers in their two different inertial frames must obtain identically the same experimental results—to within the accuracy of their experimental techniques. Since the experimental results are governed by the (nongravitational) laws of physics, this is equivalent to the statement that all physical laws are the same in the two inertial frames.

Perhaps the most central of special relativistic laws is the one stating that *the speed of light c in vacuum is frame-independent*, i.e., is a constant, independent of the inertial reference frame in which it is measured. In other words, there is no *aether* that supports light's vibrations and in the process influences its speed — a remarkable fact that came as a great experimental surprise to physicists at the end of the nineteenth century.

The constancy of the speed of light is built into Maxwell's equations. In order for the Maxwell equations to be frame independent, the speed of light, which appears in them, must also be frame independent. In this sense, the constancy of the speed of light follows from the Principle of Relativity; it is not an independent postulate. This is illustrated in Box 1.2.

Box 1.2
Measuring the Speed of Light Without Light



In some inertial reference frame we perform two experiments using two particles, one with a large charge Q; the other, a test particle, with a much smaller charge q and mass μ. In the first experiment we place the two particles at rest, separated by a distance $|\Delta x| \equiv r$ and measure the electrical repulsive acceleration a_e of q (left diagram). In Gaussian cgs units (where the speed of light shows up explicitly instead of via $\epsilon_o\mu_o = 1/c^2$), the acceleration is $a_e = qQ/r^2\mu$. In the second experiment, we connect Q to ground by a long wire, and we place q at the distance $|\Delta x| = r$ from the wire and set it moving at speed v parallel to the wire. The charge Q flows down the wire with an e-folding time τ so the current is $I = dQ/d\tau = (Q/\tau)e^{-t/\tau}$. At early times $0 < t \ll \tau$, this current $I = Q/\tau$ produces a solenoidal magnetic field at q with field strength $B = (2/cr)(Q/\tau)$, and this field exerts a magnetic force on q, giving it an acceleration $a_m = q(v/c)B/\mu = 2vqQ/c^2\tau r/\mu$. The ratio of the electric acceleration in the first experiment to the magnetic acceleration in the second experiment is $a_e/a_m = c^2\tau/2rv$. Therefore, we can measure the speed of light c in our chosen inertial frame by performing this pair of experiments, carefully measuring the separation r, speed v, current Q/τ, and accelerations, and then simply computing $c = \sqrt{(2rv/\tau)(a_e/a_m)}$. The principle of relativity insists that the result of this pair of experiments should be independent of the inertial frame in which they are performed. Therefore, the speed of light c which appears in Maxwell's equations must be frame-independent. In this sense, the constancy of the speed of light follows from the Principle of Relativity as applied to Maxwell's equations.

The constancy of the speed of light was verified with very high precision in an era when the units of length (centimeters) and the units of time (seconds) were defined independently. By 1983, the constancy had become so universally accepted that it was used to redefine the meter (which is hard to measure precisely) in terms of the second (which is much easier to measure with modern technology[4]): The meter is now related to the second in such a way

[4]The second is defined as the duration of 9,192,631,770 periods of the radiation produced by a certain hyperfine transition in the ground state of a ^{133}Cs atom that is at rest in empty space. Today (2008) all fundamental physical units except mass units (e.g. the kilogram) are defined similarly in terms of fundamental constants of nature.

that the speed of light is precisely $c = 299,792,458$ m s^{-1} ; i.e., one meter is the distance traveled by light in $(1/299,792,458)$ seconds.

Because of this constancy of the light speed, it is permissible when studying special relativity to set c to unity. Doing so is equivalent to the relationship

$$c = 2.99792458 \times 10^{10} \text{cm}\,\text{s}^{-1} = 1 \tag{1.3a}$$

between seconds and centimeters; i.e., equivalent to

$$1 \text{ second} = 2.99792458 \times 10^{10} \text{ cm} . \tag{1.3b}$$

We shall refer to units in which $c = 1$ as *geometrized units*, and we shall adopt them throughout this book, when dealing with relativistic physics, since they make equations look much simpler. Occasionally it will be useful to restore the factors of c to an equation, thereby converting it to ordinary (SI or Gaussian-cgs) units. This restoration is achieved easily using dimensional considerations. For example, the equivalence of mass m and energy $\mathcal{E}$ is written in geometrized units as $\mathcal{E} = m$. In cgs units $\mathcal{E}$ has dimensions ergs = gram cm^2 sec^{-2}, while m has dimensions of grams, so to make $\mathcal{E} = m$ dimensionally correct we must multiply the right side by a power of c that has dimensions cm^2/sec^2, i.e. by c^2; thereby we obtain $\mathcal{E} = mc^2$.

We turn, next, to another fundamental concept, the *interval* $(\Delta s)^2$ between the two events $\mathcal{P}$ and $\mathcal{Q}$ whose separation vector is $\Delta\vec{x}$. In a specific but arbitrary inertial reference frame and in geometrized units, $(\Delta s)^2$ is given by

$$\boxed{(\Delta s)^2 \equiv -(\Delta t)^2 + (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 = -(\Delta t)^2 + \sum_{i,j} \delta_{ij} \Delta x^i \Delta x^j \ ;} \tag{1.4a}$$

cf. Eq. (1.2). Here δ_{ij} is the Kronecker delta, (unity if $i = j$; zero otherwise) and the spatial indices i and j are summed over 1, 2, 3. If $(\Delta s)^2 > 0$, the events $\mathcal{P}$ and $\mathcal{Q}$ are said to have a *spacelike* separation; if $(\Delta s)^2 = 0$, their separation is *null* or *lightlike*; and if $(\Delta s)^2 < 0$, their separation is *timelike*. For timelike separations, $(\Delta s)^2 < 0$ implies that Δs is imaginary; to avoid dealing with imaginary numbers, we describe timelike intervals by

$$\boxed{(\Delta\tau)^2 \equiv -(\Delta s)^2 \ ,} \tag{1.4b}$$

whose square root $\Delta\tau$ is real.

The coordinate separation between $\mathcal{P}$ and $\mathcal{Q}$ depends on one's reference frame; i.e., if $\Delta x^{\alpha'}$ and Δx^{α} are the coordinate separations in two different frames, then $\Delta x^{\alpha'} \neq \Delta x^{\alpha}$. Despite this frame dependence, the principle of relativity forces the interval $(\Delta s)^2$ to be the same in all frames:

$$\begin{aligned} (\Delta s)^2 &= -(\Delta t)^2 + (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 \\ &= -(\Delta t')^2 + (\Delta x')^2 + (\Delta y')^2 + (\Delta z')^2 \end{aligned} \tag{1.5}$$

We shall sketch a proof for the case of two events $\mathcal{P}$ and $\mathcal{Q}$ whose separation is timelike:

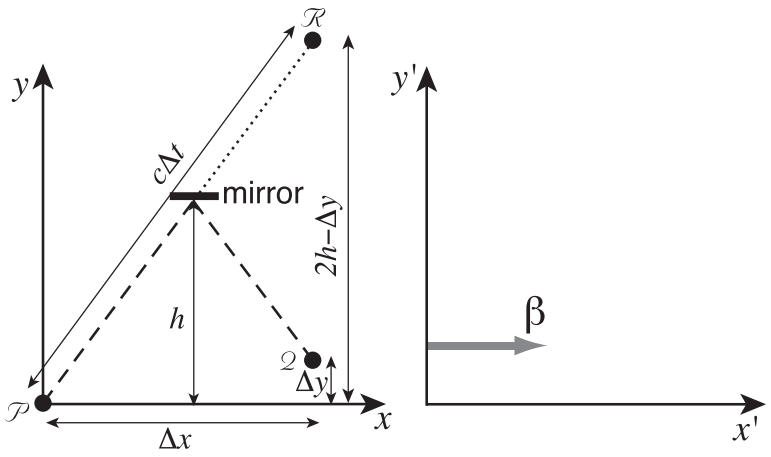


Fig. 1.5: Geometry for proving the invariance of the interval.

Choose the spatial coordinate systems of the primed and unprimed frames in such a way that (i) their relative motion (with speed β that will not enter into our analysis) is along the x direction and the x' direction, (ii) event $\mathcal{P}$ lies on the x and x' axes, and (iii) event $\mathcal{Q}$ lies in the x-y plane and in the x'-y' plane, as shown in Fig. 1.5. Then evaluate the interval between $\mathcal{P}$ and $\mathcal{Q}$ in the unprimed frame by the following construction: Place a mirror parallel to the x-z plane at precisely the height h that permits a photon, emitted from $\mathcal{P}$, to travel along the dashed line of Fig. 1.5 to the mirror, then reflect off the mirror and continue along the dashed path, arriving at event $\mathcal{Q}$. If the mirror were placed lower, the photon would arrive at the spatial location of $\mathcal{Q}$ sooner than the time of $\mathcal{Q}$; if placed higher, it would arrive later. Then the distance the photon travels (the length of the two-segment dashed line) is equal to $c\Delta t = \Delta t$, where Δt is the time between events $\mathcal{P}$ and $\mathcal{Q}$ as measured in the unprimed frame. If the mirror had not been present, the photon would have arrived at event $\mathcal{R}$ after time Δt, so $c\Delta t$ is the distance between $\mathcal{P}$ and $\mathcal{R}$. From the diagram it is easy to see that the height of $\mathcal{R}$ above the x axis is $2h - \Delta y$, and the Pythagorean theorem then implies that

$$(\Delta s)^2 = -(\Delta t)^2 + (\Delta x)^2 + (\Delta y)^2 = -(2h - \Delta y)^2 + (\Delta y)^2 \ . \tag{1.6a}$$

The same construction in the primed frame must give the same formula, but with primes

$$(\Delta s')^2 = -(\Delta t')^2 + (\Delta x')^2 + (\Delta y')^2 = -(2h' - \Delta y')^2 + (\Delta y')^2 \ . \tag{1.6b}$$

The proof that $(\Delta s')^2 = (\Delta s)^2$ then reduces to showing that the principle of relativity requires that distances perpendicular to the direction of relative motion of two frames be the same as measured in the two frames, $h' = h$, $\Delta y' = \Delta y$. We leave it to the reader to develop a careful argument for this [Ex. 1.2].

Because of its frame invariance, the interval $(\Delta s)^2$ can be regarded as a geometric property of the vector $\Delta\vec{x}$ that reaches from $\mathcal{P}$ to $\mathcal{Q}$; we shall call it the *squared length* $(\Delta\vec{x})^2$ of $\Delta\vec{x}$:

$$(\Delta\vec{x})^2 \equiv (\Delta s)^2 \ . \tag{1.7}$$

Note that this squared length, despite its name, can be negative (for timelike $\Delta\vec{x}$) or zero (for null $\Delta\vec{x}$) as well as positive (for spacelike $\Delta\vec{x}$).

This invariant interval between two events is as fundamental to Minkowski spacetime as the Euclidean distance between two points is to flat 3-space. Just as the Euclidean distance gives rise to the geometry of 3-space, as embodied, e.g., in Euclid's axioms, so the interval gives rise to the geometry of spacetime, which we shall be exploring. If this spacetime geometry were as intuitively obvious to humans as is Euclidean geometry, we would not need the crutch of inertial reference frames to arrive at it. Nature (presumably) has no need for such a crutch. To Nature (it seems evident), the geometry of Minkowski spacetime, as embodied in the invariant interval, is among the most fundamental aspects of physical law.

Before we leave this central idea, we should emphasize that vacuum electromagnetic radiation is not the only type of wave in nature. In this course, we shall encounter dispersive media, like optical fibers or plasmas, where electromagnetic signals travel slower than c, and we shall analyze sound waves and seismic waves where the governing laws do not involve electromagnetism at all. How do these fit into our special relativistic framework? The answer is simple. Each of these waves requires an underlying medium that is at rest in one particular frame (not necessarily inertial) and the velocity of the wave, specifically the group velocity, is most simply calculated in this frame *from the waves' and medium's fundamental laws*. We can then use the kinematic rules of Lorentz transformations to compute the velocity in another frame. However, if we had chosen to compute the wave speed in the second frame directly, *using the same fundamental laws*, we would have gotten the same answer, albeit perhaps with greater effort. All waves are in full compliance with the principle of relativity. What is special about vacuum electromagnetic waves and, by extension, photons, is that no medium (or "ether" as it used to be called) is needed for them to propagate. Their speed is therefore the same in all frames.

This raises an interesting question. What about other waves that do not require an underlying medium? What about electron de Broglie waves? Here the fundamental wave equation, Schrödinger's or Dirac's, is mathematically different from Maxwell's and contains an important parameter, the electron rest mass. This allows the fundamental laws of relativistic quantum mechanics to be written in a form that is the same in all inertial reference frames and that allows an electron, considered as either a wave or a particle, to travel at a different speed when measured in a different frame.

What about non-electromagnetic waves whose quanta have vanishing rest mass? For a long while, we thought that neutrinos provided a good example, but we now know from experiment that their rest masses are non-zero. However, there are other particles that have not yet been detected, including photinos (the hypothesized, supersymmetric partners to photons) and gravitons (and their associated gravitational waves which we shall discuss in Chapter 26), that are believed to exist without a rest mass (or an ether!), just like photons. Must these travel at the same speed as photons? The answer to this question, according to the principle of relativity, is "yes". The reason is simple. Suppose there were two such waves (or particles) whose governing laws led to different speeds, c and $c' < c$, each the same in all reference frames. If we then move with speed c' in the direction of propagation of the second wave, we would bring it to rest, in conflict with our hypothesis that its speed is frame-independent. Therefore all signals, whose governing laws require them to travel with a speed that has no governing parameters (no rest mass and no underlying medium with physical properties) must travel with a unique speed which we call "c". The speed of light

is more fundamental to relativity than light itself!

****

EXERCISES

Exercise 1.1 *Practice: [R] Geometrized Units*
Convert the following equations from the geometrized units in which they are written to cgs/Gaussian units:

(a) The "Planck time" t_P expressed in terms of Newton's gravitation constant G and Planck's constant $\hbar$, $t_P = \sqrt{G\hbar}$. What is the numerical value of t_P in seconds? in meters?

(b) The Lorentz force law $m d\mathbf{v}/dt = e(\mathbf{E} + \mathbf{v} \times \mathbf{B})$.

(c) The expression $\mathbf{p} = \hbar\omega\mathbf{n}$ for the momentum $\mathbf{p}$ of a photon in terms of its angular frequency ω and direction $\mathbf{n}$ of propagation.

How tall are you, in seconds? How old are you, in centimeters?

Exercise 1.2 *Derivation and Example: [R] Invariance of the Interval*
Complete the derivation of the invariance of the interval given in the text [Eqs. (1.6)], using the principle of relativity in the form that the laws of physics must be the same in the primed and unprimed frames. Hints, if you need them:

(a) Having carried out the construction shown in Fig. 1.5 in the unprimed frame, use the same mirror and photons for the analogous construction in the primed frame. Argue that, independently of the frame in which the mirror is at rest (unprimed or primed), the fact that the reflected photon has (angle of reflection) = (angle of incidence) in the primed frame implies that this is also true for this same photon in the unprimed frame. Thereby conclude that the construction leads to Eq. (1.6b) as well as to (1.6a).

(b) Then argue that the perpendicular distance of an event from the common x and x' axis must be the same in the two reference frames, so $h' = h$ and $\Delta y' = \Delta y$; whence Eqs. (1.6b) and (1.6a) imply the invariance of the interval. [For a leisurely version of this argument, see Secs. 3.6 and 3.7 of Taylor and Wheeler (1992).]

****

1.3 [N & R] Tensor Algebra Without a Coordinate System

We now pause in our development of the geometric view of physical laws, to introduce, in a coordinate-free way, some fundamental concepts of differential geometry: tensors, the inner product, the metric tensor, the tensor product, and contraction of tensors. In this section we shall allow the space in which the concepts live to be either 4-dimensional Minkowski spacetime, or 3-dimensional Euclidean space; we shall denote its dimensionality by N; and we shall use spacetime's arrowed notation $\vec{A}$ for vectors even though the space might be Euclidean 3-space.

We have already defined a vector $\vec{A}$ as a straight arrow from one point, say $\mathcal{P}$, in our space to another, say $\mathcal{Q}$. Because our space is flat, there is a unique and obvious way to transport such an arrow from one location to another, keeping its length and direction unchanged.[5] Accordingly, we shall regard vectors as unchanged by such transport. This enables us to ignore the issue of where in space a vector actually resides; it is completely determined by its direction and its length.

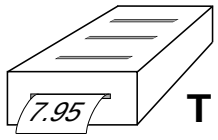


Fig. 1.6: A rank-3 tensor T.

A *rank-n tensor* T is, by definition, a real-valued, linear function of n vectors. Pictorially we shall regard T as a box (Fig. 1.6) with n slots in its top, into which are inserted n vectors, and one slot in its end, out of which rolls computer paper with a single real number printed on it: the value that the tensor T has when evaluated as a function of the n inserted vectors. Notationally we shall denote the tensor by a bold-face sans-serif character T

$$\mathsf{T}(\underbrace{\_,\_,\_,\_}_{\text{n slots in which to put the vectors}}) \ . \tag{1.8a}$$

If T is a rank-3 tensor (has 3 slots) as in Fig. 1.6, then its value on the vectors $\vec{A}, \vec{B}, \vec{C}$ will be denoted $\mathsf{T}(\vec{A}, \vec{B}, \vec{C})$. Linearity of this function can be expressed as

$$\mathsf{T}(e\vec{E} + f\vec{F}, \vec{B}, \vec{C}) = e\mathsf{T}(\vec{E}, \vec{B}, \vec{C}) + f\mathsf{T}(\vec{F}, \vec{B}, \vec{C}) \ , \tag{1.8b}$$

where e and f are real numbers, and similarly for the second and third slots.

We have already defined the *squared length* $(\vec{A})^2 \equiv \vec{A}^2$ of a vector $\vec{A}$ as the squared distance (in 3-space) or interval (in spacetime) between the points at its tail and its tip. The *inner product* $\vec{A} \cdot \vec{B}$ of two vectors is defined in terms of the squared length by

$$\vec{A} \cdot \vec{B} \equiv \frac{1}{4}\left[(\vec{A} + \vec{B})^2 - (\vec{A} - \vec{B})^2\right] \ . \tag{1.9a}$$

[5] This is not so in curved spaces, as we shall see in Sec. 24.7.

In Euclidean space this is the standard inner product, familiar from elementary geometry.

Because the inner product $\vec{A}\cdot\vec{B}$ is a linear function of each of its vectors, we can regard it as a tensor of rank 2. When so regarded, the inner product is denoted $\mathbf{g}(\_\,,\_\,)$ and is called the *metric tensor*. In other words, the metric tensor $\mathbf{g}$ is that linear function of two vectors whose value is given by

$$\mathbf{g}(\vec{A},\vec{B}) \equiv \vec{A}\cdot\vec{B}\ . \tag{1.9b}$$

Notice that, because $\vec{A}\cdot\vec{B}=\vec{B}\cdot\vec{A}$, the metric tensor is *symmetric* in its two slots; i.e., one gets the same real number independently of the order in which one inserts the two vectors into the slots:

$$\mathbf{g}(\vec{A},\vec{B}) = \mathbf{g}(\vec{B},\vec{A}) \tag{1.9c}$$

With the aid of the inner product, we can regard any vector $\vec{A}$ as a tensor of rank one: The real number that is produced when an arbitrary vector $\vec{C}$ is inserted into $\vec{A}$'s slot is

$$\vec{A}(\vec{C}) \equiv \vec{A}\cdot\vec{C}\ . \tag{1.9d}$$

Second-rank tensors appear frequently in the laws of physics—often in roles where one sticks a single vector into the second slot and leaves the first slot empty thereby producing a single-slotted entity, a vector. A familiar example is a rigid body's (Newtonian) moment-of-inertia tensor $\mathsf{I}(\_\,,\_\,)$. Insert the body's angular velocity vector $\mathbf{\Omega}$ into the second slot, and you get the body's angular momentum vector $\mathbf{J}(\_\,)=\mathsf{I}(\_\,,\mathbf{\Omega})$. Other examples are the stress tensor of a solid, a fluid, a plasma or a field (Sec. 1.12 below) and the electromagnetic field tensor (Secs. 1.4.3 and 1.10 below).

From three (or any number of) vectors $\vec{A}$, $\vec{B}$, $\vec{C}$ we can construct a tensor, their *tensor product* (also called *outer product* in contradistinction to the inner product $\vec{A}\cdot\vec{B}$), defined as follows:

$$\vec{A}\otimes\vec{B}\otimes\vec{C}(\vec{E},\vec{F},\vec{G}) \equiv \vec{A}(\vec{E})\vec{B}(\vec{F})\vec{C}(\vec{G}) = (\vec{A}\cdot\vec{E})(\vec{B}\cdot\vec{F})(\vec{C}\cdot\vec{G})\ . \tag{1.10a}$$

Here the first expression is the notation for the value of the new tensor, $\vec{A}\otimes\vec{B}\otimes\vec{C}$ evaluated on the three vectors $\vec{E}$, $\vec{F}$, $\vec{G}$; the middle expression is the ordinary product of three real numbers, the value of $\vec{A}$ on $\vec{E}$, the value of $\vec{B}$ on $\vec{F}$, and the value of $\vec{C}$ on $\vec{G}$; and the third expression is that same product with the three numbers rewritten as scalar products. Similar definitions can be given (and should be obvious) for the tensor product of any two or more tensors of any rank; for example, if T has rank 2 and S has rank 3, then

$$\mathsf{T}\otimes\mathsf{S}(\vec{E},\vec{F},\vec{G},\vec{H},\vec{J}) \equiv \mathsf{T}(\vec{E},\vec{F})\mathsf{S}(\vec{G},\vec{H},\vec{J})\ . \tag{1.10b}$$

One last geometric (i.e. frame-independent) concept we shall need is *contraction*. We shall illustrate this concept first by a simple example, then give the general definition. From two vectors $\vec{A}$ and $\vec{B}$ we can construct the tensor product $\vec{A}\otimes\vec{B}$ (a second-rank tensor), and we can also construct the scalar product $\vec{A}\cdot\vec{B}$ (a real number, i.e. a *scalar*, i.e. a *rank-0 tensor*). The process of contraction is the construction of $\vec{A}\cdot\vec{B}$ from $\vec{A}\otimes\vec{B}$

$$\text{contraction}(\vec{A}\otimes\vec{B}) \equiv \vec{A}\cdot\vec{B}\ . \tag{1.11a}$$

One can show fairly easily using component techniques (Sec. 1.5 below) that any second-rank tensor $\mathbf{T}$ can be expressed as a sum of tensor products of vectors, $\mathbf{T} = \vec{A}\otimes\vec{B} + \vec{C}\otimes\vec{D} + \ldots$; and correspondingly, it is natural to define the contraction of $\mathbf{T}$ to be contraction($\mathbf{T}$) $= \vec{A}\cdot\vec{B} + \vec{C}\cdot\vec{D} + \ldots$. Note that this contraction process lowers the rank of the tensor by two, from 2 to 0. Similarly, for a tensor of rank n one can construct a tensor of rank $n-2$ by contraction, but in this case one must specify which slots are to be contracted. For example, if $\mathbf{T}$ is a third rank tensor, expressible as $\mathbf{T} = \vec{A}\otimes\vec{B}\otimes\vec{C} + \vec{E}\otimes\vec{F}\otimes\vec{G} + \ldots$, then the contraction of $\mathbf{T}$ on its first and third slots is the rank-1 tensor (vector)

$$1\&3\text{contraction}(\vec{A}\otimes\vec{B}\otimes\vec{C} + \vec{E}\otimes\vec{F}\otimes\vec{G} + \ldots) \equiv (\vec{A}\cdot\vec{C})\vec{B} + (\vec{E}\cdot\vec{G})\vec{F} + \ldots \ . \qquad (1.11b)$$

All the concepts developed in this section (vectors, tensors, metric tensor, inner product, tensor product, and contraction of a tensor) can be carried over, with no change whatsoever, into *any* vector space[6] that is endowed with a concept of squared length.

1.4 Particle Kinetics and Lorentz Force Without a Reference Frame

In this section we shall illustrate our geometric viewpoint by formulating the laws of motion for particles, first in Newtonian physics and then in special relativity.

1.4.1 [N] Newtonian Particle Kinetics

In Newtonian physics, a classical particle moves through Euclidean 3-space as universal time t passes. At time t it is located at some point $\mathbf{x}(t)$ (its *position*). The function $\mathbf{x}(t)$ represents a curve in 3-space, the particle's *trajectory*. The particle's *velocity* $\mathbf{v}(t)$ is the time derivative of its position, its *momentum* $\mathbf{p}(t)$ is the product of its mass m and velocity, and its *acceleration* $\mathbf{a}(t)$ is the time derivative of its velocity

$$\mathbf{v}(t) = d\mathbf{x}/dt \ , \quad \mathbf{p}(t) = m\mathbf{v}(t), \quad \mathbf{a}(t) = d\mathbf{v}/dt = d^2\mathbf{x}/dt^2 \ . \qquad (1.12)$$

Since points in 3-space are geometric objects (defined independently of any coordinate system), so also are the trajectory $\mathbf{x}(t)$, the velocity, the momentum, and the acceleration. (Physically, of course, the velocity has an ambiguity; it depends on one's standard of rest.)

Newton's second law of motion states that the particle's momentum can change only if a force $\mathbf{F}$ acts on it, and that its change is given by

$$d\mathbf{p}/dt = m\mathbf{a} = \mathbf{F} \ . \qquad (1.13)$$

If the force is produced by an electric field $\mathbf{E}$ and magnetic field $\mathbf{B}$, then this law of motion takes the familiar Lorentz-force form

$$d\mathbf{p}/dt = q(\mathbf{E} + \mathbf{v}\times\mathbf{B}) \qquad (1.14)$$

[6] or, more precisely, any vector space over the real numbers. If the vector space's scalars are complex numbers, as in quantum mechanics, then slight changes are needed.

(here we have used the vector cross product, which will not be introduced formally until Sec. 1.9 below). Obviously, these laws of motion are geometric relationships between geometric objects.

1.4.2 [R] Relativistic Particle Kinetics: World Lines, 4-Velocity, 4-Momentum and its Conservation, 4-Force

In special relativity, a particle moves through 4-dimensional spacetime along a curve (its *world line*) which we shall denote, in frame-independent notation, by $\vec{x}(\tau)$. Here τ is time as measured by an ideal clock that the particle carries (the particle's *proper time*), and $\vec{x}$ is the location of the particle in spacetime when its clock reads τ (or, equivalently, the vector from the arbitrary origin to that location).

The particle typically will experience an acceleration as it moves—e.g., an acceleration produced by an external electromagnetic field. This raises the question of how the acceleration affects the ticking rate of the particle's clock. We define the accelerated clock to be *ideal* if its ticking rate is totally unaffected by its acceleration, i.e., if it ticks at the same rate as a freely moving (inertial) ideal clock that is momentarily at rest with respect to it. *The builders of inertial guidance systems for airplanes and missiles always try to make their clocks as acceleration-independent, i.e., as ideal, as possible.*

We shall refer to the inertial frame in which a particle is momentarily at rest as its *momentarily comoving inertial frame* or *momentary rest frame*. Now, the particle's clock (which measures τ) is ideal and so are the inertial frame's clocks (which measure coordinate time t). Therefore, a tiny interval $\Delta\tau$ of the particle's proper time is equal to the lapse of coordinate time in the particle's momentary rest frame, $\Delta\tau = \Delta t$. Moreover, since the two events $\vec{x}(\tau)$ and $\vec{x}(\tau + \Delta\tau)$ on the clock's world line occur at the same spatial location in its momentary rest frame, $\Delta x^i = 0$ (where $i = 1,\ 2,\ 3$), the invariant interval between those events is $(\Delta s)^2 = -(\Delta t)^2 + \sum_{i,j} \Delta x^i \Delta x^j \delta_{ij} = -(\Delta t)^2 = -(\Delta\tau)^2$. This shows that *the particle's proper time τ is equal to the square root of the invariant interval, $\tau = \sqrt{-s^2}$, along its world line.*

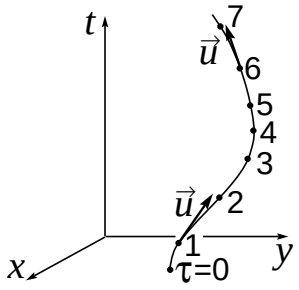


Fig. 1.7: Spacetime diagram showing the world line $\vec{x}(\tau)$ and 4-velocity $\vec{u}$ of an accelerated particle. Note that the 4-velocity is tangent to the world line.

Figure 1.7 shows the world line of the accelerated particle in a spacetime diagram where the axes are coordinates of an *arbitrary* Lorentz frame. This diagram is intended to emphasize

the world line as a frame-independent, geometric object. Also shown in the figure is the particle's 4-*velocity* $\vec{u}$, which (by analogy with the velocity in 3-space) is the time derivative of its position:

$$\boxed{\vec{u} \equiv d\vec{x}/d\tau \ .} \tag{1.15}$$

This derivative is defined by the usual limiting process

$$\frac{d\vec{x}}{d\tau} \equiv \lim_{\Delta\tau\to 0} \frac{\vec{x}(\tau+\Delta\tau) - \vec{x}(\tau)}{\Delta\tau} \ . \tag{1.16}$$

The squared length of the particle's 4-velocity is easily seen to be -1:

$$\vec{u}^2 \equiv \mathbf{g}(\vec{u},\vec{u}) = \frac{d\vec{x}}{d\tau}\cdot\frac{d\vec{x}}{d\tau} = \frac{d\vec{x}\cdot d\vec{x}}{(d\tau)^2} = -1 \ . \tag{1.17}$$

The last equality follows from the fact that $d\vec{x}\cdot d\vec{x}$ is the squared length of $d\vec{x}$ which equals the invariant interval $(\Delta s)^2$ along it, and $(d\tau)^2$ is minus that invariant interval.

The particle's 4-momentum is the product of its 4-velocity and rest mass

$$\boxed{\vec{p} \equiv m\vec{u} = m d\vec{x}/d\tau \equiv d\vec{x}/d\zeta \ .} \tag{1.18}$$

Here the parameter ζ is a renormalized version of proper time,

$$\zeta \equiv \tau/m \ . \tag{1.19}$$

This ζ, and any other renormalized version of proper time with position-independent renormalization factor, are called *affine parameters* for the particle's world line. Expression (1.18), together with the unit length of the 4-velocity $\vec{u}^2 = -1$, implies that the squared length of the 4-momentum is

$$\vec{p}^{\,2} = -m^2 \ . \tag{1.20}$$

In quantum theory a particle is described by a relativistic wave function which, in the geometric optics limit (Chapter 6), has a wave vector $\vec{k}$ that is related to the classical particle's 4-momentum by

$$\vec{k} = \vec{p}/\hbar \ . \tag{1.21}$$

The above formalism is valid only for particles with nonzero rest mass, $m \neq 0$. The corresponding formalism for a *particle with zero rest mass* (e.g. a photon or a graviton[7]) can be obtained from the above by taking the limit as $m \to 0$ and $d\tau \to 0$ with the quotient $d\zeta = d\tau/m$ held finite. More specifically, the 4-momentum of a zero-rest-mass particle is well defined (and participates in the conservation law to be discussed below), and it is expressible in terms of the particle's affine parameter ζ by Eq. (1.18)

$$\vec{p} = \frac{d\vec{x}}{d\zeta} \ . \tag{1.22}$$

[7]We do not know for sure that photons and gravitons are massless, but the laws of physics as currently undertood require them to be massless and there are tight experimental limits on their rest masses.

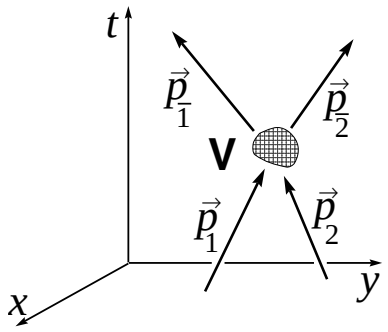


Fig. 1.8: Spacetime diagram depicting the law of 4-momentum conservation for a situation where two particles, numbered 1 and 2, enter an interaction region $\mathcal{V}$ in spacetime, there interact strongly, and produce two new particles, numbered $\bar{1}$ and $\bar{2}$. The sum of the final 4-momenta, $\vec{p}_{\bar{1}} + \vec{p}_{\bar{2}}$, must be equal to the sum of the initial 4-momenta, $\vec{p}_1 + \vec{p}_2$.

However, the particle's 4-velocity $\vec{u} = \vec{p}/m$ is infinite and thus undefined; and proper time $\tau = m\zeta$ ticks vanishingly slowly along its world line and thus is undefined. Because proper time is the square root of the invariant interval along the world line, the interval between two neighboring points on the world line vanishes identically; and correspondingly *the world line of a zero-rest-mass particle is null*. (By contrast, since $d\tau^2 > 0$ and $ds^2 < 0$ along the world line of a particle with finite rest mass, *the world line of a finite-rest-mass particle is timelike*.)

The 4-momenta of particles are important because of the *law of conservation of 4-momentum* (which, as we shall see in Sec. 1.6, is equivalent to the conservation laws for energy and ordinary momentum): If a number of "initial" particles, named $A = 1, 2, 3, \ldots$ enter a restricted region of spacetime $\mathcal{V}$ and there interact strongly to produce a new set of "final" particles, named $\bar{A} = \bar{1}, \bar{2}, \bar{3}, \ldots$ (Fig. 1.8), then the total 4-momentum of the final particles must be be the same as the total 4-momentum of the initial ones:

$$\boxed{\sum_{\bar{A}} \vec{p}_{\bar{A}} = \sum_{A} \vec{p}_A \ .} \tag{1.23}$$

Note that this law of 4-momentum conservation is expressed in frame-independent, geometric language—in accord with Einstein's insistence that all the laws of physics should be so expressible. As we shall see in Part VI, momentum conservation is a consequence of the translation symmetry of flat, 4-dimensional spacetime. In general relativity's curved spacetime, where that translation symmetry is lost, we lose momentum conservation except under special circumstances; see Sec. 24.9.4.

If a particle moves freely (no external forces and no collisions with other particles), then its 4-momentum $\vec{p}$ will be conserved along its world line, $d\vec{p}/d\zeta = 0$. Since $\vec{p}$ is tangent to the world line, this means that the direction of the world line never changes; i.e., the free particle moves along a straight line through spacetime. To change the particle's 4-momentum, one must act on it with a 4-force $\vec{F}$,

$$d\vec{p}/d\tau = \vec{F} \ . \tag{1.24}$$

If the particle is a fundamental one (e.g., photon, electron, proton), then the 4-force must leave its rest mass unchanged,

$$0 = dm^2/d\tau = -d\vec{p}^2/d\tau = -2\vec{p}\cdot d\vec{p}/d\tau = -2\vec{p}\cdot\vec{F}\ ; \tag{1.25}$$

i.e., the 4-force must be orthogonal to the 4-momentum.

1.4.3 [R] Geometric Derivation of the Lorentz Force Law

As an illustration of these physical concepts and mathematical tools, we shall use them to deduce the relativistic version of the Lorentz force law. From the outset, in accord with the principle of relativity, we insist that the law we seek be expressible in geometric, frame-independent language, i.e. in terms of vectors and tensors.

Consider a particle with charge q and rest mass $m \neq 0$, interacting with an electromagnetic field. It experiences an electromagnetic 4-force whose mathematical form we seek. The Newtonian version of the electromagnetic force $\mathbf{F} = q(\mathbf{E} + \mathbf{v}\times\mathbf{B})$ is proportional to q and contains one piece (electric) that is independent of velocity $\mathbf{v}$, and a second piece (magnetic) that is linear in $\mathbf{v}$. It is reasonable to expect that, in order to produce this Newtonian limit, the relativistic 4-force $\vec{F}$ will be proportional to q and will be linear in the 4-velocity $\vec{u}$. Linearity means there must exist some second-rank tensor $\mathsf{F}(\_\,,\_\,)$, the "electromagnetic field tensor", such that

$$d\vec{p}/d\tau = \vec{F}(\_\,) = q\mathsf{F}(\_\,,\vec{u})\ . \tag{1.26}$$

Because the 4-force $\vec{F}$ must be orthogonal to the particle's 4-momentum and thence also to its 4-velocity, $\vec{F}\cdot\vec{u} \equiv \vec{F}(\vec{u}) = 0$, expression (1.26) must vanish when $\vec{u}$ is inserted into its empty slot. In other words, for all timelike unit-length vectors $\vec{u}$,

$$\mathsf{F}(\vec{u},\vec{u}) = 0\ . \tag{1.27}$$

It is an instructive exercise (Ex. 1.3) to show that this is possible only if F is antisymmetric, so the electromagnetic 4-force is

$$\boxed{d\vec{p}/d\tau = q\mathsf{F}(\_\,,\vec{u})\ , \quad \text{where } \mathsf{F}(\vec{A},\vec{B}) = -\mathsf{F}(\vec{B},\vec{A}) \text{ for all } \vec{A} \text{ and } \vec{B}\ .} \tag{1.28}$$

This must be the relativistic form of the Lorentz force law. In Sec. 1.10 below, we shall deduce the relationship of the electromagnetic field tensor F to the more familiar electric and magnetic fields, and the relationship of this relativistic Lorentz force to its Newtonian form (1.14).

This discussion of particle kinematics and the electromagnetic force is elegant, but perhaps unfamiliar. In Secs. 1.6 and 1.10 we shall see that it is equivalent to the more elementary (but more complex) formalism based on components of vectors.

****

EXERCISES

Exercise 1.3 *Derivation and Example: [R] Antisymmetry of Electromagnetic Field Tensor*
Show that Eq. (1.27) can be true for all timelike, unit-length vectors $\vec{u}$ if and only if $\mathbf{F}$ is antisymmetric. [Hints: (i) Show that the most general second-rank tensor $\mathbf{F}$ can be written as the sum of a symmetric tensor $\mathbf{S}$ and an antisymmetric tensor $\mathbf{A}$, and that the antisymmetric piece contributes nothing to Eq. (1.27). (ii) Let $\vec{B}$ and $\vec{C}$ be any two vectors such that $\vec{B}+\vec{C}$ and $\vec{B}-\vec{C}$ are both timelike; show that $\mathbf{S}(\vec{B},\vec{C})=0$. (iii) Convince yourself (if necessary using the component tools developed in the next section) that this result, together with the 4-dimensionality of spacetime and the large arbitrariness inherent in the choice of $\vec{A}$ and $\vec{B}$, implies $\mathbf{S}$ vanishes (i.e., it gives zero when *any* two vectors are inserted into its slots).]

Exercise 1.4 *[R] Problem: Relativistic Gravitational Force Law*
In Newtonian theory the gravitational potential Φ exerts a force $\mathbf{F} = d\mathbf{p}/dt = -m\boldsymbol{\nabla}\Phi$ on a particle with mass m and momentum $\mathbf{p}$. Before Einstein formulated general relativity, some physicists constructed relativistic theories of gravity in which a Newtonian-like scalar gravitational field Φ exerted a 4-force $\vec{F} = d\vec{p}/d\tau$ on any particle with rest mass m, 4-velocity $\vec{u}$ and 4-momentum $\vec{p} = m\vec{u}$. What must that force law have been, in order to (i) obey the principle of relativity, (ii) reduce to Newton's law in the non-relativistic limit, and (iii) preserve the particle's rest mass as time passes?

****

1.5 Component Representation of Tensor Algebra

1.5.1 [N] Euclidean 3-space

In the Euclidean 3-space of Newtonian physics, there is a unique set of *orthonormal basis vectors* $\{\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z\} \equiv \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ associated with any *Cartesian coordinate system* $\{x, y, z\} \equiv \{x^1, x^2, x^3\} \equiv \{x_1, x_2, x_3\}$. [In Cartesian coordinates in Euclidean space, we will usually place indices down, but occasionally we will place them up. It doesn't matter. By definition, in Cartesian coordinates a quantity is the same whether its index is down or up.] The basis vector $\mathbf{e}_j$ points along the x_j coordinate direction, which is orthogonal to all the other coordinate directions, and it has unit length, so

$$\mathbf{e}_j \cdot \mathbf{e}_k = \delta_{jk} \ . \tag{1.29a}$$

Any vector $\mathbf{A}$ in 3-space can be expanded in terms of this basis,

$$\mathbf{A} = A_j \mathbf{e}_j \ . \tag{1.29b}$$

Here and throughout this book, we adopt the Einstein summation convention: repeated indices (in this case j) are to be summed (in this 3-space case over $j = 1, 2, 3$). By virtue

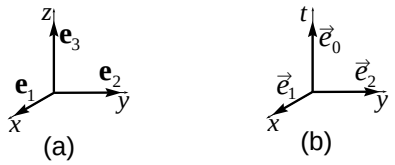


Fig. 1.9: (a) The orthonormal basis vectors $\mathbf{e}_j$ associated with a Euclidean coordinate system in 3-space; (b) the orthonormal basis vectors $\vec{e}_\alpha$ associated with an inertial (Lorentz) reference frame in Minkowski spacetime.

of the orthonormality of the basis, the components A_j of $\mathbf{A}$ can be computed as the scalar product

$$A_j = \mathbf{A} \cdot \mathbf{e}_j \ . \tag{1.29c}$$

(The proof of this is straightforward: $\mathbf{A} \cdot \mathbf{e}_j = (A_k \mathbf{e}_k) \cdot \mathbf{e}_j = A_k (\mathbf{e}_k \cdot \mathbf{e}_j) = A_k \delta_{kj} = A_j$.)

Any tensor, say the third-rank tensor $\mathsf{T}(\_\,,\_\,,\_\,)$, can be expanded in terms of tensor products of the basis vectors:

$$\mathsf{T} = T_{ijk} \mathbf{e}_i \otimes \mathbf{e}_j \otimes \mathbf{e}_k \ . \tag{1.29d}$$

The components T_{ijk} of T can be computed from T and the basis vectors by the generalization of Eq. (1.29c)

$$T_{ijk} = \mathsf{T}(\mathbf{e}_i, \mathbf{e}_j, \mathbf{e}_k) \ . \tag{1.29e}$$

(This equation can be derived using the orthonormality of the basis in the same way as Eq. (1.29c) was derived.) As an important example, the components of the metric are $g_{jk} = \mathbf{g}(\mathbf{e}_j, \mathbf{e}_k) = \mathbf{e}_j \cdot \mathbf{e}_k = \delta_{jk}$ [where the first equality is the method (1.29e) of computing tensor components, the second is the definition (1.9b) of the metric, and the third is the orthonormality relation (1.29a)]:

$$g_{jk} = \delta_{jk} \quad \text{in any orthonormal basis in 3-space.} \tag{1.29f}$$

In Part VI we shall often use bases that are not orthonormal; in such bases, the metric components will not be δ_{jk}.

The components of a tensor product, e.g. $\mathsf{T}(\_\,,\_\,,\_\,) \otimes \mathsf{S}(\_\,,\_\,)$, are easily deduced by inserting the basis vectors into the slots [Eq. (1.29e)]; they are $\mathsf{T}(\mathbf{e}_i, \mathbf{e}_j, \mathbf{e}_k) \otimes \mathsf{S}(\mathbf{e}_l, \mathbf{e}_m) = T_{ijk} S_{lm}$ [cf. Eq. (1.10a)]. In words, the components of a tensor product are equal to the ordinary arithmetic product of the components of the individual tensors.

In component notation, the inner product of two vectors and the value of a tensor when vectors are inserted into its slots are given by

$$\mathbf{A} \cdot \mathbf{B} = A_j B_j \ , \quad \mathsf{T}(\mathbf{A}, \mathbf{B}, \mathbf{C}) = T_{ijk} A_i B_j C_k \ , \tag{1.29g}$$

as one can easily show using previous equations. Finally, the contraction of a tensor [say, the fourth rank tensor $\mathsf{R}(\_\,,\_\,,\_\,,\_\,)$] on two of its slots [say, the first and third] has components that are easily computed from the tensor's own components:

$$\text{Components of [1\&3contraction of } \mathsf{R}] = R_{ijik} \tag{1.29h}$$

Note that R_{ijik} is summed on the i index, so it has only two free indices, j and k, and thus is the component of a second rank tensor, as it must be if it is to represent the contraction of a fourth-rank tensor.

1.5.2 [R] Minkowski Spacetime

In Minkowski spacetime, associated with any *inertial reference frame* (Fig. 1.3 and associated discussion in Sec. 1.2.2), there is a *Lorentz coordinate system* $\{t, x, y, z\} = \{x^0, x^1, x^2, x^3\}$ generated by the frame's rods and clocks. And associated with these coordinates is a set of *orthonormal basis vectors* $\{\vec{e}_t, \vec{e}_x, \vec{e}_y, \vec{e}_z\} = \{\vec{e}_0, \vec{e}_1, \vec{e}_2, \vec{e}_3\}$; cf. Fig. 1.9. (The reason for putting the indices up on the coordinates but down on the basis vectors will become clear below.) The basis vector $\vec{e}_\alpha$ points along the x^α coordinate direction, which is orthogonal to all the other coordinate directions, and it has squared length -1 for $\alpha = 0$ (vector pointing in a timelike direction) and $+1$ for $\alpha = 1, 2, 3$ (spacelike):

$$\vec{e}_\alpha \cdot \vec{e}_\beta = \eta_{\alpha\beta} \ . \tag{1.30}$$

Here $\eta_{\alpha\beta}$, the *orthonormality values* (a spacetime analog of the Kronecker delta) are defined by

$$\eta_{00} \equiv -1 \ , \quad \eta_{11} \equiv \eta_{22} \equiv \eta_{33} \equiv 1 \ , \quad \eta_{\alpha\beta} \equiv 0 \text{ if } \alpha \neq \beta \ . \tag{1.31}$$

The fact that $\vec{e}_\alpha \cdot \vec{e}_\beta \neq \delta_{\alpha\beta}$ prevents many of the Euclidean-space component-manipulation formulas (1.29c)–(1.29h) from holding true in Minkowski spacetime. There are two approaches to recovering these formulas. One approach, used in many old textbooks (including the first and second editions of Goldstein's *Classical Mechanics* and Jackson's *Classical Electrodynamics*), is to set $x^0 = it$, where $i = \sqrt{-1}$ and correspondingly make the time basis vector be imaginary, so that $\vec{e}_\alpha \cdot \vec{e}_\beta = \delta_{\alpha\beta}$. When this approach is adopted, the resulting formalism does not care whether indices are placed up or down; one can place them wherever one's stomach or liver dictate without asking one's brain. However, this $x^0 = it$ approach has severe disadvantages: (i) it hides the true physical geometry of Minkowski spacetime, (ii) it cannot be extended in any reasonable manner to non-orthonormal bases in flat spacetime, and (iii) it cannot be extended in any reasonable manner to the curvilinear coordinates that one must use in general relativity. For these reasons, most modern texts (including the third editions of Goldstein and Jackson) take an alternative approach, one always used in general relativity. This alternative, which we shall adopt, requires introducing two different types of components for vectors, and analogously for tensors: *contravariant components* denoted by superscripts, e.g. $T^{\alpha\beta\gamma}$, and *covariant components* denoted by subscripts, e.g. $T_{\alpha\beta\gamma}$. In Parts I–V of this book we introduce these components only for orthonormal bases; in Part VI we develop a more sophisticated version of them, valid for nonorthonormal bases.

A vector or tensor's *contravariant components* are defined as its expansion coefficients in the chosen basis [analog of Eq. (1.29d) in Euclidean 3-space]:

$$\vec{A} \equiv A^\alpha \vec{e}_\alpha \ , \quad \mathbf{T} \equiv T^{\alpha\beta\gamma} \vec{e}_\alpha \otimes \vec{e}_\beta \otimes \vec{e}_\gamma \ . \tag{1.32a}$$

Here and throughout this book, *Greek (spacetime) indices are to be summed whenever they are repeated with one up and the other down.* The *covariant components* are defined as the

numbers produced by evaluating the vector or tensor on its basis vectors [analog of Eq. (1.29e) in Euclidean 3-space]:

$$A_\alpha \equiv \vec{A}(\vec{e}_\alpha) = \vec{A}\cdot\vec{e}_\alpha \ , \quad T_{\alpha\beta\gamma} \equiv \mathbf{T}(\vec{e}_\alpha,\vec{e}_\beta,\vec{e}_\gamma) \ . \tag{1.32b}$$

These definitions have a number of important consequences. We shall derive them one after another and then at the end shall summarize them succinctly with equation numbers:

(i) The covariant components of the metric tensor are $g_{\alpha\beta} = \mathbf{g}(\vec{e}_\alpha,\vec{e}_\beta) = \vec{e}_\alpha\cdot\vec{e}_\beta = \eta_{\alpha\beta}$. Here the first equality is the definition (1.32b) of the covariant components and the second equality is the orthonormality relation (1.30) for the basis vectors.

(ii) The covariant components of any tensor can be computed from the contravariant components by $T_{\lambda\mu\nu} = \mathbf{T}(\vec{e}_\lambda,\vec{e}_\mu,\vec{e}_\nu) = T^{\alpha\beta\gamma}\vec{e}_\alpha\otimes\vec{e}_\beta\otimes\vec{e}_\gamma(\vec{e}_\lambda,\vec{e}_\mu,\vec{e}_\nu) = T^{\alpha\beta\gamma}(\vec{e}_\alpha\cdot\vec{e}_\lambda)(\vec{e}_\beta\cdot\vec{e}_\mu)(\vec{e}_\gamma\cdot\vec{e}_\nu) = T^{\alpha\beta\gamma}g_{\alpha\lambda}g_{\beta\mu}g_{\gamma\nu}$. The first equality is the definition (1.32b) of the covariant components, the second is the expansion (1.32a) of $\mathbf{T}$ on the chosen basis, the third is the definition (1.10a) of the tensor product, and the fourth is one version of our result (i) for the covariant components of the metric.

(iii) This result, $T_{\lambda\mu\nu} = T^{\alpha\beta\gamma}g_{\alpha\lambda}g_{\beta\mu}g_{\gamma\nu}$, together with the numerical values (i) of $g_{\alpha\beta}$, implies that when one lowers a spatial index there is no change in the numerical value of a component, and when one lowers a temporal index, the sign changes: $T_{ijk} = T^{ijk}$, $T_{0jk} = -T^{0jk}$, $T_{0j0} = +T^{0j0}$, $T_{000} = -T^{000}$. We shall call this the "sign-flip-if-temporal" rule. As a special case, $-1 = g_{00} = g^{00}$, $0 = g_{0j} = -g^{0j}$, $\delta_{jk} = g_{jk} = g^{jk}$ — i.e., the metric's covariant and contravariant components are numerically identical; they are both equal to the orthonormality values $\eta_{\alpha\beta}$.

(iv) It is easy to see that this sign-flip-if-temporal rule for lowering indices implies the same sign-flip-if-temporal rule for raising them, which in turn can be written in terms of metric components as $T^{\alpha\beta\gamma} = T_{\lambda\mu\nu}g^{\lambda\alpha}g^{\mu\beta}g^{\nu\gamma}$.

(v) It is convenient to define *mixed components* of a tensor, components with some indices up and others down, as having numerical values obtained by raising or lowering some but not all of its indices using the metric, e.g. $T^\alpha{}_{\mu\nu} = T^{\alpha\beta\gamma}g_{\beta\mu}g_{\gamma\nu} = T_{\lambda\mu\nu}g^{\lambda\alpha}$. Numerically, this continues to follow the sign-flip-if-temporal rule: $T^0{}_{0k} = -T^{00k}$, $T^0{}_{jk} = T^{0jk}$, and it implies, in particular, that the mixed components of the metric are $g^\alpha{}_\beta = \delta_{\alpha\beta}$ (the Kronecker-delta values; plus one if $\alpha = \beta$ and zero otherwise).

Summarizing these results: *The numerical values of the components of the metric in Minkowski spacetime are*

$$g_{\alpha\beta} = \eta_{\alpha\beta} \ , \quad g^\alpha{}_\beta = \delta_{\alpha\beta} \ , \quad g_\alpha{}^\beta = \delta_{\alpha\beta} \ , \quad g^{\alpha\beta} = \eta_{\alpha\beta} \ ; \tag{1.32c}$$

and *indices on all vectors and tensors can be raised and lowered using these components of the metric*

$$A_\alpha = g_{\alpha\beta}A^\beta \ , \quad A^\alpha = g^{\alpha\beta}A_\beta \ , \quad T^\alpha{}_{\mu\nu} \equiv g_{\mu\beta}g_{\nu\gamma}T^{\alpha\beta\gamma} \quad T^{\alpha\beta\gamma} \equiv g^{\beta\mu}g^{\gamma\nu}T^\alpha{}_{\mu\nu} \ , \tag{1.32d}$$

which says numerically that *lowering a temporal index changes the component's sign and lowering a spatial index leaves the component unchanged—and similarly for raising indices;* the **sign-flip-if-temporal rule**.

This index notation gives rise to formulas for tensor products, inner products, values of tensors on vectors, and tensor contractions, that are the obvious analogs of those in Euclidean space:

$$[\text{Contravariant components of } \mathbf{T}(\_,\_,\_)\otimes\mathbf{S}(\_,\_)] = T^{\alpha\beta\gamma}S^{\delta\epsilon}\ , \tag{1.32e}$$

$$\vec{A}\cdot\vec{B} = A^\alpha B_\alpha = A_\alpha B^\alpha\ , \quad \mathbf{T}(\mathbf{A},\mathbf{B},\mathbf{C}) = T_{\alpha\beta\gamma}A^\alpha B^\beta C^\gamma = T^{\alpha\beta\gamma}A_\alpha B_\beta C_\gamma\ , \tag{1.32f}$$

$$\begin{aligned} \text{Covariant components of [1\&3contraction of } \mathbf{R}] &= R^\mu{}_{\alpha\mu\beta}\ , \\ \text{Contravariant components of [1\&3contraction of } \mathbf{R}] &= R^{\mu\alpha}{}_\mu{}^\beta\ . \end{aligned} \tag{1.32g}$$

Notice the very simple pattern in Eqs. (1.32), which universally permeates the rules of index gymnastics, a pattern that permits one to reconstruct the rules without any memorization: *Free indices (indices not summed over) must agree in position (up versus down) on the two sides of each equation.* In keeping with this pattern, one often regards the two indices in a pair that is summed (one index up and the other down) as "strangling each other" and thereby being destroyed, and one speaks of "lining up the indices" on the two sides of an equation to get them to agree.

In Part VI, when we use non-orthonormal basis, all of these index-notation equations (1.32) will remain valid unchanged except for the numerical values (1.32c) of the metric components and the sign-flip-if-temporal rule.

1.5.3 [N & R] Slot-Naming Index Notation

[*Note:* In this and other sections marked "N&R", the Newtonian reader should mentally lower all indices on tensor components and make them Latin; e.g. should mentally change $T^{\alpha\beta}{}_\alpha = T^{\alpha\beta\gamma}g_{\alpha\gamma}$ in Eq. (1.33) to $T_{aba} = T_{abc}g_{ac}$.]

We now pause, in our development of the component version of tensor algebra, to introduce a very important new viewpoint:

Consider the rank-2 tensor $\mathbf{F}(\_,\_)$. We can define a new tensor $\mathbf{G}(\_,\_)$ to be the same as $\mathbf{F}$, but with the slots interchanged; i.e., for any two vectors $\vec{A}$ and $\vec{B}$ it is true that $\mathbf{G}(\vec{A},\vec{B}) = \mathbf{F}(\vec{B},\vec{A})$. We need a simple, compact way to indicate that $\mathbf{F}$ and $\mathbf{G}$ are equal except for an interchange of slots. The best way is to give the slots names, say α and β—i.e., to rewrite $\mathbf{F}(\_,\_)$ as $\mathbf{F}(\_{}_\alpha,\_{}_\beta)$ or more conveniently as $F_{\alpha\beta}$; and then to write the relationship between $\mathbf{G}$ and $\mathbf{F}$ as $G_{\alpha\beta} = F_{\beta\alpha}$. *"NO!"* some readers might object. This notation is indistinguishable from our notation for components on a particular basis. *"GOOD!"* a more astute reader will exclaim. The relation $G_{\alpha\beta} = F_{\beta\alpha}$ in a particular basis is a true statement if and only if "$\mathbf{G} = \mathbf{F}$ with slots interchanged" is true, so why not use the same notation to symbolize both? This, in fact, we shall do. We shall ask our readers to look at any "index equation" such as $G_{\alpha\beta} = F_{\beta\alpha}$ like they would look at an Escher drawing: momentarily think of it as a relationship between components of tensors in a specific basis;

then do a quick mind-flip and regard it quite differently, as a relationship between geometric, basis-independent tensors with the indices playing the roles of names of slots. This mind-flip approach to tensor algebra will pay substantial dividends.

As an example of the power of this *slot-naming index notation,* consider the contraction of the first and third slots of a third-rank tensor $\mathbf{T}$. In any basis the components of 1&3contraction($\mathbf{T}$) are $T^{\alpha\beta}{}_{\alpha}$; cf. Eq. (1.32g). Correspondingly, in slot-naming index notation we denote 1&3contraction($\mathbf{T}$) by the simple expression $T^{\alpha\beta}{}_{\alpha}$. We say that the first and third slots are "strangling each other" by the contraction, leaving free only the second slot (named β) and therefore producing a rank-1 tensor (a vector).

By virtue of the "index-lowering" role of the metric, we can also write the contraction as

$$T^{\alpha\beta}{}_{\alpha} = T^{\alpha\beta\gamma} g_{\alpha\gamma} \ , \tag{1.33}$$

and we can look at this relation from either of two viewpoints: The component viewpoint says that the components of the contraction of $\mathbf{T}$ in any chosen basis are obtained by taking a product of components of $\mathbf{T}$ and of the metric $\mathbf{g}$ and then summing over the appropriate indices. The slot-naming viewpoint says that the contraction of $\mathbf{T}$ can be achieved by taking a tensor product of $\mathbf{T}$ with the metric $\mathbf{g}$ to get $\mathbf{T} \otimes \mathbf{g}(\_\,,\_\,,\_\,,\_\,,\_\,)$ (or $T^{\alpha\beta\gamma} g_{\mu\nu}$ in slot-naming index notation), and by then strangling on each other the first and fourth slots [named α in Eq. (1.33)], and also strangling on each other the third and fifth slots [named γ in Eq. (1.33)].

****

EXERCISES

Exercise 1.5 *Derivation: [N & R] Component Manipulation Rules*
If you are studying only the Newtonian part of this book, derive the component manipulation rules (1.29g) and (1.29h); otherwise, derive the relativistic rules (1.32e)–(1.32g). As you proceed, abandon any piece of the exercise when it becomes trivial for you.

Exercise 1.6 *Practice: [N & R] Numerics of Component Manipulations*

(a) In Euclidean space, in some Cartesian basis, the third rank tensor $\mathbf{S}(\_\,,\_\,,\_\,)$ and vectors $\mathbf{A}$ and $\mathbf{B}$ have as their only nozero components $S_{123} = S_{231} = S_{312} = +1$, $A_1 = 3$, $B_1 = 4$, $B_2 = 5$. What are the components of the vector $\mathbf{S}(\mathbf{A}, \mathbf{B}, \_\,)$, the vector $\mathbf{S}(\mathbf{A}, \_\,, \mathbf{B})$ and the tensor $\mathbf{A} \otimes \mathbf{B}$?

(b) In Minkowski spacetime, in some inertial reference frame, the vector $\vec{A}$ and second-rank tensor $\mathbf{T}$ have as their only nonzero components $A^0 = 1$, $A^1 = 2$, $A^2 = A^3 = 0$; $T^{00} = 3$, $T^{01} = T^{10} = 2$, $T^{11} = -1$. Evaluate $\mathbf{T}(\vec{A}, \vec{A})$ and the components of $\mathbf{T}(\vec{A}, \_\,)$ and $\vec{A} \otimes \mathbf{T}$.

Exercise 1.7 *Practice: [N & R] Meaning of Slot-Naming Index Notation*

(a) In Euclidean space, the following expressions and equations are written in slot-naming index notation; convert them to geometric, index-free notation: $A_i B_{jk}$; $A_i B_{ji}$, $S_{ijk} = S_{kji}$, $A_i B_i = A_i B_j g_{ij}$.

(b) In Euclidean space, the following expressions are written in geometric, index-free notation; convert them to slot-naming index notation: $\mathsf{T}(\_\,, \_\,, \mathbf{A})$; $\mathsf{T}(\_\,, \mathsf{S}(\mathbf{B}, \_\,), \_\,)$.

(c) In Minkowski spacetime, convert $\mathsf{T}(\_\,, \mathsf{S}(\mathsf{R}(\vec{C}, \_\,), \_\,), \_\,)$ into slot-naming index notation.

Exercise 1.8 *Practice: [R] Index Gymnastics*

(a) Simplify the following expression so that the metric does not appear in it: $A^{\alpha\beta\gamma} g_{\beta\rho} S_{\gamma\lambda} g^{\rho\delta} g^{\lambda}{}_{\alpha}$.

(b) The quantity $g_{\alpha\beta} g^{\alpha\beta}$ is a scalar since it has no free indices. What is its numerical value?

(c) What is wrong with the following expression and equation? $A_{\alpha}{}^{\beta\gamma} S_{\alpha\gamma}$; $A_{\alpha}{}^{\beta\gamma} S_{\beta} T_{\gamma} = R_{\alpha\beta\delta} S^{\beta}$.

****

1.6 [R] Particle Kinetics in Index Notation and in a Lorentz Frame

As an illustration of the component representation of tensor algebra, let us return to the relativistic, accelerated particle of Fig. 1.7 and, from the frame-independent equations for the particle's 4-velocity $\vec{u}$ and 4-momentum $\vec{p}$ (Sec. 1.4), derive the component description given in elementary textbooks.

We introduce a specific inertial reference frame and associated Lorentz coordinates x^α and basis vectors $\{\vec{e}_\alpha\}$. In this Lorentz frame, the particle's world line $\vec{x}(\tau)$ is represented by its coordinate location $x^\alpha(\tau)$ as a function of its proper time τ. The contravariant components of the separation vector $d\vec{x}$ between two neighboring events along the particle's world line are the events' coordinate separations dx^α [Eq. (1.2)—which is why we put the indices up on coordinates]; and correspondingly, the components of the particle's 4-velocity $\vec{u} = d\vec{x}/d\tau$ are

$$u^\alpha = \frac{dx^\alpha}{d\tau} \tag{1.34a}$$

(the time derivatives of the particle's spacetime coordinates). Note that Eq. (1.34a) implies

$$v^j \equiv \frac{dx^j}{dt} = \frac{dx^j/d\tau}{dt/d\tau} = \frac{u^j}{u^0}\ . \tag{1.34b}$$

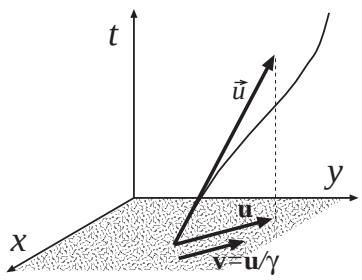


Fig. 1.10: Spacetime diagram in a specific Lorentz frame, showing the frame's 3-space $t = 0$ (stippled region), the 4-velocity $\vec{u}$ of a particle as it passes through that 3-space (i.e., at time $t = 0$); and two 3-dimensional vectors that lie in the 3-space: the spatial part of the particle's 4-velocity, **u**, and the particle's ordinary velocity **v**.

Here v^j are the components of the *ordinary velocity* as measured in the Lorentz frame. This relation, together with the unit norm of $\vec{u}$, $\vec{u}^2 = g_{\alpha\beta}u^\alpha u^\beta = -(u^0)^2 + \delta_{ij}u^i u^j = -1$, implies that the components of the 4-velocity have the forms familiar from elementary textbooks:

$$u^0 = \gamma \ , \quad u^j = \gamma v^j \ , \quad \text{where} \quad \gamma = \frac{1}{(1 - \delta_{ij}v^i v^j)^{\frac{1}{2}}} \ . \tag{1.34c}$$

It is useful to think of v^j as the components of a 3-dimensional vector **v**, the ordinary velocity, that lives in the 3-dimensional Euclidean space $t =$ const of the chosen Lorentz frame. As we shall see below, this 3-space is not well defined until a Lorentz frame has been chosen, and correspondingly, **v** relies for its existence on a specific choice of frame. However, once the frame has been chosen, **v** can be regarded as a coordinate-independent, basis-independent 3-vector lying in the frame's 3-space $t =$const. Similarly, the spatial part of the 4-velocity $\vec{u}$ (the part with components u^j in our chosen frame) can be regarded as a 3-vector **u** lying in the frame's 3-space; and Eqs. (1.34c) become the component versions of the coordinate-independent, basis-independent 3-space relations

$$\mathbf{u} = \gamma \mathbf{v} \ , \quad \gamma = \frac{1}{\sqrt{1 - \mathbf{v}^2}} \ . \tag{1.34d}$$

Figure 1.10 shows stippled the 3-space $t = 0$ of a specific Lorentz frame, and the 4-velocity $\vec{u}$ and ordinary velocity **v** of a particle as it passes through that 3-space.

The components of the particle's 4-momentum $\vec{p}$ in our chosen Lorentz frame have special names and special physical significances: The time component of the 4-momentum is the particle's *energy* $\mathcal{E}$ as measured in that frame

$$\begin{aligned} \mathcal{E} \equiv p^0 = mu^0 = m\gamma = \frac{m}{\sqrt{1 - \mathbf{v}^2}} = \ &\text{(the particle's energy)} \\ \simeq m + \frac{1}{2}m\mathbf{v}^2 \quad &\text{for } v \equiv |\mathbf{v}| \ll 1 \ . \end{aligned} \tag{1.35a}$$

Note that this energy is the sum of the particle's *rest mass-energy* $m = mc^2$ and its *kinetic energy* $m\gamma - m$ (which, for low velocities, reduces to the familiar nonrelativistic kinetic

energy $E = \frac{1}{2}m\mathbf{v}^2$). The spatial components of the 4-momentum, when regarded from the viewpoint of 3-dimensional physics, are the same as the components of the *momentum*, a 3-vector residing in the chosen Lorentz frame's 3-space:

$$p^j = mu^j = m\gamma v^j = \frac{mv^j}{\sqrt{1-\mathbf{v}^2}} = \mathcal{E}v_j = (j\text{-component of particle's momentum}) \; ; \quad (1.35b)$$

or, in basis-independent, 3-dimensional vector notation,

$$\mathbf{p} = m\mathbf{u} = m\gamma\mathbf{v} = \frac{m\mathbf{v}}{\sqrt{1-\mathbf{v}^2}} = \mathcal{E}\mathbf{v} = \text{ (particle's momentum)} \; . \quad (1.35c)$$

For a zero-rest-mass particle, as for one with finite rest mass, we identify the time component of the 4-momentum, in a chosen Lorentz frame, as the particle's energy, and the spatial part as its momentum. Moreover, if—appealing to quantum theory—we regard a zero-rest-mass particle as a quantum associated with a monochromatic wave, then quantum theory tells us that the wave's angular frequency ω as measured in a chosen Lorentz frame will be related to its energy by

$$\mathcal{E} \equiv p^0 = \hbar\omega = (\text{particle's energy}) \; ; \quad (1.36a)$$

and, since the particle has $\vec{p}^2 = -(p^0)^2 + \mathbf{p}^2 = -m^2 = 0$ (in accord with the lightlike nature of its world line), its momentum as measured in the chosen Lorentz frame will be

$$\mathbf{p} = \mathcal{E}\mathbf{n} = \hbar\omega\mathbf{n} \; . \quad (1.36b)$$

Here $\mathbf{n}$ is the unit 3-vector that points in the direction of travel of the particle, as measured in the chosen frame; i.e. (since the particle moves at the speed of light $v = 1$), $\mathbf{n}$ is the particle's ordinary velocity. Eqs. (1.36a) and (1.36b) are the temporal and spatial components of the geometric, frame-independent relation $\vec{p} = \hbar\vec{k}$ [Eq. (1.21), which is valid for zero-rest-mass particles as well as finite-mass ones].

The introduction of a specific Lorentz frame into spacetime can be said to produce a "3+1" split of every 4-vector into a 3-dimensional vector plus a scalar (a real number). The 3+1 split of a particle's 4-momentum $\vec{p}$ produces its momentum $\mathbf{p}$ plus its energy $\mathcal{E} = p^0$; and correspondingly, the 3+1 split of the law of 4-momentum conservation (1.23) produces a law of conservation of momentum plus a law of conservation of energy:

$$\sum_{\bar{A}} \mathbf{p}_{\bar{A}} = \sum_{A} \mathbf{p}_A \; , \quad \sum_{\bar{A}} \mathcal{E}_{\bar{A}} = \sum_{A} \mathcal{E}_A \; . \quad (1.37)$$

Here the unbarred quantities are the momenta or energies of the particles entering the interaction region, and the barred quantities are the momenta or energies of those leaving; cf. Fig. 1.8.

Because the concept of energy does not even exist until one has chosen a Lorentz frame, and neither does that of momentum, the laws of energy conservation and momentum conservation separately are frame-dependent laws. In this sense they are far less fundamental than their combination, the frame-independent law of 4-momentum conservation.

Box 1.3
[N] Relativistic Particles for Newtonian Readers

Readers who are skipping the relativistic parts of this book will need to know two important pieces of relativity: (i) geometrized units, as embodied in Eqs. (1.3), and (ii) the (relativistic) energy and momentum of a moving particle, as described here:

A particle with rest mass m, moving with ordinary velocity $\mathbf{v} = d\mathbf{x}/dt$ and speed $v = |\mathbf{v}|$, has energy $\mathcal{E}$ (including its rest-mass), energy E and momentum $\mathbf{p}$ given by

$$\mathcal{E} = \frac{m}{\sqrt{1-v^2}} \equiv \frac{m}{\sqrt{1-v^2/c^2}} \equiv E + m \ , \quad \mathbf{p} = \mathcal{E}\mathbf{v} = \frac{m\mathbf{v}}{\sqrt{1-v^2}} \ . \tag{1}$$

[Eqs. (1.35)]. In the low-velocity, Newtonian limit, the energy E with rest mass removed and the momentum $\mathbf{p}$ and take their familiar, Newtonian forms:

$$\text{When } v \ll c \equiv 1, \quad E \to \frac{1}{2}mv^2 \quad \text{ and } \mathbf{p} \to m\mathbf{v} \ . \tag{2}$$

A particle with zero rest mass (a photon or a graviton[7]) always moves with the speed of light $v = c = 1$, and like other particles it has has momentum $\mathbf{p} = \mathcal{E}\mathbf{v}$, so the magnitude of its momentum is equal to its energy: $|\mathbf{p}| = \mathcal{E}v = \mathcal{E}$.

When particles interact (e.g. in chemical reactions, nuclear reactions, and elementary-particle collisons) the sum of the particle energies $\mathcal{E}$ is conserved, as is the sum of the particle momenta $\mathbf{p}$: Eq. (1.37).

By learning to think about the 3+1 split in a geometric, frame-independent way, one can gain much conceptual and computational power. As a example, consider a particle with 4-momentum $\vec{p}$, as studied by an observer with 4-velocity $\vec{U}$. In the observer's own Lorentz reference frame, her 4-velocity has components $U^0 = 1$ and $U^j = 0$, and therefore, her 4-velocity is $\vec{U} = U^\alpha \vec{e}_\alpha = \vec{e}_0$, i.e. it is identically equal to the time basis vector of her Lorentz frame. This means that the particle energy that she measures is $\mathcal{E} = p^0 = -p_0 = -\vec{p}\cdot\vec{e}_0 = -\vec{p}\cdot\vec{U}$. This equation, derived in the observer's Lorentz frame, is actually a geometric, frame-independent relation: the inner product of two 4-vectors. It says that *when an observer with 4-velocity $\vec{U}$ measures the energy of a particle with 4-momentum $\vec{p}$, the result she gets (the time part of the 3+1 split of $\vec{p}$ as seen by her) is*

$$\boxed{\mathcal{E} = -\vec{p}\cdot\vec{U} \ .} \tag{1.38}$$

We shall use this equation fairly often in later chapters. In Exs. 1.9 and 1.10, the reader can get experience at deriving and interpreting other frame-independent equations for 3+1 splits. Exercise 1.11 exhibits the power of this geometric way of thinking by using it to derive the Doppler shift of a photon.

****

EXERCISES

Exercise 1.9 ***Practice: [R] Frame-Independent Expressions for Energy, Momentum, and Velocity*[8]
An observer with 4-velocity $\vec{U}$ measures the properties of a particle with 4-momentum $\vec{p}$. The energy she measures is $\mathcal{E} = -\vec{p}\cdot\vec{E}$, Eq. (1.38).

(a) Show that the rest mass the observer measures is computable from

$$m^2 = -\vec{p}^2 \ . \tag{1.39a}$$

(b) Show that the momentum the observer measures has the magnitude

$$|\mathbf{p}| = [(\vec{p}\cdot\vec{U})^2 + \vec{p}\cdot\vec{p}]^{\frac{1}{2}} \ . \tag{1.39b}$$

(c) Show that the ordinary velocity the observer measures has the magnitude

$$|\mathbf{v}| = \frac{|\mathbf{p}|}{\mathcal{E}}, \tag{1.39c}$$

where $|\mathbf{p}|$ and $\mathcal{E}$ are given by the above frame-independent expressions.

(d) Show that the ordinary velocity $\mathbf{v}$, thought of as a 4-vector that happens to lie in the observer's 3-space of constant time, is given by

$$\vec{v} = \frac{\vec{p} + (\vec{p}\cdot\vec{U})\vec{U}}{-\vec{p}\cdot\vec{U}} \ . \tag{1.39d}$$

Exercise 1.10 ***Example: [R] 3-Metric as a Projection Tensor*
Consider, as in Exercise 1.9, an observer with 4-velocity $\vec{U}$ who measures the properties of a particle with 4-momentum $\vec{p}$.

(a) Show that the Euclidean metric of the observer's 3-space, when thought of as a tensor in 4-dimensional spacetime, has the form

$$\boxed{\mathbf{P} \equiv \mathbf{g} + \vec{U}\otimes\vec{U} \ .} \tag{1.40a}$$

Show, further, that if $\vec{A}$ is an arbitrary vector in spacetime, then $-\vec{A}\cdot\vec{U}$ is the component of $\vec{A}$ along the observer's 4-velocity $\vec{U}$, and

$$\mathbf{P}(\_\,, \vec{A}) = \vec{A} + (\vec{A}\cdot\vec{U})\vec{U} \tag{1.40b}$$

is the projection of $\vec{A}$ into the observer's 3-space; i.e., it is the spatial part of $\vec{A}$ as seen by the observer. For this reason, $\mathbf{P}$ is called a *projection tensor*. In quantum mechanics one introduces the concept of a *projection operator* $\hat{P}$ as an operator that satisfies the equation $\hat{P}^2 = \hat{P}$. Show that the projection tensor $\mathbf{P}$ is a projection operator in the quantum mechanical sense:

$$P_{\alpha\mu}P^{\mu}{}_{\beta} = P_{\alpha\beta} \ . \tag{1.40c}$$

[8] Exercises marked with double stars are important expansions of the material presented in the text.

(b) Show that Eq. (1.39d) for the particle's ordinary velocity, thought of as a 4-vector, can be rewritten as

$$\vec{v} = \frac{\mathbf{P}(\_, \vec{p})}{-\vec{p} \cdot \vec{U}} \ . \tag{1.41}$$

Exercise 1.11 ***Example: [R] Doppler Shift Derived without Lorentz Transformations*

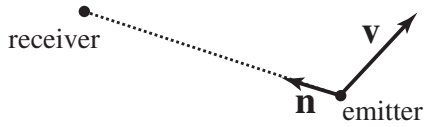


Fig. 1.11: Geometry for Doppler shift.

(a) An observer at rest in some inertial frame receives a photon that was emitted in a direction $\mathbf{n}$ by an atom moving with ordinary velocity $\mathbf{v}$ (Fig. 1.11). The photon frequency and energy as measured by the emitting atom are $\nu_{\rm em}$ and $\mathcal{E}_{\rm em}$; those measured by the receiving observer are $\nu_{\rm rec}$ and $\mathcal{E}_{\rm em}$. By a calculation carried out solely in the receiver's inertial frame (the frame of Fig. 1.11), and without the aid of any Lorentz transformation, derive the standard formula for the photon's Doppler shift,

$$\frac{\nu_{\rm rec}}{\nu_{\rm em}} = \frac{\sqrt{1-v^2}}{1-\mathbf{v}\cdot\mathbf{n}} \ . \tag{1.42}$$

Hint: Use Eq. (1.38) to evaluate $\mathcal{E}_{\rm em}$ using receiver-frame expressions for the emitting atom's 4-velocity $\vec{U}$ and the photon's 4-momentum $\vec{p}$.

(b) Suppose that instead of emitting a photon, the emitter ejects a particle with finite rest mass m. Using the same method, derive an expression for the ratio of received energy to emitted energy, $\mathcal{E}_{\rm rec}/\mathcal{E}_{\rm em}$, expressed in terms of the emitter's ordinary velocity $\mathbf{v}$ and the particle's ordinary velocity $\mathbf{V}$ (both as measured in the receiver's frame).

****

1.7 Orthogonal and Lorentz Transformations of Bases, and Spacetime Diagrams

1.7.1 [N] Euclidean 3-space: Orthogonal Transformations

Consider two different Cartesian coordinate systems $\{x, y, z\} \equiv \{x_1, x_2, x_3\}$, and $\{\bar{x}, \bar{y}, \bar{z}\} \equiv \{x_{\bar{1}}, x_{\bar{2}}, x_{\bar{3}}\}$. Denote by $\{\mathbf{e}_i\}$ and $\{\mathbf{e}_{\bar{p}}\}$ the corresponding bases. It must be possible to expand

the basis vectors of one basis in terms of those of the other. We shall denote the expansion coefficients by the letter R and shall write

$$\mathbf{e}_i = \mathbf{e}_{\bar{p}} R_{\bar{p}i} \ , \quad \mathbf{e}_{\bar{p}} = \mathbf{e}_i R_{i\bar{p}} \ . \tag{1.43}$$

The quantities $R_{\bar{p}i}$ and $R_{i\bar{p}}$ are *not* the components of a tensor; rather, they are the elements of transformation matrices

$$[R_{\bar{p}i}] = \begin{bmatrix} R_{\bar{1}1} & R_{\bar{1}2} & R_{\bar{1}3} \\ R_{\bar{2}1} & R_{\bar{2}2} & R_{\bar{2}3} \\ R_{\bar{3}1} & R_{\bar{3}2} & R_{\bar{3}3} \end{bmatrix} \ , \quad [R_{i\bar{p}}] = \begin{bmatrix} R_{1\bar{1}} & R_{1\bar{2}} & R_{1\bar{3}} \\ R_{2\bar{1}} & R_{2\bar{2}} & R_{2\bar{3}} \\ R_{3\bar{1}} & R_{3\bar{2}} & R_{3\bar{3}} \end{bmatrix} \ . \tag{1.44a}$$

(Here and throughout this book we use square brackets to denote matrices.) These two matrices must be the inverse of each other, since one takes us from the barred basis to the unbarred, and the other in the reverse direction, from unbarred to barred:

$$R_{\bar{p}i} R_{i\bar{q}} = \delta_{\bar{p}\bar{q}} \ , \quad R_{i\bar{p}} R_{\bar{p}j} = \delta_{ij} \ . \tag{1.44b}$$

The orthonormality requirement for the two bases implies that $\delta_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j = (\mathbf{e}_{\bar{p}} R_{\bar{p}i}) \cdot (\mathbf{e}_{\bar{q}} R_{\bar{q}j}) = R_{\bar{p}i} R_{\bar{q}j} (\mathbf{e}_{\bar{p}} \cdot \mathbf{e}_{\bar{q}}) = R_{\bar{p}i} R_{\bar{q}j} \delta_{\bar{p}\bar{q}} = R_{\bar{p}i} R_{\bar{p}j}$. This says that the transpose of $[R_{\bar{p}i}]$ is its inverse—which we have already denoted by $[R_{i\bar{p}}]$;

$$R_{i\bar{p}} = R_{\bar{p}i} \ . \tag{1.44c}$$

This property implies that the transformation matrix is orthogonal; i.e., the transformation is a reflection or a rotation [see, e.g., Goldstein (1980)]. Thus (as should be obvious and familiar), the bases associated with any two Euclidean coordinate systems are related by a reflection or rotation. Note: Eq. (1.44c) does *not* say that $[R_{i\bar{p}}]$ is a symmetric matrix; in fact, it typically is not. Rather, (1.44c) says that $[R_{i\bar{p}}]$ is the transpose of $[R_{\bar{p}i}]$.]

The fact that a vector $\mathbf{A}$ is a geometric, basis-independent object implies that $\mathbf{A} = A_i \mathbf{e}_i = A_i (\mathbf{e}_{\bar{p}} R_{\bar{p}i}) = (R_{\bar{p}i} A_i) \mathbf{e}_{\bar{p}} = A_{\bar{p}} \mathbf{e}_{\bar{p}}$; i.e.,

$$A_{\bar{p}} = R_{\bar{p}i} A_i \ , \quad \text{and similarly } A_i = R_{i\bar{p}} A_{\bar{p}} \ ; \tag{1.45a}$$

and correspondingly for the components of a tensor

$$T_{\bar{p}\bar{q}\bar{r}} = R_{\bar{p}i} R_{\bar{q}j} R_{\bar{r}k} T_{ijk} \ , \quad T_{ijk} = R_{i\bar{p}} R_{j\bar{q}} R_{k\bar{r}} T_{\bar{p}\bar{q}\bar{r}} \ . \tag{1.45b}$$

It is instructive to compare the transformation law (1.45a) for the components of a vector with those (1.43) for the bases. To make these laws look natural, we have placed the transformation matrix on the left in the former and on the right in the latter. In Minkowski spacetime, the placement of indices, up or down, will automatically tell us the order.

If we choose the origins of our two coordinate systems to coincide, then the vector $\mathbf{x}$ reaching from the their common origin to some point $\mathcal{P}$ whose coordinates are x_j and $x_{\bar{p}}$ has components equal to those coordinates; and as a result, the coordinates themselves obey the same transformation law as any other vector

$$x_{\bar{p}} = R_{\bar{p}i} x_i \ , \quad x_i = R_{i\bar{p}} x_{\bar{p}} \ ; \tag{1.45c}$$

The product of two rotation matrices, $[R_{i\bar{p}} R_{\bar{p}\bar{s}}]$ is another rotation matrix $[R_{i\bar{s}}]$, which transforms the Cartesian bases $\mathbf{e}_{\bar{s}}$ to $\mathbf{e}_i$. Under this product rule, the rotation matrices form a mathematical *group*: the *rotation group*, whose "representations" play an important role in quantum theory.

1.7.2 [R] Minkowski Spacetime: Lorentz Transformations

Consider two different inertial reference frames in Minkowski spacetime. Denote their Lorentz coordinates by $\{x^\alpha\}$ and $\{x^{\bar\mu}\}$ and their bases by $\{\mathbf{e}_\alpha\}$ and $\{\mathbf{e}_{\bar\mu}\}$, and write the transformation from one basis to the other as

$$\vec{e}_\alpha = \vec{e}_{\bar\mu} L^{\bar\mu}{}_\alpha \ , \quad \vec{e}_{\bar\mu} = \vec{e}_\alpha L^\alpha{}_{\bar\mu} \ . \tag{1.46}$$

As in Euclidean 3-space, $L^{\bar\mu}{}_\alpha$ and $L^\alpha{}_{\bar\mu}$ are elements of two different transformation matrices, and since these matrices operate in opposite directions, they must be the inverse of each other:

$$L^{\bar\mu}{}_\alpha L^\alpha{}_{\bar\nu} = \delta^{\bar\mu}{}_{\bar\nu} \ , \quad L^\alpha{}_{\bar\mu} L^{\bar\mu}{}_\beta = \delta^\alpha{}_\beta \ . \tag{1.47a}$$

Notice the up/down placement of indices on the elements of the transformation matrices: the first index is always up, and the second is always down. This is just a convenient convention which helps systematize the index shuffling rules in a way that can be easily remembered. Our rules about summing on the same index when up and down, and matching unsummed indices on the two sides of an equation, automatically dictate the matrix to use in each of the transformations (1.46); and similarly for all other equations in this section.

In Euclidean 3-space the orthonormality of the two bases dictated that the transformations must be orthogonal, i.e. must be reflections or rotations. In Minkowski spacetime, orthonormality implies $g_{\alpha\beta} = \vec{e}_\alpha \cdot \vec{e}_\beta = (\vec{e}_{\bar\mu} L^{\bar\mu}{}_\alpha) \cdot (\vec{e}_{\bar\nu} L^{\bar\nu}{}_\beta) = L^{\bar\mu}{}_\alpha L^{\bar\nu}{}_\beta g_{\bar\mu\bar\nu}$; i.e.,

$$g_{\bar\mu\bar\nu} L^{\bar\mu}{}_\alpha L^{\bar\nu}{}_\beta = g_{\alpha\beta} \ , \quad \text{and similarly } g_{\alpha\beta} L^\alpha{}_{\bar\mu} L^\beta{}_{\bar\nu} = g_{\bar\mu\bar\nu} \ . \tag{1.47b}$$

Any matrices whose elements satisfy these equations is a *Lorentz transformation*.

From the fact that vectors and tensors are geometric, frame-independent objects, one can derive the Minkowski-space analogs of the Euclidean transformation laws for components (1.45a), (1.45b):

$$A^{\bar\mu} = L^{\bar\mu}{}_\alpha A^\alpha \ , \quad T^{\bar\mu\bar\nu\bar\rho} = L^{\bar\mu}{}_\alpha L^{\bar\nu}{}_\beta L^{\bar\rho}{}_\gamma T^{\alpha\beta\gamma} \ , \quad \text{and similarly in the opposite direction.} \tag{1.48a}$$

Notice that here, as elsewhere, these equations can be constructed by lining up indices in accord with our standard rules.

If (as is conventional) we choose the spacetime origins of the two Lorentz coordinate systems to coincide, then the vector $\vec{x}$ extending from the origin to some event $\mathcal{P}$, whose coordinates are x^α and $x^{\bar\alpha}$, has components equal to those coordinates. As a result, the transformation law for the coordinates takes the same form as that (1.48a) for components of a vector:

$$x^\alpha = L^\alpha{}_{\bar\mu} x^{\bar\mu} \ , \quad x^{\bar\mu} = L^{\bar\mu}{}_\alpha x^\alpha \ . \tag{1.48b}$$

The product $[L^\alpha{}_{\bar\mu} L^{\bar\mu}{}_{\bar{\bar\rho}}]$ of two Lorentz transformation matrices is a Lorentz transformation matrix; and under this product rule, the Lorentz transformations form a mathematical group, the *Lorentz group*, whose "representations" play an important role in quantum field theory.

An important specific example of a Lorentz transformation is the following

$$|| L^\alpha{}_{\bar\mu} || = \begin{bmatrix} \gamma & \beta\gamma & 0 & 0 \\ \beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \ , \quad || L^{\bar\mu}{}_\alpha || = \begin{bmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \ , \tag{1.49a}$$

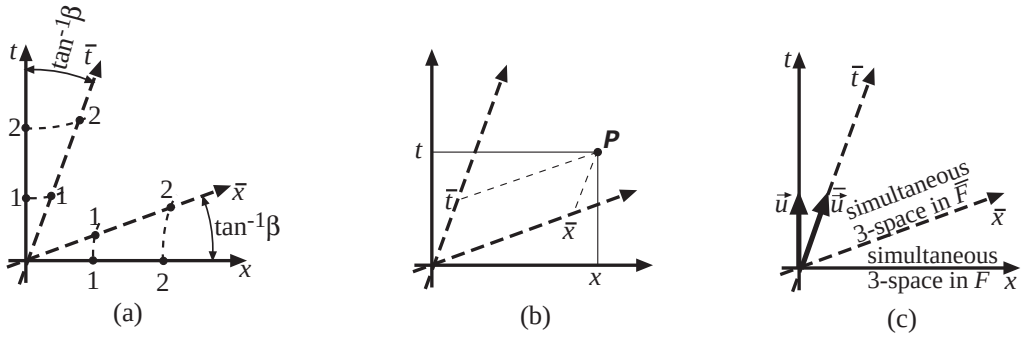


Fig. 1.12: Spacetime diagrams illustrating the pure boost (1.49c) from one Lorentz reference frame to another.

where β and γ are related by

$$|\beta| < 1 \ , \qquad \gamma \equiv (1-\beta^2)^{-\frac{1}{2}} \ . \tag{1.49b}$$

One can readily verify that these matrices are the inverses of each other and that they satisfy the Lorentz-transformation relation (1.47b). These transformation matrices produce the following change of coordinates [Eq. (1.48b)]

$$\begin{aligned} t &= \gamma(\bar{t} + \beta\bar{x}) \ , \quad x = \gamma(\bar{x} + \beta\bar{t}) \ , \quad y = \bar{y} \ , \quad z = \bar{z} \ , \\ \bar{t} &= \gamma(t - \beta x) \ , \quad \bar{x} = \gamma(x - \beta t) \ , \quad \bar{y} = y \ , \quad \bar{z} = z \ . \end{aligned} \tag{1.49c}$$

These expressions reveal that any point at rest in the unbarred frame (a point with fixed, time-independent x, y, z) is seen in the barred frame to move along the world line $\bar{x} = \text{const} - \beta\bar{t}$, $\bar{y} = \text{const}$, $\bar{z} = \text{const}$. In other words, the unbarred frame is seen by observers at rest in the barred frame to move with uniform velocity $\vec{v} = -\beta\vec{e}_{\bar{x}}$, and correspondingly the barred frame is seen by observers at rest in the unbarred frame to move with the opposite uniform velocity $\vec{v} = +\beta\vec{e}_x$. This special Lorentz transformation is called a *pure boost* along the x direction.

1.7.3 [R] Spacetime Diagrams for Boosts

Figure 1.12 illustrates the pure boost (1.49c). Diagram (a) in that figure is a two-dimensional spacetime diagram, with the y- and z-coordinates suppressed, showing the $\bar{t}$ and $\bar{x}$ axes of the boosted Lorentz frame $\bar{F}$ in the t, x Lorentz coordinate system of the unboosted frame F. That the barred axes make angles $\tan^{-1}\beta$ with the unbarred axes, as shown, can be inferred from the Lorentz transformation equation (1.49c). Note that invariance of the interval guarantees that the event $\bar{x} = a$ on the $\bar{x}$-axis lies at the intersection of that axis with the dashed hyperbola $x^2 - t^2 = a^2$; and similarly, the event $\bar{t} = a$ on the $\bar{t}$-axis lies at the intersection of that axis with the dashed hyperbola $t^2 - x^2 = a^2$. As is shown in diagram (b) of the figure, the barred coordinates $\bar{t}$, $\bar{x}$ of an event $\mathcal{P}$ can be inferred by projecting from $\mathcal{P}$ onto the $\bar{t}$- and $\bar{x}$-axes, with the projection going parallel to the $\bar{x}$- and $\bar{t}$-axes respectively. Diagram (c) shows the 4-velocity $\vec{u}$ of an observer at rest in frame F and

that, $\bar{\vec{u}}$ of an observer in frame $\bar{F}$. The events which observer F regards as all simultaneous, with time $t = 0$, lie in a 3-space that is orthogonal to $\vec{u}$ and includes the x-axis. This is the *Euclidean 3-space of reference frame* F and is also sometimes called F's *3-space of simultaneity*. Similarly, the events which observer $\bar{F}$ regards as all simultaneous, with $\bar{t} = 0$, live in the 3-space that is orthogonal to $\bar{\vec{u}}$ and includes the $\bar{x}$-axis. This is the Euclidean 3-space (3-space of simultaneity) of frame $\bar{F}$.

Exercise 1.14 uses spacetime diagrams, similar to Fig. 1.12, to deduce a number of important relativistic phenomena, including the contraction of the length of a moving object ("length contraction"), the breakdown of simultaneity as a universally agreed upon concept, and the dilation of the ticking rate of a moving clock ("time dilation"). This exercise is extremely important; every reader who is not already familiar with it should study it.

****

EXERCISES

Exercise 1.12 *Derivation: [R] The Inverse of a Lorentz Boost*
Show that, if the Lorentz coordinates of an inertial frame F are expressed in terms of those of the frame $\bar{F}$ by Eq. (1.49c), then the inverse transformation from F to $\bar{F}$ is given by the same equation with the sign of β reversed. Write down the corresponding transformation matrix $L^{\bar{\mu}}{}_{\alpha}$ [analog of Eq. (1.49a)].

Exercise 1.13 *Problem: [R] Allowed and Forbidden Electron-Photon Reactions*
Show, using spacetime diagrams and also using frame-independent calculations, that the law of conservation of 4-momentum forbids a photon to be absorbed by an electron, $e + \gamma \to e$ and also forbids an electron and a positron to annihilate and produce a single photon $e^+ + e^- \to \gamma$ (in the absence of any other particles to take up some of the 4-momentum); but the annihilation to form two photons, $e^+ + e^- \to 2\gamma$, is permitted.

Exercise 1.14 ***Example: [R] Spacetime Diagrams*
Use spacetime diagrams to prove the following:

(a) *Two events that are simultaneous in one inertial frame are not necessarily simultaneous in another.* More specifically, if frame $\bar{F}$ moves with velocity $\vec{v} = \beta\vec{e}_x$ as seen in frame F, where $\beta > 0$, then of two events that are simultaneous in $\bar{F}$ the one farther "back" (with the more negative value of $\bar{x}$) will occur in F before the one farther "forward".

(b) Two events that occur at the same spatial location in one inertial frame do not necessarily occur at the same spatial location in another.

(c) If $\mathcal{P}_1$ and $\mathcal{P}_2$ are two events with a timelike separation, then there exists an inertial reference frame in which they occur at the same spatial location; and in that frame the time lapse between them is equal to the square root of the negative of their invariant interval, $\Delta t = \Delta\tau \equiv \sqrt{-\Delta s^2}$.

(d) If $\mathcal{P}_1$ and $\mathcal{P}_2$ are two events with a spacelike separation, then there exists an inertial reference frame in which they are simultaneous; and in that frame the spatial distance between them is equal to the square root of their invariant interval, $\sqrt{g_{ij}\Delta x^i \Delta x^j} = \Delta s \equiv \sqrt{\Delta s^2}$.

(e) If the inertial frame $\bar{F}$ moves with speed β relative to the frame F, then a clock at rest in $\bar{F}$ ticks more slowly as viewed from F than as viewed from $\bar{F}$—more slowly by a factor $\gamma^{-1} = (1-\beta^2)^{\frac{1}{2}}$. This is called *relativistic time dilation.*

(f) If the inertial frame $\bar{F}$ moves with velocity $\vec{v} = \beta \vec{e}_x$ relative to the frame F and the two frames are related by a pure boost, then an object at rest in $\bar{F}$ as studied in F appears shortened by a factor $\gamma^{-1} = (1-\beta^2)^{\frac{1}{2}}$ along the x direction, but its length along the y and z directions is unchanged. This is called *Lorentz contraction.*

Exercise 1.15 *Example: [R] General Boosts and Rotations*

(a) Show that, if n^j is a 3-dimensional unit vector and β and γ are defined as in Eq. (1.49b), then the following is a Lorentz transformation; i.e., it satisfies Eq. (1.47b).

$$L^{0}{}_{\bar{0}} = \gamma \ , \quad L^{0}{}_{\bar{j}} = L^{j}{}_{\bar{0}} = \beta\gamma n^j \ , \quad L^{j}{}_{\bar{k}} = L^{k}{}_{\bar{j}} = (\gamma - 1) n^j n^k + \delta^{jk} \ . \tag{1.50}$$

Show, further, that this transformation is a *pure boost along the direction* $\mathbf{n}$ *with speed* β, and show that the inverse matrix $[L^{\bar{\mu}}{}_{\alpha}]$ for this boost is the same as $[L^{\alpha}{}_{\bar{\mu}}]$, but with β changed to $-\beta$.

(b) Show that the following is also a Lorentz transformation:

$$|| \, L^{\alpha}{}_{\bar{\mu}} \, || = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & & & \\ 0 & & [R_{i\bar{j}}] & \\ 0 & & & \end{bmatrix} \ , \tag{1.51}$$

where $[R_{i\bar{j}}]$ is a three-dimensional rotation matrix for Euclidean 3-space. Show, further, that this Lorentz transformation rotates the inertial frame's spatial axes (its latticework of measuring rods), while leaving the frame's velocity unchanged; i.e., the new frame is at rest with respect to the old.

One can show (not surprisingly) that the general Lorentz transformation [i.e., the general solution of Eqs. (1.47b)] can be expressed as a sequence of pure boosts, pure rotations, and pure inversions (in which one or more of the coordinate axes are reflected through the origin, so $x^{\alpha} = -x^{\bar{\alpha}}$).

****

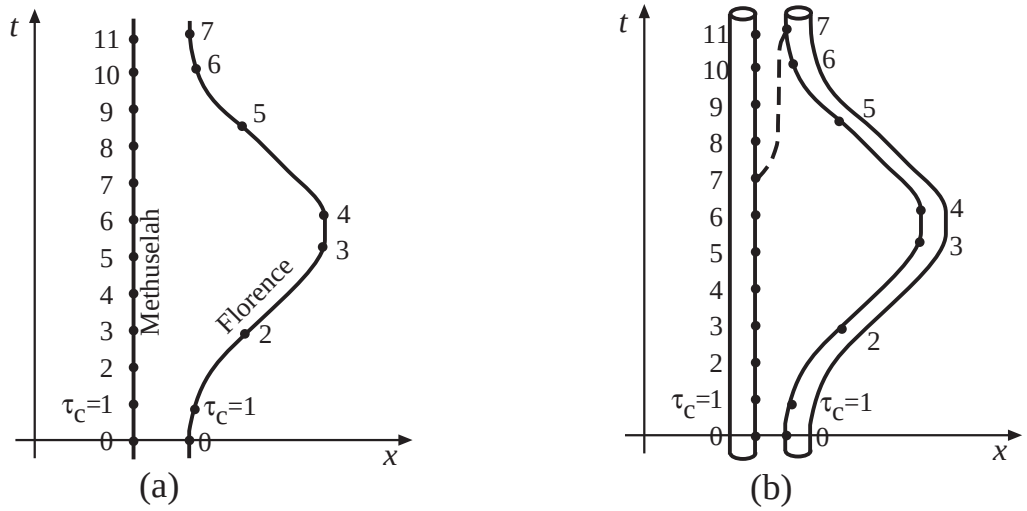


Fig. 1.13: (a) Spacetime diagram depicting the twins paradox. Marked along the two world lines are intervals of proper time as measured by the two twins. (b) Spacetime diagram depicting the motions of the two mouths of a wormhole. Marked along the mouths' world tubes are intervals of proper time τ_c as measured by the single clock that sits on the common mouths.

1.8 [R] Time Travel

Time dilation is one facet of a more general phenomenon: Time, as measured by ideal clocks, is a "personal thing," different for different observers who move through spacetime on different world lines. This is well illustrated by the infamous "twins paradox," in which one twin, Methuselah, remains forever at rest in an inertial frame and the other, Florence, makes a spacecraft journey at high speed and then returns to rest beside Methuselah.

The twins' world lines are depicted in Fig. 1.13(a), a spacetime diagram whose axes are those of Methuselah's inertial frame. The time measured by an ideal clock that Methuselah carries is the coordinate time t of his inertial frame; and its total time lapse, from Florence's departure to her return, is $t_{\rm return} - t_{\rm departure} \equiv T_{\rm Methuselah}$. By contrast, the time measured by an ideal clock that Florence carries is the proper time τ, i.e. the square root of the invariant interval (1.7), along her world line; and thus her total time lapse from departure to return is

$$T_{\rm Florence} = \int d\tau = \int \sqrt{dt^2 - \delta_{ij} dx^i dx^j} = \int_0^{T_{\rm Methuselah}} \sqrt{1-v^2}\, dt \ . \qquad (1.52)$$

Here $(t,\ x^i)$ are the time and space coordinates of Methuselah's inertial frame, and v is Florence's ordinary speed, $v = \sqrt{\delta_{ij}(dx^i/dt)(dx^j/dt)}$, relative to Methuselah's frame. Obviously, Eq. (1.52) predicts that $T_{\rm Florence}$ is less than $T_{\rm Methuselah}$. In fact (cf. Exercise 1.16), even if Florence's acceleration is kept no larger than one Earth gravity throughout her trip, and her trip lasts only $T_{\rm Florence} =$ (a few tens of years), $T_{\rm Methuselah}$ can be hundreds or thousands or millions or billions of years.

Does this mean that Methuselah actually "experiences" a far longer time lapse, and actually ages far more than Florence? Yes. The time experienced by humans and the aging

of the human body are governed by chemical processes, which in turn are governed by the natural oscillation rates of molecules, rates that are constant to high accuracy when measured in terms of ideal time (or, equivalently, proper time τ). Therefore, a human's experiential time and aging time are the same as the human's proper time—so long as the human is not subjected to such high accelerations as to damage her body.

In effect, then, Florence's spacecraft has functioned as a time machine to carry her far into Methuselah's future, with only a modest lapse of her own proper time (ideal time; experiential time; aging time).

Is it also possible, at least in principle, for Florence to construct a time machine that carries her into Methuselah's past—and also her own past? At first sight, the answer would seem to be Yes. Figure 1.13(b) shows one possible method, using a *wormhole*. [Papers on other methods are cited in Thorne (1993) and Friedman and Higuchi (2006).]

Wormholes are hypothetical "handles" in the topology of space. A simple model of a wormhole can be obtained by taking a flat 3-dimensional space, removing from it the interiors of two identical spheres, and identifying the spheres' surfaces so that if one enters the surface of one of the spheres, one immediately finds oneself exiting through the surface of the other. When this is done, there is a bit of strongly localized spatial curvature at the spheres' common surface, so to analyze such a wormhole properly, one must use general relativity rather than special relativity. In particular, it is the laws of general relativity, combined with the laws of quantum field theory, that tell one how to construct such a wormhole and what kinds of materials (quantum fields) are required to "hold it open" so things can pass through it. Unfortunately, despite considerable effort, theoretical physicists have not yet deduced definitively whether those laws permit such wormholes to exist.[9] On the other hand, assuming such wormholes *can* exist, the following special relativistic analysis shows how one might be used to construct a machine for backward time travel.[10]

The two identified spherical surfaces are called the wormhole's mouths. Ask Methuselah to keep one mouth with himself, forever at rest in his inertial frame, and ask Florence to take the other mouth with herself on her high-speed journey. The two mouths' *world tubes* (analogs of world lines for a 3-dimensional object) then have the forms shown in Fig. 1.13(b). Suppose that a single ideal clock sits on the wormhole's identified mouths, so that from the external Universe one sees it both on Methuselah's wormhole mouth and on Florence's. As seen on Methuselah's mouth, the clock measures his proper time, which is equal to the coordinate time t [see tick marks along the left world tube in Fig. 1.13(b)]. As seen on Florence's mouth, the clock measures her proper time, Eq. (1.52) [see tick marks along the right world tube in Fig. 1.13(b)]. The result should be obvious, if surprising: When Florence returns to rest beside Methuselah, the wormhole has become a time machine. If she travels through the wormhole when the clock reads $\tau_c = 7$, she goes backward in time as seen in Methuselah's (or anyone else's) inertial frame; and then, in fact, traveling along the everywhere timelike, dashed world line, she is able to meet her younger self before she entered the wormhole.

This scenario is profoundly disturbing to most physicists because of the dangers of science-

[9] See, e.g., Morris and Thorne (1987), Thorne (1993), Borde, Ford and Roman (2002), and references therein.

[10] Morris, Thorne, and Yurtsever (1988).

fiction-type paradoxes (e.g., the older Florence might kill her younger self, thereby preventing herself from making the trip through the wormhole and killing herself). Fortunately perhaps, it now seems moderately likely (though not certain) that vacuum fluctuations of quantum fields will destroy the wormhole at the moment when its mouths' motion first makes backward time travel possible; and it may be that this mechanism will *always* prevent the construction of backward-travel time machines, no matter what tools one uses for their construction.[11]

****

EXERCISES

Exercise 1.16 *Example: [R] Twins Paradox*

(a) The 4-acceleration of a particle or other object is defined by $\vec{a} \equiv d\vec{u}/d\tau$, where $\vec{u}$ is its 4-velocity and τ is proper time along its world line. Show that, if an observer carries an accelerometer, the magnitude of the acceleration a measured by the accelerometer will always be equal to the magnitude of the observer's 4-acceleration, $a = |\vec{a}| \equiv \sqrt{\vec{a} \cdot \vec{a}}$.

(b) In the twins paradox of Fig. 1.13(a), suppose that Florence begins at rest beside Methuselah, then accelerates in Methuselah's x-direction with an acceleration a equal to one Earth gravity, "1g", for a time $T_{\rm Florence}/4$ as measured by her, then accelerates in the $-x$-direction at 1g for a time $T_{\rm Florence}/2$ thereby reversing her motion, and then accelerates in the $+x$-direction at 1g for a time $T_{\rm Florence}/4$ thereby returning to rest beside Methuselah. (This is the type of motion shown in the figure.) Show that the total time lapse as measured by Methuselah is

$$T_{\rm Methuselah} = \frac{4}{g} \sinh\left(\frac{gT_{\rm Florence}}{4}\right) . \tag{1.53}$$

(b) Show that in the geometrized units used here, Florence's acceleration (equal to acceletion of gravity at the surface of the Earth) is $g = 1.033/\text{yr}$. Plot $T_{\rm Methuselah}$ as a function of $T_{\rm Florence}$, and from your plot deduce that, if $T_{\rm Florence}$ is several tens of years, then $T_{\rm Methuselah}$ can be hundreds or thousands or millions or even billions of years.

Exercise 1.17 *Challenge: [R] Around the World on TWA*
In a long-ago era when an airline named Trans World Airlines (TWA) flew around the world, J. C. Hafele and R. E. Keating carried out a real live twins paradox experiment: They synchronized two atomic clocks, and then flew one around the world eastward on TWA, and on a separate trip, around the world westward, while the other clock remained at home at the Naval Research Laboratory near Washington D.C. When the clocks were compared after each trip, they were found to have aged differently. Making reasonable estimates for the airplane routing and speeds, compute the difference in aging, and compare your result

[11] Kim and Thorne (1991), Hawking (1992), Thorne (1993). But see also contrary indications in more recent research reviewed by Friedman and Higuchi (2006).

with the experimental data (Hafele and Keating, 1972). [Note: The rotation of the Earth is important, as is the general relativistic gravitational reshift associated with the clocks' altitudes; but the gravitational redshift drops out of the *difference* in aging, if the time spent at high altitude is the same eastward as westward.]

****

1.9 [N & R] Directional Derivatives, Gradients, Levi-Civita Tensor, Cross Product and Curl

[See note at the beginning of Sec. 1.5.3.]

Let us return to the formalism of differential geometry. We shall use the vector notation $\vec{A}$ of Minkowski spacetime, but our discussion will be valid simultaneously for spacetime and for Euclidean 3-space.

Consider a tensor field $\mathbf{T}(\mathcal{P})$ in spacetime or 3-space and a vector $\vec{A}$. We define the *directional derivative* of $\mathbf{T}$ along $\vec{A}$ by the obvious limiting procedure

$$\nabla_{\vec{A}}\mathbf{T} \equiv \lim_{\epsilon\to 0}\frac{1}{\epsilon}[\mathbf{T}(\vec{x}_{\mathcal{P}}+\epsilon\vec{A}) - \mathbf{T}(\vec{x}_{\mathcal{P}})] \tag{1.54a}$$

and similarly for the directional derivative of a vector field $\vec{B}(\mathcal{P})$ and a scalar field $\psi(\mathcal{P})$. In this definition we have denoted points, e.g. $\mathcal{P}$, by the vector $\vec{x}_{\mathcal{P}}$ that reaches from some arbitrary origin to the point.

It should not be hard to convince oneself that the directional derivative of any tensor field $\mathbf{T}$ is linear in the vector $\vec{A}$ along which one differentiates. Correspondingly, if $\mathbf{T}$ has rank n (n slots), then there is another tensor field, denoted $\nabla\mathbf{T}$, with rank $n+1$, such that

$$\nabla_{\vec{A}}\mathbf{T} = \nabla\mathbf{T}(\_, \_, \_, \vec{A}) \ . \tag{1.54b}$$

Here on the right side the first n slots (3 in the case shown) are left empty, and $\vec{A}$ is put into the last slot (the "differentiation slot"). The quantity $\nabla\mathbf{T}$ is called the *gradient* of $\mathbf{T}$. In slot-naming index notation, it is conventional to denote this gradient by $T_{\alpha\beta\gamma;\delta}$, where in general the number of indices preceding the semicolon is the rank of $\mathbf{T}$. Using this notation, the directional derivative of $\mathbf{T}$ along $\vec{A}$ reads [cf. Eq. (1.54b)] $T_{\alpha\beta\gamma;\delta}A^{\delta}$.

It is not hard to show that in any orthonormal (i.e., Cartesian or Lorentz) coordinate system, the components of the gradient are nothing but the partial derivatives of the components of the original tensor,

$$T_{\alpha\beta\gamma;\delta} = \frac{\partial T_{\alpha\beta\gamma}}{\partial x^{\delta}} \equiv T_{\alpha\beta\gamma,\delta} \ . \tag{1.54c}$$

(Here and henceforth all indices that follow a subscript comma represent partial derivatives, e.g., $S_{\alpha,\mu\nu} \equiv \partial^2 S_{\alpha}/\partial x^{\mu}\partial x^{\nu}$.) In a non-Cartesian and non-Lorentz basis, the components of the gradient typically are *not* obtained by simple partial differentiation [Eq. (1.54c) fails]

because of turning and length changes of the basis vectors as we go from one location to another. In Sec. 10.3 we shall learn how to deal with this by using objects called *connection coefficients*. Until then, however, we shall confine ourselves to Cartesian and Lorentz bases, so subscript semicolons and subscript commas can be used interchangeably.

Because the gradient and the directional derivatives are defined by the same standard limiting process as one uses when defining elementary derivatives, they obey the standard Leibniz rule for differentiating products:

$$\nabla_{\vec{A}}(\mathbf{S}\otimes\mathbf{T}) = (\nabla_{\vec{A}}\mathbf{S})\otimes\mathbf{T} + \mathbf{S}\otimes\nabla_{\vec{A}}\mathbf{T} ,$$
$$\text{i.e., } (S^{\alpha\beta}T^{\gamma\delta\epsilon})_{;\mu}A^\mu = (S^{\alpha\beta}{}_{;\mu}A^\mu)T^{\gamma\delta\epsilon} + S^{\alpha\beta}(T^{\gamma\delta\epsilon}{}_{;\mu}A^\mu) ; \tag{1.55a}$$

and

$$\nabla_{\vec{A}}(f\mathbf{T}) = (\nabla_{\vec{A}}f)\mathbf{T} + f\nabla_{\vec{A}}\mathbf{T} , \quad \text{i.e., } (fT^{\alpha\beta\gamma})_{;\mu}A^\mu = (f_{,\mu}A^\mu)T^{\alpha\beta\gamma} + fT^{\alpha\beta\gamma}{}_{;\mu}A^\mu . \tag{1.55b}$$

In an orthonormal basis these relations should be obvious: They follow from the Leibniz rule for partial derivatives.

Because the components $g_{\alpha\beta}$ of the metric tensor are constant in any Lorentz or Cartesian coordinate system, Eq. (1.54c) (which is valid in such coordinates) guarantees that $g_{\alpha\beta;\gamma} = 0$; i.e., the metric has vanishing gradient:

$$\nabla\mathbf{g} = 0 , \quad \text{i.e., } \; g_{\alpha\beta;\mu} = 0 . \tag{1.56}$$

From the gradient of any vector or tensor we can construct several other important derivatives by contracting on indices: (*i*) Since the gradient $\nabla\vec{A}$ of a vector field $\vec{A}$ has two slots, $\nabla\vec{A}(\_\,,\_\,)$, we can strangle (contract) its slots on each other to obtain a scalar field. That scalar field is the *divergence* of $\vec{A}$ and is denoted

$$\nabla\cdot\vec{A} \equiv (\text{contraction of } \nabla\vec{A}) = A^\alpha{}_{;\alpha} . \tag{1.57}$$

(*ii*) Similarly, if $\mathbf{T}$ is a tensor field of rank three, then $T^{\alpha\beta\gamma}{}_{;\gamma}$ is its divergence on its third slot, and $T^{\alpha\beta\gamma}{}_{;\beta}$ is its divergence on its second slot. (*iii*) By taking the double gradient and then contracting on the two gradient slots we obtain, from any tensor field $\mathbf{T}$, a new tensor field with the same rank,

$$\nabla^2\mathbf{T} \equiv (\nabla\cdot\nabla)\mathbf{T} , \quad \text{or, in index notation, } T_{\alpha\beta\gamma;\mu}{}^{;\mu} . \tag{1.58}$$

In Euclidean space ∇^2 is called the *Laplacian*; in spacetime it is called the *d'Alembertian*.

The metric tensor is a fundamental property of the space in which it lives; it embodies the inner product and thence the space's notion of distance or interval and thence the space's geometry. In addition to the metric, there is one (and only one) other fundamental tensor that embodies a piece of the space's geometry: the *Levi-Civita tensor* $\boldsymbol{\epsilon}$.

The Levi-Civita tensor has a number of slots equal to the dimensionality N of the space in which it lives, 4 slots in 4-dimensional spacetime and 3 slots in 3-dimensional Euclidean space; and $\boldsymbol{\epsilon}$ is antisymmetric in each and every pair of its slots. These properties determine $\boldsymbol{\epsilon}$ uniquely up to a multiplicative constant. That constant is fixed by a compatibility relation

between $\boldsymbol{\epsilon}$ and the metric $\mathbf{g}$: If $\{\vec{e}_\alpha\}$ is an orthonormal basis [orthonormality being defined with the aid of the metric, $\vec{e}_\alpha \cdot \vec{e}_\beta = \mathbf{g}(\vec{e}_\alpha, \vec{e}_\beta) = \eta_{\alpha\beta}$ in spacetime and $= \delta_{\alpha\beta}$ in Euclidean space], and if this basis is right-handed (a new property, not determined by the metric), then

$$\boldsymbol{\epsilon}(\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_N) = +1 \text{ in a space of } N \text{ dimensions}; \quad \boldsymbol{\epsilon}(\vec{e}_0, \vec{e}_1, \vec{e}_2, \vec{e}_3) = +1 \text{ in spacetime.} \tag{1.59a}$$

The concept of right handedness should be familiar in Euclidean 2-space or 3-space. In spacetime, the basis is right handed if $\{\vec{e}_1, \vec{e}_2, \vec{e}_3\}$ is right handed and $\vec{e}_0$ points to the future. Equation (1.59a) and the antisymmetry of $\boldsymbol{\epsilon}$ imply that in an orthonormal, right-handed basis, the only nonzero covariant components of $\boldsymbol{\epsilon}$ are

$$\begin{aligned} \epsilon_{12\ldots N} &= +1 \ , \\ \epsilon_{\alpha\beta\ldots\nu} &= +1 \text{ if } \alpha, \beta, \ldots, \nu \text{ is an even permutation of } 1, 2, \ldots, N \\ &= -1 \text{ if } \alpha, \beta, \ldots, \nu \text{ is an odd permutation of } 1, 2, \ldots, N \\ &= 0 \text{ if } \alpha, \beta, \ldots, \nu \text{ are not all different;} \end{aligned} \tag{1.59b}$$

(In spacetime the indices run from 0 to 3 rather than 1 to $N = 4$.) One can show that these components in one right-handed orthonormal frame imply these same components in all other right-handed orthonormal frames by virtue of the fact that the orthogonal (3-space) and Lorentz (spacetime) transformation matrices have unit determinant; and that in a left-handed orthormal frame the signs of these components are reversed.

In 3-dimensional Euclidean space, the Levi-Civita tensor is used to define the cross product and the curl:

$$\mathbf{A} \times \mathbf{B} \equiv \boldsymbol{\epsilon}(\_, \mathbf{A}, \mathbf{B}) \quad \text{i.e., in slot-naming index notation, } \epsilon_{ijk} A_j B_k \ ; \tag{1.60a}$$

$$\boldsymbol{\nabla} \times \mathbf{A} \equiv (\text{the vector field whose slot-naming index form is } \epsilon_{ijk} A_{k;j}) \ . \tag{1.60b}$$

[Equation (1.60b) is an example of an expression that is complicated if written in index-free notation; it says that $\boldsymbol{\nabla} \times \mathbf{A}$ is the double contraction of the rank-5 tensor $\boldsymbol{\epsilon} \otimes \nabla\mathbf{A}$ on its second and fifth slots, and on its third and fourth slots.]

Although Eqs. (1.60a) and (1.60b) look like complicated ways to deal with concepts that most readers regard as familiar and elementary, they have great power. The power comes from the following property of the Levi-Civita tensor in Euclidean 3-space [readily derivable from its components (1.59b)]:

$$\boxed{\epsilon_{ijm}\epsilon_{klm} = \delta^{ij}_{kl} \equiv \delta^i_k \delta^j_l - \delta^i_l \delta^j_k \ .} \tag{1.61}$$

Here δ^i_k is the Kronecker delta. Examine the 4-index delta function δ^{ij}_{kl} carefully; it says that either the indices above and below each other must be the same ($i = k$ and $j = l$) with a + sign, or the diagonally related indices must be the same ($i = l$ and $j = k$) with a − sign. [We have put the indices ij of δ^{ij}_{kl} up solely to facilitate remembering this rule. Recall (first paragraph of Sec. 1.5) that in Euclidean space and Cartesian coordinates, it does not matter whether indices are up or down.] With the aid of Eq. (1.61) and the index-notation expressions for the cross product and curl, one can quickly and easily derive a wide variety of useful vector identities; see the very important Exercise 1.18.

****

EXERCISES

Exercise 1.18 ***Example and Practice: [N] Vectorial Identities for the Cross Product and Curl*
Here is an example of how to use index notation to derive a vector identity for the double cross product $\mathbf{A}\times(\mathbf{B}\times\mathbf{C})$: In index notation this quantity is $\epsilon_{ijk}A_j(\epsilon_{klm}B_lC_m)$. By permuting the indices on the second ϵ and then invoking Eq. (1.61), we can write this as $\epsilon_{ijk}\epsilon_{lmk}A_jB_lC_m = \delta^{lm}_{ij}A_jB_lC_m$. By then invoking the meaning (1.61) of the 4-index delta function, we bring this into the form $A_jB_iC_j - A_jB_jC_i$, which is the index-notation form of $(\mathbf{A}\cdot\mathbf{C})\mathbf{B} - (\mathbf{A}\cdot\mathbf{B})\mathbf{C}$. Thus, it must be that $\mathbf{A}\times(\mathbf{B}\times\mathbf{C}) = (\mathbf{A}\cdot\mathbf{C})\mathbf{B} - (\mathbf{A}\cdot\mathbf{B})\mathbf{C}$.
Use similar techniques to evaluate the following quantities:

(a) $\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\mathbf{A})$

(b) $(\mathbf{A}\times\mathbf{B})\cdot(\mathbf{C}\times\mathbf{D})$

(c) $(\mathbf{A}\times\mathbf{B})\times(\mathbf{C}\times\mathbf{D})$

****

1.10 [R] Nature of Electric and Magnetic Fields; Maxwell's Equations

Now that we have introduced the gradient and the Levi-Civita tensor, we are prepared to study the relationship of the relativistic version of electrodynamics to the nonrelativistic ("Newtonian") version.

Consider a particle with charge q, rest mass m and 4-velocity $\vec{u}$ interacting with an electromagnetic field $\mathsf{F}(\_,\_)$. In index notation, the electromagnetic 4-force acting on the particle [Eq. (1.28)] is

$$dp^\alpha/d\tau = qF^{\alpha\beta}u_\beta \ . \tag{1.62}$$

Let us examine this 4-force in some arbitrary inertial reference frame in which particle's ordinary-velocity components are $v^j = v_j$ and its 4-velocity components are $u^0 = \gamma$, $u^j = \gamma v^j$ [Eqs. (1.34c)]. Anticipating the connection with the nonrelativistic viewpoint, we introduce the following notation for the contravariant components of the antisymmetric electromagnetic field tensor:

$$F^{0j} = -F^{j0} = E_j \ , \quad F^{ij} = \epsilon_{ijk}B_k \ . \tag{1.63}$$

(Recall that spatial indices, being Euclidean, can be placed up or down freely with no change in sign of the indexed quantity.) Inserting these components of F and $\vec{u}$ into Eq. (1.62) and using the relationship $dt/d\tau = u^0 = \gamma$ between t and τ derivatives, we obtain for the

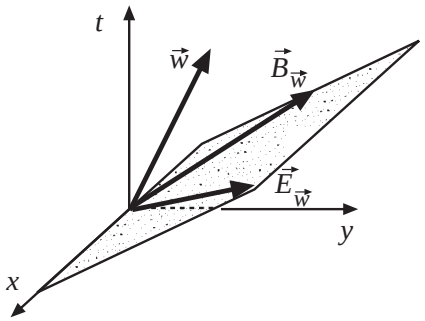


Fig. 1.14: The electric and magnetic fields measured by an observer with 4-velocity $\vec{w}$, shown as 4-vectors $\vec{E}_{\vec{w}}$ and $\vec{B}_{\vec{w}}$ that lie in the observer's 3-surface of simultaneity (stippled 3-surface orthogonal to $\vec{w}$).

components of the 4-force $dp_j/d\tau = \gamma dp_j/dt = q\gamma(E_j + \epsilon_{ijk}v_jB_k)$ and $dp^0/d\tau = \gamma dp^0/dt = \gamma E_j v_j$. Dividing by γ, converting into 3-space index notation, and denoting the particle's energy by $\mathcal{E} = p^0$, we bring these into the familiar Lorentz-force form

$$d\mathbf{p}/dt = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \ , \quad d\mathcal{E}/dt = \mathbf{v} \cdot \mathbf{E} \ . \tag{1.64}$$

Evidently **E** is the electric field and **B** the magnetic field as measured in our chosen Lorentz frame.

This may be familiar from standard electrodynamics textbooks, e.g. Jackson (1999). Not so familiar, but quite important, is the following geometric interpretation of **E** and **B**:

The electric and magnetic fields **E** and **B** are spatial vectors as measured in the chosen inertial frame. We can also regard them as 4-vectors that lie in the 3-surface of simultaneity $t =$ const of the chosen frame, i.e. that are orthogonal to the 4-velocity (denote it $\vec{w}$) of the frame's observers (cf. Figs. 1.12 and 1.14). We shall denote this 4-vector version of **E** and **B** by $\vec{E}_{\vec{w}}$ and $\vec{B}_{\vec{w}}$, where the subscript $\vec{w}$ identifies the 4-velocity of the observers who measure these fields. These fields are depicted in Fig. 1.14.

In the rest frame of the observer $\vec{w}$, the components of $\vec{E}_{\vec{w}}$ are $E^0_{\vec{w}} = 0$, $E^j_{\vec{w}} = E_j$ [the E_j appearing in Eqs. (1.63)], and similarly for $\vec{B}_{\vec{w}}$; and the components of $\vec{w}$ are $w^0 = 1$, $w^j = 0$. Therefore, in this frame Eqs. (1.63) can be rewritten as

$$\boxed{E^\alpha_{\vec{w}} = F^{\alpha\beta}w_\beta \ , \quad B^\beta_{\vec{w}} = \frac{1}{2}\epsilon^{\alpha\beta\gamma\delta}F_{\gamma\delta}w_\alpha \ .} \tag{1.65a}$$

(To verify this, insert the above components of F and $\vec{w}$ into these equations and, after some algebra, recover Eqs. (1.63) along with $E^0_{\vec{w}} = B^0_{\vec{w}} = 0$.) Equations (1.65a) say that in one special reference frame, that of the observer $\vec{w}$, the components of the 4-vectors on the left and on the right are equal. This implies that in every Lorentz frame the components of these 4-vectors will be equal; i.e., it implies that Eqs. (1.65a) are true when one regards them as geometric, frame-independent equations written in slot-naming index notation. *These equations enable one to compute the electric and magnetic fields measured by an observer (viewed as 4-vectors in the observer's 3-surface of simultaneity) from the observer's 4-velocity and the electromagnetic field tensor, without the aid of any basis or reference frame.*

Equations (1.65a) embody explicitly the following important fact: Although the electromagnetic field tensor $\mathbf{F}$ is a geometric, frame-independent quantity, the electric and magnetic fields $\vec{E}_{\vec{w}}$ and $\vec{B}_{\vec{w}}$ individually depend for their existence on a specific choice of observer (with 4-velocity $\vec{w}$), i.e., a specific choice of inertial reference frame, i.e., a specific choice of the split of spacetime into a 3-space (the 3-surface of simultaneity orthogonal to the observer's 4-velocity $\vec{w}$) and corresponding time (the Lorentz time of the observer's reference frame). *Only after making such an observer-dependent "3+1 split" of spacetime into space plus time do the electric field and the magnetic field come into existence as separate entities.* Different observers with different 4-velocities $\vec{w}$ make this spacetime split in different ways, thereby resolving the frame-independent $\mathbf{F}$ into different electric and magnetic fields $\vec{E}_{\vec{w}}$ and $\vec{B}_{\vec{w}}$.

By the same procedure as we used to derive Eqs. (1.65a), one can derive the inverse relationship, the following expression for the electromagnetic field tensor in terms of the (4-vector) electric and magnetic fields measured by some observer:

$$F^{\alpha\beta} = w^\alpha E^\beta_{\vec{w}} - E^\alpha_{\vec{w}} w^\beta + \epsilon^{\alpha\beta}{}_{\gamma\delta} w^\gamma B^\delta_{\vec{w}} \ . \tag{1.65b}$$

Maxwell's equations in geometric, frame-independent form are

$$F^{\alpha\beta}{}_{;\beta} = \begin{cases} 4\pi J^\alpha \text{ in Gaussian units} \\ J^\alpha/\epsilon_o = \mu_o J^\alpha \text{ in SI units} \ , \end{cases}$$
$$\epsilon^{\alpha\beta\gamma\delta} F_{\gamma\delta;\beta} = 0 \ . \tag{1.66}$$

(Since we are setting the speed of light to unity, $\epsilon_o = 1/\mu_o$.) Here $\vec{J}$ is the charge-current 4-vector, which in any inertial frame has components

$$J^0 = \rho_e = (\text{charge density}) \ , \quad J^i = j_i = (\text{current density}). \tag{1.67}$$

Exercise 1.20 describes how to think about this charge density and current density as geometric objects determined by the observer's 4-velocity or 3+1 split of spacetime into space plus time. Exercise 1.21 shows how the frame-independent Maxwell equations (1.66) reduce to the more familiar ones in terms of $\mathbf{E}$ and $\mathbf{B}$. Exercise 1.22 explores potentials for the electromagnetic field in geometric, frame-independent language and the 3+1 split.

****

EXERCISES

Exercise 1.19 *Derivation and Practice: [R] Reconstruction of* $\mathbf{F}$
Derive Eq. (1.65b) by the same method as was used to derive (1.65a).

Exercise 1.20 *Problem: [R] 3+1 Split of Charge-Current 4-Vector*
Just as the electric and magnetic fields measured by some observer can be regarded as 4-vectors $\vec{E}_{\vec{w}}$ and $\vec{B}_{\vec{w}}$ that live in the observer's 3-space of simultaneity, so also the charge density and current density that the observer measures can be regarded as a scalar $\rho_{\vec{w}}$ and 4-vector $\vec{j}_{\vec{w}}$ that live in the 3-space of simultaneity. Derive geometric, frame-independent equations for $\rho_{\vec{w}}$ and $\vec{j}_{\vec{w}}$ in terms of the charge-current 4-vector $\vec{J}$ and the observer's 4-velocity $\vec{w}$, and derive a geometric expression for $\vec{J}$ in terms of $\rho_{\vec{w}}$, $\vec{j}_{\vec{w}}$, and $\vec{w}$.

Exercise 1.21 *Problem: [R] Frame-Dependent Version of Maxwell's Equations*
From the geometric version of Maxwell's equations (1.66), derive the elementary, frame-dependent version

$$\boldsymbol{\nabla}\cdot\mathbf{E} = \begin{cases} 4\pi\rho_e \text{ in Gaussian units} \\ \rho_e/\epsilon_o \text{ in SI units,} \end{cases} \qquad \boldsymbol{\nabla}\times\mathbf{B} - \frac{\partial\mathbf{E}}{\partial t} = \begin{cases} 4\pi\mathbf{j} \text{ in Gaussian units} \\ \mu_o\mathbf{j} \text{ in SI units,} \end{cases}$$
$$\boldsymbol{\nabla}\cdot\mathbf{B} = 0\ , \qquad \boldsymbol{\nabla}\times\mathbf{E} + \frac{\partial\mathbf{B}}{\partial t} = 0\ . \tag{1.68}$$

Exercise 1.22 *Problem: [R] Potentials for the Electromagnetic Field*

(a) Express the electromagnetic field tensor as an antisymmetrized gradient of a 4-vector potential: in slot-naming index notation

$$F_{\alpha\beta} = A_{\beta;\alpha} - A_{\alpha;\beta}\ . \tag{1.69a}$$

Show that, whatever may be the 4-vector potential $\vec{A}$, the second of the Maxwell equations (1.66) is automatically satisfied. Show further that the electromagnetic field tensor is unaffected by a gauge change of the form

$$\vec{A}_{\rm new} = \vec{A}_{\rm old} + \nabla\psi\ , \tag{1.69b}$$

where ψ is a scalar field (the generator of the gauge change). Show, finally, that it is possible to find a gauge-change generator that enforces "Lorenz gauge"

$$\nabla\cdot\vec{A} = 0 \tag{1.69c}$$

on the new 4-vector potential, and show that in this gauge, the first of the Maxwell equations (1.66) becomes

$$\nabla^2\vec{A} = 4\pi\vec{J}\,; \qquad \text{i.e. } A^{\alpha;\mu}{}_{\mu} = 4\pi J^{\alpha}\ . \tag{1.69d}$$

(b) Introduce an inertial reference frame, and in that frame split $\mathbf{F}$ into the electric and magnetic fields $\mathbf{E}$ and $\mathbf{B}$, split $\vec{J}$ into the charge and current densities ρ_e and $\mathbf{j}$, and split the vector potential into a scalar potential and a 3-vector potential

$$\phi \equiv A_0\ , \quad \mathbf{A} = \text{spatial part of } \vec{A}\ . \tag{1.69e}$$

Deduce the 3+1 splits of Eqs. (1.69a)–(1.69d) and show that they take the form given in standard textbooks on electrodynamics.

****

1.11 Volumes, Integration, and Integral Conservation Laws

1.11.1 [N] Newtonian Volumes and Integration

The Levi-Civita tensor is the foundation for computing volumes and performing volume integrals in any number of dimensions. In Cartesian coordinates of 2-dimensional Euclidean space, the area (i.e. 2-dimensional volume) of a parallelogram whose sides are $\mathbf{A}$ and $\mathbf{B}$ is

$$\text{2-Volume} = \epsilon_{ab}A_aB_b = A_1B_2 - A_2B_1 = \det\begin{bmatrix} A_1 & B_1 \\ A_2 & B_2 \end{bmatrix} , \tag{1.70a}$$

a relation that should be familiar from elementary geometry. Equally familiar should be the expression for the 3-dimensional volume of a parallelopiped with legs $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$:

$$\text{3-Volume} = \epsilon_{ijk}A_iB_jC_k = \mathbf{A}\cdot(\mathbf{B}\times\mathbf{C}) = \det\begin{bmatrix} A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \\ A_3 & B_3 & C_3 \end{bmatrix} . \tag{1.70b}$$

Recall that this volume has a sign: it is positive if $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ is a right handed set of vectors and negative if left-handed. Equations (1.70a) and (1.70b) are foundations from which one can derive the usual formulae $dA = dx\,dy$ and $dV = dx\,dy\,dz$ for the area and volume of elementary surface and volume elements with Cartesian side lengths dx, dy and dz.

In Euclidean 3-space, we define the vectorial surface area of a 2-dimensional parallelogram with legs $\mathbf{A}$ and $\mathbf{B}$ to be

$$\mathbf{\Sigma} = \mathbf{A}\times\mathbf{B} = \boldsymbol{\epsilon}(\_, \mathbf{A}, \mathbf{B}) . \tag{1.70c}$$

This vectorial surface area has a magnitude equal to the area of the parallelogram and a direction perpendicular to it. Such vectorial surface areas are the foundation for surface integrals in 3-dimensional space, and for the familiar Gauss theorem

$$\int_{\mathcal{V}_3} (\boldsymbol{\nabla}\cdot\mathbf{A})dV = \int_{\partial\mathcal{V}_3} \mathbf{A}\cdot d\mathbf{\Sigma} \tag{1.71a}$$

(where $\mathcal{V}_3$ is a compact 3-dimensional region and $\partial\mathcal{V}_3$ is its closed two-dimensional boundary) and Stokes theorem

$$\int_{\mathcal{V}_2} \boldsymbol{\nabla}\times\mathbf{A}\cdot d\mathbf{\Sigma} = \int_{\partial\mathcal{V}_2} \mathbf{A}\cdot d\mathbf{l} \tag{1.71b}$$

(where $\mathcal{V}_2$ is a compact 2-dimensional region, $\partial\mathcal{V}_2$ is the 1-dimensional closed curve that bounds it, and the last integral is a line integral around that curve).

Notice that in Euclidean 3-space, the vectorial surface area $\boldsymbol{\epsilon}(\_, \mathbf{A}, \mathbf{B})$ can be thought of as an object that is waiting for us to insert a third leg $\mathbf{C}$ so as to compute a volume $\boldsymbol{\epsilon}(\mathbf{C}, \mathbf{A}, \mathbf{B})$—the volume of the parallelopiped with legs $\mathbf{C}$, $\mathbf{A}$, and $\mathbf{B}$.

This mathematics is illustrated by the integral and differential conservation laws for electric charge and for particles: The total charge and the total number of particles inside a three dimensional region of space $\mathcal{V}_3$ are $\int_{\mathcal{V}_3}\rho_e dV$ and $\int_{\mathcal{V}_3} n dV$, where ρ_e is the charge density

and n the number density of particles. The rates that charge and particles flow out of $\mathcal{V}_3$ are the integrals of the current density $\mathbf{j}$ and the particle flux vector $\mathbf{S}$ over its boundary $\partial\mathcal{V}_3$. Therefore, the laws of charge conservation and particle conservation say

$$\frac{d}{dt}\int_{\mathcal{V}_3}\rho_e dV + \int_{\partial\mathcal{V}_3}\mathbf{j}\cdot d\mathbf{\Sigma} = 0\ , \qquad \frac{d}{dt}\int_{\mathcal{V}_3} n dV + \int_{\partial\mathcal{V}_3}\mathbf{S}\cdot d\mathbf{\Sigma} = 0\ . \tag{1.72}$$

Pull the time derivative inside each volume integral (where it becomes a partial derivative), and apply Gauss's law to each surface integral; the results are $\int_{\mathcal{V}_3}(\partial\rho_e/\partial t + \boldsymbol{\nabla}\cdot\mathbf{j})dV = 0$ and similarly for particles. The only way these equations can be true for all choices of $\mathcal{V}_3$ is by the integrands vanishing:

$$\partial\rho_e/\partial t + \boldsymbol{\nabla}\cdot\mathbf{j} = 0\ , \quad \partial n/\partial t + \boldsymbol{\nabla}\cdot\mathbf{S} = 0\ . \tag{1.73}$$

These are the differential conservation laws for charge and for particles. They have a standard, universal form: the time derivative of the density of a quantity plus the divergence of its flux vanishes.

1.11.2 [R] Spacetime Volumes and Integration

The generalization to 4-dimensional spacetime should be obvious: The 4-dimensional parallelopiped whose legs are the four vectors $\vec{A}, \vec{B}, \vec{C}, \vec{D}$ has a 4-dimensional volume given by the analog of Eqs. (1.70a) and (1.70b):

$$\text{4-Volume} = \epsilon_{\alpha\beta\gamma\delta}A^\alpha B^\beta C^\gamma D^\delta = \boldsymbol{\epsilon}(\vec{A},\vec{B},\vec{C},\vec{D}) = \det\begin{bmatrix} A^0 & B^0 & C^0 & D^0 \\ A^1 & B^1 & C^1 & D^1 \\ A^2 & B^2 & C^2 & D^2 \\ A^3 & B^3 & C^3 & D^3 \end{bmatrix}\ . \tag{1.74}$$

Note that this 4-volume is positive if the set of vectors $\{\vec{A},\ \vec{B},\ \vec{C},\ \vec{D}\}$ is right-handed and negative if left-handed.

Just as Eqs. (1.70a) and (1.70b) give us a way to perform area and volume integrals in 2- and 3-dimensional Euclidean space, so Equation (1.74) provides us a way to perform volume integrals over 4-dimensional Minkowski spacetime: To integrate a tensor field $\mathbf{T}$ over some region $\mathcal{V}$ of spacetime, we need only divide spacetime up into tiny parallelopipeds, multiply the 4-volume $d\Sigma$ of each parallelopiped by the value of $\mathbf{T}$ at its center, and add. It is not hard to see from Eq. (1.74) that in any right-handed Lorentz coordinate system, the 4-volume of a tiny parallelopiped whose edges are dx^α along the four orthogonal coordinate axes is $d\Sigma = dt\,dx\,dy\,dz$ (the analog of $dV = dx\,dy\,dz$), and correspondingly the integral of $\mathbf{T}$ over $\mathcal{V}$ can be expressed as

$$\int_{\mathcal{V}} T^{\alpha\beta\gamma}d\Sigma = \int_{\mathcal{V}} T^{\alpha\beta\gamma}dt\,dx\,dy\,dz\ . \tag{1.75}$$

By analogy with the vectorial area (1.70c) of a parallelogram in 3-space, any 3-dimensional parallelopiped in spacetime with legs $\vec{A}, \vec{B}, \vec{C}$ has a vectorial 3-volume $\vec{\Sigma}$ (not to be confused with the scalar 4-volume Σ) defined by

$$\vec{\Sigma}(\_\,) = \boldsymbol{\epsilon}(\_\,,\ \vec{A},\ \vec{B},\ \vec{C})\ ; \quad \Sigma_\mu = \epsilon_{\mu\alpha\beta\gamma}A^\alpha B^\beta C^\gamma\ . \tag{1.76}$$

Here we have written the volume vector both in abstract notation and in component notation. This volume vector has one empty slot, ready and waiting for a fourth vector ("leg") to be inserted, so as to compute the 4-volume Σ of a 4-dimensional parallelopiped.

Notice that the volume vector $\vec{\Sigma}$ is orthogonal to each of its three legs (because of the antisymmetry of $\boldsymbol{\epsilon}$), and thus (unless it is null) it can be written as $\vec{\Sigma} = V\vec{n}$ where V is the magnitude of the volume and $\vec{n}$ is the unit normal to the three legs.

Interchanging any two legs of the parallelopiped reverses the 3-volume's sign. Consequently, the 3-volume is characterized not only by its legs but also by the order of its legs, or equally well, in two other ways: (i) by the direction of the vector $\vec{\Sigma}$ (reverse the order of the legs, and the direction of $\vec{\Sigma}$ will reverse); and (ii) by the *sense* of the 3-volume, defined as follows. Just as a 2-volume (i.e., a segment of a plane) in 3-dimensional space has two sides, so a 3-volume in 4-dimensional spacetime has two sides; cf. Fig. 1.15. Every vector $\vec{D}$ for which $\vec{\Sigma} \cdot \vec{D} > 0$ points out of one side of the 3-volume $\vec{\Sigma}$. We shall call that side the "positive side" of $\vec{\Sigma}$; and we shall call the other side, the one out of which point vectors $\vec{D}$ with $\vec{\Sigma} \cdot \vec{D} < 0$, its "negative side". When something moves through or reaches through or points through the 3-volume from its negative side to its positive side, we say that this thing is moving or reaching or pointing in the "positive sense"; and similarly for "negative sense". The examples shown in Fig. 1.15 should make this more clear.

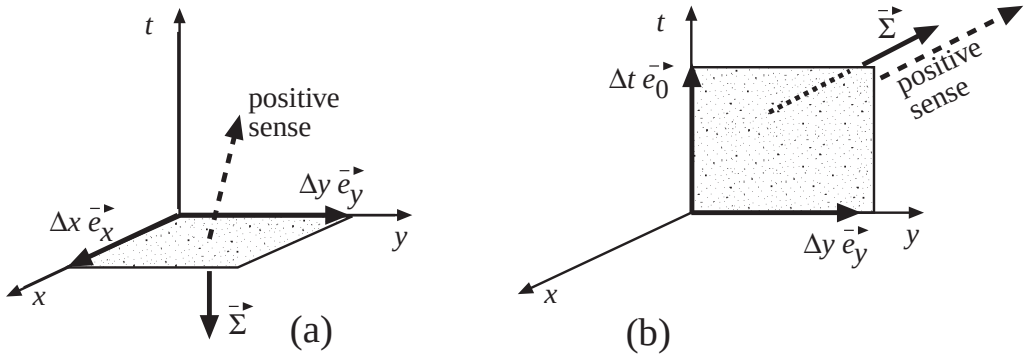


Fig. 1.15: Spacetime diagrams depicting 3-volumes in 4-dimensional spacetime, with one spatial dimension (that along the z-direction) suppressed.

Figure 1.15(a) shows two of the three legs of the volume vector $\vec{\Sigma} = \boldsymbol{\epsilon}(\_, \Delta x\vec{e}_x, \Delta y\vec{e}_y, \Delta z\vec{e}_z)$, where x, y, z are the spatial coordinates of a specific Lorentz frame. It is easy to show that this vector can also be written as $\vec{\Sigma} = -\Delta V\vec{e}_0$, where ΔV is the ordinary volume of the parallelopiped as measured by an observer in the chosen Lorentz frame, $\Delta V = \Delta x \Delta y \Delta z$. Thus, the direction of the vector $\vec{\Sigma}$ is toward the past (direction of decreasing Lorentz time t). From this, and the fact that timelike vectors have negative squared length, it is easy to infer that $\vec{\Sigma} \cdot \vec{D} > 0$ if and only if the vector $\vec{D}$ points out of the "future" side of the 3-volume (the side of increasing Lorentz time t); therefore, the positive side of $\vec{\Sigma}$ is the future side. This means that the vector $\vec{\Sigma}$ points in the negative sense of its own 3-volume.

Figure 1.15(b) shows two of the three legs of the volume vector $\vec{\Sigma} = \boldsymbol{\epsilon}(\_, \Delta t\vec{e}_t, \Delta y\vec{e}_y, \Delta z\vec{e}_z) = -\Delta t \Delta A \vec{e}_x$ (with $\Delta A = \Delta y \Delta z$). In this case, $\vec{\Sigma}$ points in its own positive sense.

This peculiar behavior is completely general: When the normal to a 3-volume is timelike, its volume vector $\vec{\Sigma}$ points in the negative sense; when the normal is spacelike, $\vec{\Sigma}$ points in the positive sense; and—it turns out—when the normal is null, $\vec{\Sigma}$ lies in the 3-volume (parallel to its one null leg) and thus points neither in the positive sense nor the negative.[12]

Note the physical interpretations of the 3-volumes of Fig. 1.15: That in Fig. 1.15(a) is an instantaneous snapshot of an ordinary, spatial, parallelopiped, while that in Fig. 1.15(b) is the 3-dimensional region in spacetime swept out during time Δt by the parallelogram with legs $\Delta y \vec{e}_y$, $\Delta z \vec{e}_z$ and with area $\Delta A = \Delta y \Delta z$.

Just as in 3-dimensional Euclidean space, vectorial surface areas can be used to construct 2-dimensional surface integrals, so also (and in identically the same manner) in 4-dimensional spacetime, vectorial volume elements can be used to construct integrals over 3-dimensional volumes (also called 3-dimensional surfaces), e.g. $\int_{\mathcal{V}_3} \vec{A} \cdot d\vec{\Sigma}$. More specifically: Let (a, b, c) be (possibly curvilinear) coordinates in the 3-surface $\mathcal{V}_3$, and denote by $\vec{x}(a, b, c)$ the spacetime point $\mathcal{P}$ on $\mathcal{V}_3$ whose coordinate values are (a, b, c). Then $(\partial\vec{x}/\partial a)da$, $(\partial\vec{x}/\partial b)db$, $(\partial\vec{x}/\partial c)dc$ are the vectorial legs of the elementary parallelopiped whose corners are at (a, b, c), $(a+da, b, c)$, $(a, b+db, c)$, etc; and the spacetime components of these vectorial legs are $(\partial x^\alpha/\partial a)da$, $(\partial x^\alpha/\partial b)db$, $(\partial x^\alpha/\partial c)dc$. The 3-volume of this elementary parallelopiped is $d\vec{\Sigma} = \boldsymbol{\epsilon}\big(\_, (\partial\vec{x}/\partial a)da, (\partial\vec{x}/\partial b)db, (\partial\vec{x}/\partial c)dc\big)$, which has spacetime components

$$d\Sigma_\mu = \epsilon_{\mu\alpha\beta\gamma} \frac{\partial x^\alpha}{\partial a} \frac{\partial x^\beta}{\partial b} \frac{\partial x^\gamma}{\partial c} da db dc \, . \tag{1.77}$$

This is the integration element to be used when evaluating

$$\int_{\mathcal{V}_3} \vec{A} \cdot d\vec{\Sigma} = \int_{\partial\mathcal{V}_3} A^\mu d\Sigma_\mu \, . \tag{1.78}$$

Just as there are Gauss and Stokes theorems (1.71a) and (1.71b) for integrals in Euclidean 3-space, so also there are Gauss and Stokes theorems in spacetime. The Gauss theorem has the obvious form

$$\boxed{\int_{\mathcal{V}_4} (\nabla \cdot \vec{A}) d\Sigma = \int_{\partial\mathcal{V}_4} \vec{A} \cdot d\vec{\Sigma} \, ,} \tag{1.79}$$

where the first integral is over a 4-dimensional region $\mathcal{V}_4$ in spacetime, and the second is over the 3-dimensional boundary of $\mathcal{V}_4$, with the boundary's positive sense pointing outward, away from $\mathcal{V}_4$ (just as in the 3-dimensional case). We shall not write down the 4-dimensional Stokes theorem because it is complicated to formulate with the tools we have developed thus far; easy formulation requires the concept of a *differential form,* which we shall not introduce in this book.

1.11.3 [R] Conservation of Charge in Spacetime

We can use integration over a 3-dimensional region (3-surface) in 4-dimensional spacetime to construct an elegant, frame-independent formulation of the law of conservation of electric charge:

[12] This peculiar behavior gets replaced by a simpler description if one uses one-forms rather than vectors to describe 3-volumes; see, e.g., Box 5.2 of Misner, Thorne, and Wheeler (1973).

We begin by examining the geometric meaning of the charge-current 4-vector $\vec{J}$. We defined $\vec{J}$ in Eq. (1.67) in terms of its components. The spatial component $J^x = J_x = J(\vec{e}_x)$ is equal to the x component of current density j_x; i.e. it is the amount Q of charge that flows across a unit surface area lying in the y-z plane, in a unit time; i.e., the charge that flows across the unit 3-surface $\vec{\Sigma} = \vec{e}_x$. In other words, $\vec{J}(\vec{\Sigma}) = \vec{J}(\vec{e}_x)$ *is the total charge* Q *that flows across* $\vec{\Sigma} = \vec{e}_x$ *in* $\vec{\Sigma}$*'s positive sense*; and similarly for the other spatial directions. The temporal component $J^0 = -J_0 = \vec{J}(-\vec{e}_0)$ is the charge density ρ_e; i.e., it is the total charge Q in a unit spatial volume. This charge is carried by particles that are traveling through spacetime from past to future, and pass through the unit 3-surface (3-volume) $\vec{\Sigma} = -\vec{e}_0$. Therefore, $\vec{J}(\vec{\Sigma}) = \vec{J}(-\vec{e}_0)$ *is the total charge* Q *that flows through* $\vec{\Sigma} = -\vec{e}_0$ *in its positive sense.* This is precisely the same interpretation as we deduced for the spatial components of $\vec{J}$.

This makes it plausible, and indeed one can show, that *for any small* 3*-surface* $\vec{\Sigma}$, $\vec{J}(\vec{\Sigma}) \equiv J^\alpha \Sigma_\alpha$ *is the total charge* Q *that flows across* $\vec{\Sigma}$ *in its positive sense.*

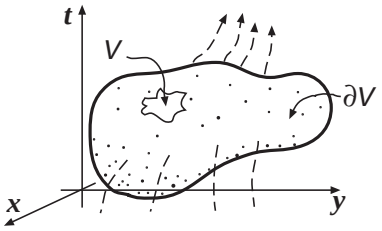


Fig. 1.16: The 4-dimensional region $\mathcal{V}$ in spacetime, and its closed 3-boundary $\partial\mathcal{V}$, used in formulating the law of charge conservation. The dashed lines symbolize, heuristically, the flow of charge from past toward future.

This property of the charge-current 4-vector is the foundation for our frame-independent formulation of the law of charge conservation. Let $\mathcal{V}$ be a compact, 4-dimensional region of spacetime and denote by $\partial\mathcal{V}$ its boundary, a closed 3-surface in 4-dimensional spacetime (Fig. 1.16). The charged media (fluids, solids, particles, ...) present in spacetime carry electric charge through $\mathcal{V}$, from the past toward the future. The law of charge conservation says that all the charge that enters $\mathcal{V}$ through the past part of its boundary $\partial\mathcal{V}$ must exit through the future part of its boundary. If we choose the positive sense of the boundary's infinitesimal 3-volume $d\vec{\Sigma}$ to point out of $\mathcal{V}$ (toward the past on the bottom boundary and toward the future on the top), then this *global law of charge conservation* can be expressed mathematically as

$$\boxed{\int_{\partial\mathcal{V}} J^\alpha d\Sigma_\alpha = 0 \ .} \tag{1.80}$$

When each tiny charge q enters $\mathcal{V}$ through its past boundary, it contributes negatively to the integral, since it travels through $\partial\mathcal{V}$ in the negative sense (from positive side of $\partial\mathcal{V}$ toward negative side); and when that same charge exits $\mathcal{V}$ through its future boundary, it contributes positively to the integral. Therefore its net contribution is zero, and similarly for all other charges.

In exercise 1.24, we show when this *global law of charge conservation* (1.80) is subjected to a 3+1 split of spacetime into space plus time, it becomes the nonrelativistic integral law of charge conservation (1.72).

This global conservation law can be converted into a *local conservation law* with the help of the 4-dimensional Gauss theorem (1.79), $\int_{\partial\mathcal{V}} J^\alpha d\Sigma_\alpha = \int_{\mathcal{V}} J^\alpha{}_{;\alpha} d\Sigma$. Since the left-hand side vanishes, so must the right-hand side; and in order for this 4-volume integral to vanish for every choice of $\mathcal{V}$, it is necessary that the integrand vanish everywhere in spacetime:

$$\boxed{J^\alpha{}_{;\alpha} = 0 \ ; \quad \text{i.e.} \vec{\boldsymbol{\nabla}} \cdot \vec{J} = 0 \ .} \tag{1.81}$$

In a specific but arbitrary Lorentz frame (i.e., in a 3+1 split of spacetime into space plus time), this becomes the standard differential law of charge conservation (1.73).

1.11.4 [R] Conservation of Particles, Baryons and Rest Mass

Any conserved scalar quantity obeys conservation laws of the same form as those for electric charge. For example, if the number of particles of some species (e.g. electrons or protons or photons) is conserved, then we can introduce for that species a *number-flux 4-vector* $\vec{S}$ (analog of charge-current 4-vector $\vec{J}$): In any Lorentz coordinate system S^0 is the number density of particles n and S^j is the particle flux. If $\vec{\Sigma}$ is a small 3-volume (3-surface) in spacetime, then $\vec{S}(\vec{\Sigma}) = SW^\alpha \Sigma_\alpha$ is the number of particles that pass through Σ from its negative side to its positive side. The frame-invariant global and local conservation laws for these particles take the same form as those for electric charge:

$$\int_{\partial\mathcal{V}} S^\alpha d\Sigma_\alpha = 0, \quad \text{where } \partial\mathcal{V} \text{ is any closed 3-surface in spacetime,} \tag{1.82a}$$

$$S^\alpha{}_{;\alpha} = 0 \ ; \quad i.e. \ \nabla \cdot \vec{S} = 0 \ . \tag{1.82b}$$

When fundamental particles (e.g. protons and antiprotons) are created and destroyed by quantum processes, the total baryon number (number of baryons minus number of antibaryons) is still conserved—or, at least this is so to the accuracy of all experiments performed thus far. We shall assume it so in this book. This law of baryon-number conservation takes the forms (1.82a) and (1.82b), with $\vec{S}$ the number-flux 4-vector for baryons (with antibaryons counted negatively).

It is useful to reexpress this baryon-number conservation law in Newtonian-like language by introducing a universally agreed upon mean rest mass per baryon $\bar{m}_B$ This $\bar{m}_B$ is often taken to be 1/56 the mass of an ^{56}Fe (iron-56) atomic nucleus, since ^{56}Fe is the nucleus with the tightest nuclear binding, i.e. the endpoint of thermonuclear evolution in stars. We multiply the baryon number-flux 4-vector $\vec{S}$ by this mean rest mass per baryon to obtain a rest-mass-flux 4-vector

$$\vec{S}_{\rm rm} = \bar{m}_B \vec{S} \ , \tag{1.83}$$

which (since $\bar{m}_B$ is, by definition, a constant) satisfies the same conservation laws (1.82a) and (1.82b) as baryon number.

For media such as fluids and solids, in which the particles travel only short distances between collisions or strong interactions, it is often useful to resolve the particle number-flux 4-vector and the rest-mass-flux 4-vector into a 4-velocity of the medium $\vec{u}$ (i.e., the 4-velocity of the frame in which there is a vanishing net spatial flux of particles), and the particle number density n_o or rest mass density ρ_o as measured in the medium's rest frame:

$$\vec{S} = n_o \vec{u} \ , \quad \vec{S}_{\rm rm} = \rho_o \vec{u} \ . \tag{1.84}$$

See Exercise 1.25.

We shall make use of the conservation laws $\nabla \cdot \vec{S} = 0$ and $\nabla \cdot \vec{S}_{\rm rm} = 0$ for particles and rest mass later in this book, e.g. when studying relativistic fluids; and we shall find the expressions (1.84) for the number-flux 4-vector and rest-mass-flux 4-vector quite useful. See, e.g., the discussion of relativistic shock waves in Ex. 16.11, and the nonrelativistic limit of a relativistic fluid in Sec. 23.4 .

****

EXERCISES

Exercise 1.23 *Practice and Example: [R] Evaluation of 3-Surface Integral in Spacetime*
In Minkowski spacetime the set of all events separated from the origin by a timelike interval a^2 is a 3-surface, the hyperboloid $t^2 - x^2 - y^2 - z^2 = a^2$, where $\{t, x, y, z\}$ are Lorentz coordinates of some inertial reference frame. On this hyperboloid introduce coordinates $\{\chi, \theta, \phi\}$ such that

$$t = a \cosh\chi \ , \quad x = a \sinh\chi \sin\theta \cos\phi \ , \quad y = a \sinh\chi \sin\theta \sin\phi; \ , \quad z = a \sinh\chi \cos\theta \ . \tag{1.85}$$

Note that χ is a radial coordinate and (θ, ϕ) are spherical polar coordinates. Denote by $\mathcal{V}_3$ the portion of the hyperboloid with $\chi \leq b$.

(a) Verify that for all values of (χ, θ, ϕ), the points (1.85) do lie on the hyperboloid.

(b) On a spacetime diagram, draw a picture of $\mathcal{V}_3$, the $\{\chi, \theta, \phi\}$ coordinates, and the elementary volume element (vector field) $d\vec{\Sigma}$ [Eq. (1.77)].

(c) Set $\vec{A} \equiv \vec{e}_0$ (the temporal basis vector), and express $\int_{\mathcal{V}_3} \vec{A} \cdot d\vec{\Sigma}$ as an integral over $\{\chi, \theta, \phi\}$. Evaluate the integral.

(d) Consider a closed 3-surface consisting of the segment $\mathcal{V}_3$ of the hyperboloid as its top, the hypercylinder $\{x^2 + y^2 + z^2 = a^2 \sinh^2 b, \ 0 < t < a \cosh b\}$ as its sides, and the sphere $\{x^2 + y^2 + z^2 \leq a^2 \sinh^2 b \ , \ t = 0\}$ as its bottom. Draw a picture of this closed 3-surface on a spacetime diagram. Use Gauss's theorem, applied to this 3-surface, to show that $\int_{\mathcal{V}_3} \vec{A} \cdot d\vec{\Sigma}$ is equal to the 3-volume of its spherical base.

Exercise 1.24 *Derivation and Example: [R] Global Law of Charge Conservation in an Inertial Frame*
Consider the global law of charge conservation $\int_{\partial V} J^\alpha d\Sigma_\alpha = 0$ for a special choice of the closed 3-surface $\partial\mathcal{V}$: The bottom of $\partial\mathcal{V}$ is the ball $\{t = 0, x^2 + y^2 + z^2 \le a^2\}$, where $\{t, x, y, z\}$ are the Lorentz coordinates of some inertial frame. The sides are the spherical world tube $\{0 \le t \le T, x^2 + y^2 + z^2 = a^2\}$. The top is the ball $\{t = T, x^2 + y^2 + z^2 \le a^2\}$.

(a) Draw this 3-surface in a spacetime diagram.

(b) Show that for this $\partial\mathcal{V}$, $\int_{\partial V} J^\alpha d\Sigma_\alpha = 0$ is the nonrelativistic integral conservation law (1.72) for charge.

Exercise 1.25 *Example: [R] Rest-mass-flux 4-vector, Lorentz contraction of rest-mass density, and rest-mass conservation for a fluid*
Consider a fluid with 4-velocity $\vec{u}$, and rest-mass density ρ_o as measured in the fluid's rest frame.

(a) From the physical meanings of $\vec{u}$, ρ_o, and the rest-mass-flux 4-vector $\vec{S}_{\rm rm}$, deduce Eq. (1.84).

(b) Examine the components of $\vec{S}_{\rm rm}$ in a reference frame where the fluid moves with ordinary velocity $\mathbf{v}$. Show that $S^0 = \rho_o\gamma$, $S^j = \rho_o\gamma v^j$, where $\gamma = 1/\sqrt{1 - \mathbf{v}^2}$. Explain the physical interpretation of these formulas in terms of Lorentz contraction.

(c) Show that the law of conservation of rest-mass $\nabla \cdot \vec{S}_{\rm rm} = 0$, takes the form

$$\frac{d\rho_o}{d\tau} = -\rho_o \nabla \cdot \vec{u} \,, \tag{1.86}$$

where $d/d\tau$ is derivative with respect to proper time moving with the fluid.

(d) Consider a small 3-dimensional volume V of the fluid, whose walls move with the fluid (so if the fluid expands, V goes up). Explain why the law of rest-mass conservation must take the form $d(\rho_o V)/d\tau = 0$. Thereby deduce that

$$\nabla \cdot \vec{u} = (1/V)(dV/d\tau) \,. \tag{1.87}$$

****

1.12 The Newtonian Stress Tensor, Relativistic Stress-energy Tensor, and Conservation of 4-Momentum

1.12.1 [N] Newtonian Stress Tensor and Momentum Conservation

Press your hands together in the x–y plane and feel the force that one hand exerts on the other across a tiny area A — say, one square millimeter of your hands' palms (Fig. 1.17). That force, of course, is a vector $\mathbf{F}$. It has a normal component (along the x direction). It also has a tangential component: if you try to slide your hands past each other, you feel a component of force along their surface, a "shear" force in the y and z directions. Not only is the force $\mathbf{F}$ vectorial; so is the 2-surface across which it acts, $\mathbf{\Sigma} = A\,\mathbf{e}_x$. (Here $\mathbf{e}_x$ is the unit vector orthogonal to the tiny area A, and we have chosen the negative side of the surface to be the $-x$ side and the positive side to be $+x$. With this choice, the force $\mathbf{F}$ is that which the negative hand, on the $-x$ side, exerts on the positive hand.)

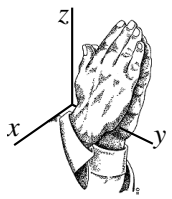


Fig. 1.17: Hands, pressed together, exert a stress on each other.

Now, it should be obvious that the force $\mathbf{F}$ is a linear function of our chosen surface $\mathbf{\Sigma}$. Therefore, there must be a tensor, the *stress tensor*, that reports the force to us when we insert the surface into its second slot:

$$\mathbf{F}(\_\,) = \mathbf{T}(\_\,, \mathbf{\Sigma})\ , \quad \text{i.e., } F_i = T_{ij}\Sigma_j\ . \tag{1.88}$$

Newton's law of action and reaction tells us that the force that the positive hand exerts on the negative hand must be equal and opposite to that that which the negative hand exerts on the positive. This shows up trivially in Eq. (1.88): By changing the sign of $\mathbf{\Sigma}$, one reverses which hand is regarded as negative and which positive; and since $\mathbf{T}$ is linear in $\mathbf{\Sigma}$, one also reverses the sign of the force.

The definition (1.88) of the stress tensor gives rise to the the following physical meaning of its components:

$$\begin{aligned} T_{jk} &= \left(\begin{array}{c} j\text{-component of force per unit area} \\ \text{across a surface perpendicular to } \vec{e}_k \end{array} \right) \\ &= \left(\begin{array}{c} j\text{-component of momentum that crosses a unit} \\ \text{area which is perpendicular to } \vec{e}_k\text{, per unit time,} \\ \text{with the crossing being from } -x^k \text{ to } +x^k \end{array} \right) . \end{aligned} \tag{1.89}$$

The stresses inside a table with a heavy weight on it are described by the stress tensor $\mathbf{T}$, as are the stresses in a flowing fluid or plasma, in the electromagnetic field, and in any other

physical medium. Accordingly, we shall use the stress tensor as an important mathematical tool in our study of force balance in kinetic theory (Chap. 2), elasticity theory (Part III), fluid mechanics (Part IV), and plasma physics (Part V).

It is not obvious from its definition, but the stress tensor $\mathbf{T}$ is always symmetric in its two slots. To see this, consider a small cube with side L in any medium (or field) (Fig. 1.18). The medium outside the cube exerts forces, and thence also torques, on the cube's faces. The z-component of the torque is produced by the shear forces on the front and back faces and on the left and right. As shown in the figure, the shear forces on the front and back faces have magnitudes $T_{xy}L^2$ and point in opposite directions, so they exert identical torques on the cube, $N_z = T_{xy}L^2(L/2)$ (where $L/2$ is the distance of each face from the cube's center). Similarly, the shear forces on the left and right faces have magnitudes $T_{yx}L^2$ and point in opposite directions, thereby exerting identical torques on the cube, $N_z = -T_{yx}L^2(L/2)$. Adding the torques from all four faces and equating them to the rate of change of angular momentum, $\frac{1}{12}\rho L^5 d\Omega_z/dt$ (where ρ is the mass density, $\frac{1}{12}\rho L^5$ is the cube's moment of inertia, and Ω_z is the z component of its angular velocity), we obtain $(T_{yx} - T_{xy})L^3 = \frac{1}{12}\rho L^5 d\Omega_z/dt$. Now, let the cube's edge length become arbitrarily small, $L \to 0$. If $T_{yx} - T_{xy}$ does not vanish, then the cube will be set into rotation with an infinitely large angular acceleration, $d\Omega_z/dt \propto 1/L^2 \to \infty$ — an obviously unphysical behavior. Therefore $T_{yx} = T_{xy}$, and similarly for all other components; *the stress tensor is always symmetric under interchange of its two slots.*

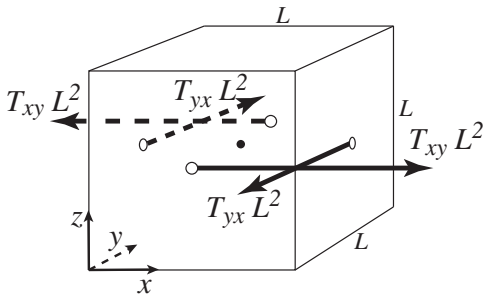


Fig. 1.18: The shear forces exerted on the left, right, front and back faces of a vanishingly small cube. The resulting torque about the z direction will set the cube into rotation with an arbitrarily large angular acceleration unless the stress tensor is symmetric.

Two examples will make the stress tensor more concrete:

Perfect fluid: Inside a perfect fluid there is an isotropic pressure P, so $T_{xx} = T_{yy} = T_{zz} = P$ (the normal forces per unit area across surfaces in the yz plane, the zx plane and the xy plane are all equal to P). The fluid cannot support any shear forces, so the off-diagonal components of $\mathbf{T}$ vanish. We can summarize this by $T_{ij} = P\delta_{ij}$ or equivalently, since δ_{ij} are the components of the Euclidean metric, $T_{ij} = Pg_{ij}$. The frame-independent version of this is $\mathbf{T} = P\mathbf{g}$. To check this result, consider a 2-surface $\mathbf{\Sigma} = A\mathbf{n}$ with area A oriented perpendicular to some arbitrary unit vector $\mathbf{n}$. The vectorial force that the fluid exerts across $\mathbf{\Sigma}$ is, in index notation, $F_j = T_{jk}\Sigma_k = Pg_{jk}An_k = PAn_j$; i.e. it is a normal force with magnitude equal to the fluid pressure P times the surface area A. This is what it should be.

Electromagnetic field: See Ex. 1.26 below.

The stress tensor plays a central role in the *Newtonian law of momentum conservation*: Recall the physical intepretation of T_{jk} as the j-component of momentum that crosses a unit area perpendicular to $\mathbf{e}_k$ per unit time [Eq. (1.89)]. Apply this definition to the little cube in Fig. 1.18. The momentum that flows *into* the cube in unit time across the front face (at $y = 0$) is $T_{jy}L^2$, and across the back face (at $y = L$) is $-T_{jy}L^2$; and their sum is $-T_{jy,y}L^3$. Adding to this the contributions from the side faces and the top and bottom faces, we find for the rate of change of total momentum inside the cube $(-T_{jx,x} - T_{jy,y} - T_{jz,z})L^3 = -T_{jk,k}L^3$. Since the cube's volume is L^3, this says that

$$\partial(\text{momentum density})/dt + \boldsymbol{\nabla}\cdot\mathbf{T} = 0\ . \tag{1.90}$$

This has the standard form for any local conservation law: the time derivative of the density of some quantity (here momentum), plus the divergence of the flux of that quantity (here momentum flux is the stress tensor), is zero. We shall make extensive use of this Newtonian *local law of momentum conservation* in Part III (elasticity theory), Part IV (fluid mechanics) and Part V (plasma physics).

****

EXERCISES

Exercise 1.26 ***Problem: [R] Electromagnetic Stress Tensor*

An electric field $\mathbf{E}$ exerts (in Gaussian cgs units) a pressure $\mathbf{E}^2/8\pi$ orthogonal to itself and a tension of this same magnitude along itself. Similarly, a magnetic field $\mathbf{B}$ exerts a pressure $\mathbf{B}^2/8\pi$ orthogonal to itself and a tension of this same magnitude along itself. Verify that the following stress tensor embodies these stresses:

$$\mathbf{T} = \frac{1}{8\pi}\left[(\mathbf{E}^2 + \mathbf{B}^2)\mathbf{g} - 2(\mathbf{E}\otimes\mathbf{E} + \mathbf{B}\otimes\mathbf{B})\right]\ . \tag{1.91}$$

****

1.12.2 [R] Relativistic Stress-Energy Tensor

We conclude this chapter by formulating the law of 4-momentum conservation in ways analogous to our laws of conservation of charge, particles, baryons and rest mass. This task is not trivial, since 4-momentum is a vector in spacetime, while charge, particle number, baryon number, and rest mass are scalar quantities. Correspondingly, the density-flux of 4-momentum must have one more slot than the density-fluxes of charge, baryon number and rest mass, $\vec{J}$, $\vec{S}$ and $\vec{S}_{\rm rm}$; it must be a second-rank tensor. We call it the *stress-energy tensor* and denote it $\mathbf{T}(\_\,,\_\,)$ (the same notation as we use for the stress-tensor in Euclidean space).

Consider a medium or field flowing through 4-dimensional spacetime. As it crosses a tiny 3-surface $\vec{\Sigma}$, it transports a net electric charge $\vec{J}(\vec{\Sigma})$ from the negative side of $\vec{\Sigma}$ to

the positive side, and net baryon number $\vec{S}(\vec{\Sigma})$ and net rest mass $\vec{S}_{\rm rm}(\vec{\Sigma})$; and similarly, it transports a net 4-momentum $\mathbf{T}(\_\,,\vec{\Sigma})$ from the negative side to the positive side:

$$\mathbf{T}(\_\,,\vec{\Sigma}) \equiv (\text{total 4-momentum } \vec{P} \text{ that flows through } \vec{\Sigma}); \quad \text{i.e., } T^{\alpha\beta}\Sigma_\beta = P^\alpha \ . \tag{1.92}$$

From this definition of the stress-energy tensor we can read off the physical meanings of its components on a specific, but arbitrary, Lorentz-coordinate basis: Making use of method (1.32b) for computing the components of a vector or tensor, we see that in a specific, but arbitrary, Lorentz frame (where $\vec{\Sigma} = -\vec{e}_0$ is a volume vector representing a parallelopiped with unit volume $\Delta V = 1$, at rest in that frame, with its positive sense toward the future):

$$\begin{aligned} -T_{\alpha 0} = \mathbf{T}(\vec{e}_\alpha, -\vec{e}_0) = \vec{P}(\vec{e}_\alpha) &= \left(\begin{array}{c} \alpha\text{-component of 4-momentum that} \\ \text{flows from past to future across a unit} \\ \text{volume } \Delta V = 1 \text{ in the 3-space } t = \text{const} \end{array} \right) \\ &= (\alpha\text{-component of density of 4-momentum }) \ . \end{aligned} \tag{1.93a}$$

Specializing α to be a time or space component and raising indices, we obtain the specialized versions of (1.93a)

$$\begin{aligned} T^{00} &= (\text{energy density as measured in the chosen Lorentz frame}), \\ T^{j0} &= (\text{density of } j\text{-component of momentum in that frame}). \end{aligned} \tag{1.93b}$$

Similarly, the αx component of the stress-energy tensor (also called the $\alpha 1$ component since $x = x^1$ and $\vec{e}_x = \vec{e}_1$) has the meaning

$$\begin{aligned} T_{\alpha 1} \equiv T_{\alpha x} \equiv \mathbf{T}(\vec{e}_\alpha, \vec{e}_x) &= \left(\begin{array}{c} \alpha\text{-component of 4-momentum that crosses} \\ \text{a unit area } \Delta y \Delta z = 1 \text{ lying in a surface of} \\ \text{constant } x, \text{ during unit time } \Delta t, \text{ crossing} \\ \text{from the } -x \text{ side toward the } +x \text{ side} \end{array} \right) \\ &= \left(\begin{array}{c} \alpha \text{ component of flux of 4-momentum} \\ \text{across a surface lying perpendicular to } \vec{e}_x \end{array} \right) \ . \end{aligned} \tag{1.93c}$$

The specific forms of this for temporal and spatial α are (after raising indices)

$$T^{0x} = \left(\begin{array}{c} \text{energy flux across a surface perpendidular to } \vec{e}_x, \\ \text{from the } -x \text{ side to the } +x \text{ side} \end{array} \right) \ , \tag{1.93d}$$

$$T^{jx} = \left(\begin{array}{c} \text{flux of } j\text{-component of momentum across a surface} \\ \text{perpendicular to } \vec{e}_x, \text{ from the } -x \text{ side to the } +x \text{ side} \end{array} \right) = \left(\begin{array}{c} jx \text{ component} \\ \text{of stress} \end{array} \right) \ . \tag{1.93e}$$

The αy and αz components have the obvious, analogous interpretations.

These interpretations, restated much more briefly, are:

$$\boxed{T^{00} = (\text{energy density}),\ T^{j0} = (\text{momentum density}),\ T^{0j} = (\text{energy flux}),\ T^{jk} = (\text{stress}).} \tag{1.93f}$$

Although it might not be obvious at first sight, *the 4-dimensional stress-energy tensor is always symmetric*: in index notation (where indices can be thought of as representing the names of slots, or equally well components on an arbitrary basis)

$$T^{\alpha\beta} = T^{\beta\alpha} \ . \tag{1.94}$$

This symmetry can be deduced by physical arguments in a specific, but arbitrary, Lorentz frame: Consider, first, the $x0$ and $0x$ components, i.e., the x-components of momentum density and energy flux. A little thought, symbolized by the following heuristic equation, reveals that they must be equal

$$T^{x0} = \begin{pmatrix} \text{momentum} \\ \text{density} \end{pmatrix} = \frac{(\Delta\mathcal{E})dx/dt}{\Delta x\Delta y\Delta z} = \frac{\Delta\mathcal{E}}{\Delta y\Delta z\Delta t} = \begin{pmatrix} \text{energy} \\ \text{flux} \end{pmatrix} , \tag{1.95}$$

and similarly for the other space-time and time-space components: $T^{j0} = T^{0j}$. [In Eq. (1.95), in the first expression $\Delta\mathcal{E}$ is the total energy (or equivalently mass) in the volume $\Delta x\Delta y\Delta z$, $(\Delta\mathcal{E})dx/dt$ is the total momentum, and when divided by the volume we get the momentum density. The third equality is just elementary algebra, and the resulting expression is obviously the energy flux.] The space-space components, being equal to the stress tensor, are also symmetric, $T^{jk} = T^{kj}$, by the argument embodied in Fig. 1.18 above. Since $T^{0j} = T^{j0}$ and $T^{jk} = T^{kj}$, all components in our chosen Lorentz frame are symmetric, $T^{\alpha\beta} = T^{\beta\alpha}$. This means that, if we insert arbitrary vectors into the slots of $\mathbf{T}$ and evaluate the resulting number in our chosen Lorentz frame, we will find

$$\mathbf{T}(\vec{A},\, \vec{B}) = T^{\alpha\beta}A_\alpha B_\beta = T^{\beta\alpha}A_\alpha B_\beta = \mathbf{T}(\vec{B},\, \vec{A}) \ ; \tag{1.96}$$

i.e., $\mathbf{T}$ is symmetric under interchange of its slots.

Let us return to the physical meanings (1.93) of the components of the stress-energy tensor. With the aid of $\mathbf{T}$'s symmetry, we can restate those meanings in the language of a 3+1 split of spacetime into space plus time: *When one chooses a specific reference frame, that choice splits the stress-energy tensor up into three parts. Its time-time part is the energy density T^{00}, Its time-space part $T^{0j} = T^{j0}$ is the energy flux or equivalently the momentum density, and its space-space part T^{jk} is the symmetric stress tensor.*

1.12.3 [R] 4-Momentum Conservation

Our interpretation of $\vec{J}(\vec{\Sigma}) \equiv J^\alpha\Sigma_\alpha$ as the net charge that flows through a small 3-surface $\vec{\Sigma}$ from its negative side to its positive side gave rise to the global conservation law for charge, $\int_{\partial\mathcal{V}} J^\alpha d\Sigma_\alpha = 0$ [Eqs. (1.80) and Fig. 1.16]. Similarly the role of $\mathbf{T}(\_\,, \vec{\Sigma})$ [$T^{\alpha\beta}\Sigma_\beta$ in slot naming index notation] as the net 4-momentum that flows through $\vec{\Sigma}$ from its negative side to positive gives rise to the following equation for conservation of 4-momentum:

$$\boxed{\int_{\partial\mathcal{V}} T^{\alpha\beta}d\Sigma_\beta = 0 \ .} \tag{1.97}$$

This equation says that all the 4-momentum that flows into the 4-volume $\mathcal{V}$ of Fig. 1.16 through its 3-surface $\partial\mathcal{V}$ must also leave $\mathcal{V}$ through $\partial\mathcal{V}$; it gets counted negatively when it

enters (since it is traveling from the positive side of $\partial\mathcal{V}$ to the negative), and it gets counted positively when it leaves, so its net contribution to the integral (1.97) is zero.

This *global law of 4-momentum conservation* can be converted into a *local law* (analogous to $\vec{\boldsymbol{\nabla}} \cdot \vec{J} = 0$ for charge) with the help of the 4-dimensional Gauss's theorem (1.79). Gauss's theorem, generalized in the obvious way from a vectorial integrand to a tensorial one, says:

$$\int_{\mathcal{V}} T^{\alpha\beta}{}_{;\beta}\, d\Sigma \;= \int_{\partial\mathcal{V}} T^{\alpha\beta} d\Sigma_\beta \;. \tag{1.98}$$

Since the right-hand side vanishes, so must the left-hand side; and in order for this 4-volume integral to vanish for every choice of $\mathcal{V}$, it is necessary that the integrand vanish everywhere in spacetime:

$$\boxed{T^{\alpha\beta}{}_{;\beta} = 0 \;; \quad \text{i.e., } \vec{\boldsymbol{\nabla}} \cdot \mathbf{T} = 0 \;.} \tag{1.99a}$$

In the second, index-free version of this local conservation law, the ambiguity about which slot the divergence is taken on is unimportant, since $\mathbf{T}$ is symmetric in its two slots: $T^{\alpha\beta}{}_{;\beta} = T^{\beta\alpha}{}_{;\beta}$.

In a specific but arbitrary Lorentz frame, the local conservation law (1.99a) for 4-momentum has as its temporal and spatial parts

$$\frac{\partial T^{00}}{\partial t} + \frac{\partial T^{0k}}{\partial x^k} = 0 \;, \tag{1.99b}$$

i.e., the time derivative of the energy density plus the 3-divergence of the energy flux vanishes; and

$$\frac{\partial T^{j0}}{\partial t} + \frac{\partial T^{jk}}{\partial x^k} = 0 \;, \tag{1.99c}$$

i.e., the time derivative of the momentum density plus the 3-divergence of the stress (i.e., of the momentum flux) vanishes. Thus, as one should expect, the geometric, frame-independent law of 4-momentum conservation includes as special cases both the conservation of energy and the conservation of momentum; and their differential conservation laws have the standard form that one expects both in Newtonian physics and in special relativity: time derivative of density plus divergence of flux vanishes; cf. Eq. (1.90) and associated discussion.

1.12.4 [R] Stress-Energy Tensors for Perfect Fluid and Electromagnetic Field

As an important example that illustrates the stress-energy tensor, consider a *perfect fluid*. A perfect fluid is a continuous medium whose stress-energy tensor, evaluated in its *local rest frame* (a Lorentz frame where $T^{j0} = T^{0j} = 0$), has the special form

$$T^{00} = \rho \;, \quad T^{jk} = P\delta^{jk} \;. \tag{1.100a}$$

Here ρ is a short-hand notation for the energy density (density of total mass-energy, including rest mass) T^{00}, as measured in the local rest frame; and the stress tensor T^{jk} as measured in that frame has the form of an isotropic pressure P, and vanishing shear stress. From this

special form of $T^{\alpha\beta}$ in the local rest frame, one can derive the following expression for the stress-energy tensor in terms of the 4-velocity $\vec{u}$ of the local rest frame, i.e., of the fluid itself, the metric tensor of spacetime $\mathbf{g}$, and the rest-frame energy density ρ and pressure P:

$$\boxed{T^{\alpha\beta} = (\rho + P)u^\alpha u^\beta + P g^{\alpha\beta} \ ; \quad \text{i.e., } \mathbf{T} = (\rho + P)\vec{u} \otimes \vec{u} + P\mathbf{g} \ .} \tag{1.100b}$$

See Ex. 1.28, below. In Part IV of this book, we shall explore in depth the implications of this stress-energy tensor, in the Newtonian limit.

Another example of a stress-energy tensor is that for the electromagnetic field, which takes the following form in Gaussian units:

$$T^{\alpha\beta} = \frac{1}{4\pi}\left(F^{\alpha\mu}F^{\beta}{}_{\mu} - \frac{1}{4}g^{\alpha\beta}F^{\mu\nu}F_{\mu\nu}\right) \tag{1.101}$$

see Exercise 1.30

****

EXERCISES

Exercise 1.27 *Example: [R] Global Conservation of 4-Momentum in an Inertial Frame*
Consider the 4-dimensional parallelopiped $\mathcal{V}$ whose legs are $\Delta t\vec{e}_t$, $\Delta x\vec{e}_x$, $\Delta y\vec{e}_y$ $\Delta z\vec{e}_z$, where $(t, x, y, z) = (x^0, x^1, x^2, x^3)$ are the coordinates of some inertial frame. The boundary $\partial\mathcal{V}$ of this $\mathcal{V}$ has eight 3-dimensional "faces". Identify these faces, and write the integral $\int_{\partial\mathcal{V}} T^{0\beta}d\Sigma_\beta$ as the sum of contributions from each of them. According to the law of energy conservation, this sum must vanish. Explain the physical interpretation of each of the eight contributions to this energy conservation law. (Cf. Ex. 1.24 for an analogous interpretation of charge conservation.)

Exercise 1.28 ***Derivation and Example: [R] Stress-Energy Tensor and Energy-Momentum Conservation for a Perfect Fluid*

(a) Derive the frame-independent expression (1.100b) for the perfect fluid stress-energy tensor from its rest-frame components (1.100a).

(b) Explain why the projection of $\nabla\cdot\mathbf{T} = 0$ along the fluid 4-velocity, $\vec{u}\cdot(\nabla\cdot\mathbf{T}) = 0$, should represent energy conservation as viewed by the fluid itself. Show that this equation reduces to

$$\frac{d\rho}{d\tau} = -(\rho + P)\nabla\cdot\vec{u} \ . \tag{1.102a}$$

With the aid of Eq. (1.87), bring this into the form

$$\frac{d(\rho V)}{d\tau} = -P\frac{dV}{d\tau} \ , \tag{1.102b}$$

where V is the 3-volume of some small fluid element as measured in the fluid's local rest frame. What are the physical interpretations of the left and right sides of this equation, and how is it related to the first law of thermodynamics?

(c) Read the discussion, in Ex. 1.10, of the tensor $\mathbf{P} = \mathbf{g} + \vec{u} \otimes \vec{u}$ that projects into the 3-space of the fluid's rest frame. Explain why $P_{\mu\alpha}T^{\alpha\beta}{}_{;\beta} = 0$ should represent the law of force balance (momentum conservation) as seen by the fluid. Show that this equation reduces to

$$(\rho + P)\vec{a} = -\mathbf{P} \cdot \nabla P \ , \tag{1.102c}$$

where $\vec{a} = d\vec{u}/d\tau$ is the fluid's 4-acceleration. This equation is a relativistic version of Newton's "$\mathbf{F} = m\mathbf{a}$". Explain the physical meanings of the left and right hand sides. Infer that $\rho + P$ must be the fluid's inertial mass per unit volume. See Ex. 1.29 for further justification of this inference.

Exercise 1.29 ***Example: [R] Inertial Mass Per Unit Volume*
Suppose that some medium has a rest frame (unprimed frame) in which its energy flux and momentum density vanish, $T^{0j} = T^{j0} = 0$. Suppose that the medium moves in the x direction with speed very small compared to light, $v \ll 1$, as seen in a (primed) laboratory frame, and ignore factors of order v^2. The "ratio" of the medium's momentum density $T^{j'0'}$ as measured in the laboratory frame to its velocity $v_i = v\delta_{ix}$ is called its total *inertial mass per unit volume*, and is denoted $\rho^{\rm inert}_{ji}$:

$$T^{j'0'} = \rho^{\rm inert}_{ji} v_i \ . \tag{1.103}$$

(a) Show, using a Lorentz transformation from the medium's (unprimed) rest frame to the (primed) laboratory frame, that

$$\rho^{\rm inert}_{ji} = T^{00}\delta_{ji} \ + \ T_{ji} \ . \tag{1.104}$$

(b) Give a physical explanation of the contribution $T_{ji}v_i$ to the momentum density.

(c) Show that for a perfect fluid [Eq. (1.100b)] the inertial mass per unit volume is isotropic and has magnitude $\rho + P$, where ρ is the mass-energy density and P is the pressure measured in the fluid's rest frame:

$$\boxed{\rho^{\rm inert}_{ji} = (\rho + P)\delta_{ji} \ .} \tag{1.105}$$

See Ex. 1.28 above for this inertial-mass role of $\rho + P$ in the law of force balance (momentum conservation) for a fluid.

Exercise 1.30 ***Example: [R] Stress-Energy Tensor, and Energy-Momentum Conservation for the Electromagnetic Field*

(a) Compute from Eq. (1.101) the components of the electromagnetic stress-energy tensor in an inertial reference frame in Gaussian units. Your answer should be the expressions given in electrodynamic textbooks:

$$\begin{aligned} T^{00} &= \frac{\mathbf{E}^2 + \mathbf{B}^2}{8\pi} \ , \quad T^{0j}\mathbf{e}_j = T^{j0}\mathbf{e}_j = \frac{\mathbf{E} \times \mathbf{B}}{4\pi} \ , \\ T^{jk} &= \frac{1}{8\pi}\left[(\mathbf{E}^2 + \mathbf{B}^2)\delta_{jk} - 2(E_jE_k + B_jB_k)\right] \ . \end{aligned} \tag{1.106}$$

See also Ex. 1.26 above for an alternative derivation of the stress tensor T_{jk}.

(b) Show that for the electromagnetic field,

$$T^{\alpha\beta}{}_{;\beta} = F^{\alpha\mu} J_\mu \ , \tag{1.107a}$$

where J_μ is the charge-current 4-vector.

(c) The matter that carries the electric charge and current can exchange energy and momentum with the electromagnetic field. Explain why Eq. (1.107a) is the rate per unit volume at which that matter feeds 4-momentum into the electromagnetic field, and conversely, $-F^{\alpha\mu} J_\mu$ is the rate per unit volume at which the electromagnetic field feeds 4-momentum into the matter. Show, further, that (as viewed in any reference frame) the time and space components of this quantity are

$$\frac{d\mathcal{E}_{\rm matter}}{dtdV} = -F^{0j} J_j = \mathbf{E}\cdot\mathbf{j} \ , \quad \frac{d\mathbf{p}_{\rm matter}}{dtdV} = \rho_e \mathbf{E} + \mathbf{j}\times\mathbf{B} \ , \tag{1.107b}$$

where ρ_e is charge density and $\mathbf{j}$ is current density [Eq. (1.67)]. The first of these equations is ohmic heating of the matter by the electric field; the second is the Lorentz force per unit volume on the matter.

****

Bibliographic Note

For an inspiring taste of the history of special relativity, see the original papers by Einstein, Lorentz, and Minkowski, translated into English and archived in Einstein et. al. (1923).

Early relativity textbooks [see the bibliography on p. 567 of Jackson (1999)] emphasized the transformation properties of physical quantities, in going from one inertial frame to another, rather than their roles as frame-invariant geometric objects. Minkowski (1908) introduced geometric thinking, but only in recent years — in large measure due to the influence of John Wheeler — has the geometric viewpoint gained ascendancy. It is still not common in texts on Newtonian physics, but it is almost universal in modern relativity texts.

In our opinion, the best elementary introduction to special relativity is the first edition of Taylor and Wheeler (1966); the more ponderous second edition (1992) is also good. Both adopt the geometric viewpoint. At an intermediate level, most physics students learn relativity from electrodynamics texts such as Griffiths (1999) and Jackson (1999), or classical mechanics texts such as Goldstein (1980). Avoid the first and second editions of Jackson and of Goldstein, which use imaginary time and obscure the geometry of spacetime! Griffiths and Jackson (like old relativity texts) adopt the "transformation" viewpoint on physical quantities, rather than the geometric viewpoint. Under John Safko's influence, the third edition of Goldstein [Goldstein, Poole and Safko (2002)] has become strongly geometric.

For fully geometric treatments of special relativity, analogous to ours, see not only the third edition of Goldstein, but also the special relativity sections in modern general relativity texts. Some we like at the undergraduate level are Schutz (1985) and especially Hartle (2002);

Box 1.4
Important Concepts in Chapter 1

- **Foundational Concepts**
  - Frameworks for physical laws (general relativity, special relativity and Newtonian physics) and their relationships to each other, Sec. 1.1.
  - Inertial reference frame, Sec. 1.2.2.
  - Invariant interval and how it defines the geometry of spacetime, Sec. 1.2.3.
- **Principle of Relativity:** Laws of physics are frame-independent geometric relations between geometric objects, Sec. 1.2.3. Important examples:
  - Newton's second law of motion $\mathbf{F} = m\mathbf{a}$, Eq. (1.13).
  - Lorentz force law in 3-dimensional Newtonian language (1.14), and in 4-dimensional geometric language (1.28) and their connection, Sec. 1.10.
  - Conservation of 4-momentum in particle interations, Eq. (1.23).
  - Global and local conservation laws for charge, baryon number, and 4-momentum, Secs. 1.11.3, 1.11.4, 1.12.3.
- **Differential geometry**
  - Tensor as a linear function of vectors, Sec. 1.3. Examples: Electromagnetic field tensor (1.26), stress tensor (1.88) and stress-energy tensor (1.92).
  - Slot-naming index notation, Sec. 1.5.3.
  - Gauss's theorem in Euclidean space (1.71a), and in spacetime (1.79).
  - Computations via geometric techniques, without coordinates or Lorentz transformations (e.g. derive Lorentz force law Ex. 1.4.3, derive Dopper shift Ex. 1.11, derive vector identities Ex. 1.18).
- **3+1 Split of spacetime into space plus time** induced by choice of inertial frame, Sec. 1.6, and resulting 3+1 split of physical quantities and laws:
  - 4-momentum $\rightarrow$ energy and momentum, Eqs. (1.35), (1.36), (1.38); Ex. 1.9.
  - Electromagnetic tensor $\rightarrow$ electric field and magnetic field, Sec. 1.10.
  - Charge-current 4-vector $\rightarrow$ charge density and current density, Ex. 1.20.
- **Spacetime diagrams** used to understand Lorentz contraction, time dilation, breakdown of simultaneity (Sec. 1.7.3, Ex. 1.14) and conservation laws (Fig. 1.16).

and at a more advanced level, Carroll (2004) and the venerable Misner, Thorne and Wheeler (1973) — often cited as MTW. In Parts I–V of our book, we minimize, so far as possible, the proliferation of mathematical concepts (avoiding, e.g., differential forms and dual bases). By contrast, other advanced treatments (e.g. MTW, Goldstein 3rd edition, and Carroll) embrace the richer mathematics.

Bibliography

Borde, Arvind, Ford, L.H., and Roman, Thomas A., 2002. "Constraints on spatial distributions of negative energy," *Physical Review D*, **65**, 084002.

Carroll, S. 2004. *An Introduction to Spacetime and Geometry*, New York: Addison Wesley.

Einstein, Albert, Lorentz, Hendrik A., Minkowski, Hermann, and Weyl, Hermann 1923. *The Principle of Relativity*, New York: Dover Publications.

Feynman, Richard P. 1966. *The Character of Physical Law*, M.I.T. Press, Cambridge, Massachusetts.

Friedman, John L. and Higuchi, A. "Topological censorship and chronology protection," *Annalen der Physik*, **15**, 109–128 (2006).

Goldstein, Herbert 1980. *Classical Mechanics*, New York: Addison Wesley, second edition.

Goldstein, Herbert, Poole, Charles and Safko, John 2002. *Classical Mechanics*, New York: Addison Wesley, third edition.

Griffiths, David J. 1999. *Introduction to Electrodynamics*, Upper Saddle River NJ: Prentice-Hall, third edition.

Hafele, J. C., and Keating, Richard E. 1972a. "Around-the-World Atomic Clocks: Predicted Relativistic Time Gains," *Science*, **177**, 166-168.

Hafele, J. C., and Keating, Richard E. 1972b. "Around-the-World Aotmic Clocks: Observed Relativistic Time Gains," *Science*, **177**, 168-170.

Hartle, J. B. 2002. *Gravity: an Introduction to Einstein's General Relativity*, New York: Addison Wesley.

Hawking, Stephen W. 1992. "The Chronology Protection Conjecture," *Physical Review D*, **46**, 603-611.

Jackson, John David 1999. *Classical Electrodynamics*, New York: Wiley, third edition.

Kim, Sung-Won and Thorne, Kip S. 1991. "Do Vacuum Fluctuations Prevent the Creation of Closed Timelike Curves?" *Physical Review D*, **43**, 3929-3947.

Lorentz, Hendrik A. 1904. "Electromagnetic Phenomena in a System Moving with Any Velocity Less than that of Light," *Proceedings of the Academy of Sciences of Amsterdam*, **6**, 809; reprinted in Einstein *et al.* (1923).

Mathews, Jon and Walker, R. L. 1964. *Mathematical Methods of Physics*, New York: Benjamin.

Minkowski, Hermann 1908. "Space and Time," Address to the 80th Assembly of German Natural Scientists and Physicians, at Cologne, 21 September 1908; text published posthumously in *Annalen der Physik*, **47**, 927 (1915); English translation in Einstein *et al.* (1923).

Misner, Charles W., Thorne, Kip S., and Wheeler, John A. 1973. *Gravitation*, San Francisco: Freeman.

Morris, Michael S. and Thorne, Kip S. 1987. "Wormholes in Spacetime and their Use for interstellar Travel—a Tool for Teaching General Relativity," *American Journal of Physics*, **56**, 395-412.

Morris, Michael S., Thorne, Kip S., and Yurtsever, Ulvi 1987. "Wormholes, Time Machines, and the Weak Energy Condition," *Physical Review Letters*, **61**, 1446-1449.

Schutz, Bernard F. 1985. *A First Course in General Relativity*, Cambridge: Cambridge University Press.

Taylor, Edwin F. and Wheeler, John A. 1966. *Spacetime Physics: Introduction to Special Relativity*, San Francisco: Freeman, first edition.

Taylor, Edwin F. and Wheeler, John A. 1992. *Spacetime Physics: Introduction to Special Relativity*, San Francisco: Freeman, second edition.

Thorne, Kip S. 1993. "Closed Timelike Curves," in *General Relativity and Gravitation 1992*, edited by R. J. Gleiser, C. N. Kozameh and O. M. Moreschi, Cambridge University Press, Cambridge, England.

Contents

VI GENERAL RELATIVITY 2

23 From Special to General Relativity 1

23.1 Overview 1
23.2 Special Relativity Once Again 3
23.2.1 Geometric, Frame-Independent Formulation 3
23.2.2 Inertial Frames and Components of Vectors, Tensors and Physical Laws 5
23.2.3 Light Speed, the Interval, and Spacetime Diagrams 7
23.3 Differential Geometry in General Bases and in Curved Manifolds 8
23.3.1 Non-Orthonormal Bases 9
23.3.2 Vectors as Differential Operators; Tangent Space; Commutators . . . 13
23.3.3 Differentiation of Vectors and Tensors; Connection Coefficients 15
23.3.4 Integration 20
23.4 The Stress-Energy Tensor Revisited 25
23.5 The Proper Reference Frame of an Accelerated Observer [MTW pp. 163–176, 327–332] 30

Part VI

GENERAL RELATIVITY

Chapter 23

From Special to General Relativity

Version 0623.1.K.pdf, 25 April 2007.
Please send comments, suggestions, and errata via email to kip@tapir.caltech.edu or on paper to Kip Thorne, 130-33 Caltech, Pasadena CA 91125

Box 23.1
Reader's Guide

- This chapter relies significantly on
  - The special relativity portions of Chap. 1.
  - The discussion of connection coefficients in Sec. 10.3.
- This chapter is a foundation for the presentation of general relativity theory in Chaps. 24–27.

23.1 Overview

We have reached the final Part of this book, in which we present an introduction to the basic concepts of general relativity and its most important applications. This subject, although a little more challenging than the material that we have covered so far, is nowhere near as formidable as its reputation. Indeed, if you have mastered the techniques developed in the first five Parts, the path to the Einstein Field Equations should be short and direct.

The General Theory of Relativity is the crowning achievement of classical physics, the last great fundamental theory created prior to the discovery of quantum mechanics; its formulation by Albert Einstein in 1915 marks the culmination of the great intellectual adventure undertaken by Newton 250 years earlier. It was created after many wrong turns and with little experimental guidance, almost by pure thought. Unlike the special theory, whose physical foundations and logical consequences were clearly appreciated by physicists

soon after Einstein's 1905 formulation, the unique and distinctive character of the general theory only came to be widely appreciated long after its creation. Ultimately, in hindsight, rival classical theories of gravitation came to seem unnatural, inelegant and arbitrary by comparison.[1] Experimental tests of Einstein's theory also were slow to come; only in the last three decades have there been striking tests of high enough precision to convince most empiricists that, in all probability, and in its domain of applicability, general relativity is essentially correct. Despite this, it is still very poorly tested compared with, for example, quantum electrodynamics.

We begin our discussion of general relativity in this chapter with a review and elaboration of relevant material already covered in earlier chapters. In Sec. 23.2, we give a brief encapsulation of the special theory drawn largely from Chap. 1, emphasizing those aspects that we must generalize to deal with non-inertial frames of reference. Then in Sec. 23.3 we collect, review and extend the fundamental ideas of differential geometry that have been scattered throughout the book and which we shall need as foundations for the mathematics of *spacetime curvature* (Chap. 24); most importantly, we generalize differential geometry to encompass coordinate systems and bases that are not orthogonal. Einstein's field equations are a relationship between the curvature of spacetime and the matter that generates it, akin to the Maxwell equations' relationship between the electromagnetic field and electric currents and charges. The matter is described using the *stress-energy tensor* that we introduced in Sec. 1.12. We revisit the stress-energy tensor in Sec. 23.4 and develop a deeper understanding of its properties. In general relativity one often wishes to describe the outcome of measurements made by observers who refuse to fall freely—e.g., an observer who hovers in a spaceship just above the horizon of a black hole, or a gravitational-wave experimenter in an earth-bound laboratory. As a foundation for treating such observers, in Sec. 23.5 we examine measurements made by accelerated observers in the flat spacetime of special relativity.

This chapter will leave us well prepared to develop, in Chap. 24, the basic concepts of general relativity, including spacetime curvature, the Einstein Field Equation, and the laws of physics in curved spacetime. In Chaps. 25–27 we shall explore the major applications of general relativity: to stars, black holes, gravitational waves, and cosmology. We begin in Chap 25 by studying the spacetime curvature around and inside highly compact stars (such as neutron stars) and showing how, in the weak field limit, non-Newtonian effects are predicted in our own solar system and in binary neutron star systems and how these predictions have been verified. We also discuss the implosion of massive stars and describe the circumstances under which the implosion inevitably produces a black hole, and we explore the surprising and, initially, counter-intuitive properties of black holes. In Chap. 26 we study gravitational waves, i.e. ripples in the curvature of spacetime that propagate with the speed of light, and we explore their close analogy with the electromagnetic waves that were first predicted by Maxwell's equations. We explore the properties of these waves, their production by binary stars and merging black holes, projects to detect them, both on earth and in space, and the prospects for using them to explore observationally the dark side of the universe and the nature of ultrastrong spacetime curvature. Finally, in Chap. 27 we draw once more upon all the previous Parts of this book, combining them with general relativity to describe the universe on the largest of scales and longest of times: cosmology. It is here, more than

[1] For a readable account, see Will 1987 Was Einstein Right.

anywhere else in classical physics, that we are conscious of reaching a frontier where the still-promised land of quantum gravity beckons.

23.2 Special Relativity Once Again

A pre-requisite to learning the theory of general relativity is to understand special relativity in geometric language. In Chap. 1, we discussed the foundations of special relativity with this in mind and it is now time to remind ourselves of what we learned.

23.2.1 Geometric, Frame-Independent Formulation

In Chap. 1 we learned that *every law of physics must be expressible as a geometric, frame-independent relationship between geometric, frame-independent objects.* This is equally true in Newtonian physics, in special relativity and in general relativity. The key difference between the three is the geometric arena: In Newtonian physics the arena is 3-dimensional Euclidean space; in special relativity it is 4-dimensional Minkowski spacetime; in general relativity (Chap. 24) it is 4-dimensional curved spacetime; see Fig. 1.1 and associated discussion.

In special relativity, the demand that the laws be geometric relationships between geometric objects in Minkowski spacetime is called the *Principle of Relativity*; see Sec. 1.2.

Examples of the geometric objects are: (i) a point $\mathcal{P}$ in spacetime (which represents an *event*); (ii) a parametrized curve in spacetime such as the world line $\mathcal{P}(\tau)$ of a particle, for which the parameter τ is the particle's *proper time*, i.e. the time measured by an ideal clock[2] that the particle carries (Fig. 23.1); (iii) vectors such as the particle's 4-velocity $\vec{u} = d\mathcal{P}/d\tau$ [the tangent vector to the curve $\mathcal{P}(\tau)$] and the particle's 4-momentum $\vec{p} = m\vec{u}$ (with m the particle's rest mass); and (iv) tensors such as the electromagnetic field tensor $\mathbf{F}(\_\,,\_\,)$. A tensor, as we recall, is a linear real-valued function of vectors; when one puts vectors $\vec{A}$ and $\vec{B}$ into the slots of $\mathbf{F}$, one obtains a real number (a scalar) $\mathbf{F}(\vec{A},\vec{B})$ that is linear in $\vec{A}$ and in $\vec{B}$ so for example $\mathbf{F}(\vec{A}, b\vec{B}+c\vec{C}) = b\mathbf{F}(\vec{A},\vec{B}) + c\mathbf{F}(\vec{A},\vec{C})$. When one puts a vector $\vec{B}$ into just one of the slots of $\mathbf{F}$ and leaves the other empty, one obtains a tensor with one empty slot, $\mathbf{F}(\_\,,\vec{B})$, i.e. a vector. The result of putting a vector into the slot of a vector is the scalar product, $\vec{D}(\vec{B}) = \vec{D}\cdot\vec{B} = \mathbf{g}(\vec{D},\vec{B})$, where $\mathbf{g}(\_\,,\_\,)$ is the metric.

In Secs. 1.2 and 1.3 we tied our definitions of the inner product and the metric to the ticking of ideal clocks: If $\Delta\vec{x}$ is the vector separation of two neighboring events $\mathcal{P}(\tau)$ and $\mathcal{P}(\tau+\Delta\tau)$ along a particle's world line, then

$$\boxed{\mathbf{g}(\Delta\vec{x},\Delta\vec{x}) \equiv \Delta\vec{x}\cdot\Delta\vec{x} \equiv -(\Delta\tau)^2\ .} \tag{23.1}$$

This relation for any particle with any timelike world line, together with the linearity of $\mathbf{g}(\_\,,\_\,)$ in its two slots, is enough to determine $\mathbf{g}$ completely and to guarantee that it is

[2]Recall that an ideal clock is one that ticks uniformly when compared, e.g., to the period of the light emitted by some standard type of atom or molecule, and that has been made impervious to accelerations so two ideal clocks momentarily at rest with respect to each other tick at the same rate independent of their relative acceleration; cf. Secs. 1.2 and 1.4, and for greater detail, pp. 23–29 and 395–399 of MTW.

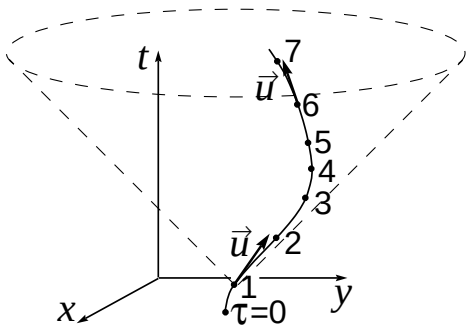


Fig. 23.1: The world line $\mathcal{P}(\tau)$ of a particle in Minkowski spacetime and the tangent vector $\vec{u} = d\mathcal{P}/d\tau$ to this world line; $\vec{u}$ is the particle's 4-velocity. The bending of the world line is produced by some force that acts on the particle, e.g. by the Lorentz force embodied in Eq. (23.3). Also shown is the light cone emitted from the event $\mathcal{P}(\tau = 1)$. Although the axes of an (arbitrary) inertial reference frame are shown, no reference frame is needed for the definition of the world line or its tangent vector $\vec{u}$ or the light cone, or for the formulation of the Lorentz force law.

symmetric, $\mathbf{g}(\vec{A}, \vec{B}) = \mathbf{g}(\vec{B}, \vec{A})$ for all $\vec{A}$ and $\vec{B}$. Since the particle's 4-velocity $\vec{u}$ is

$$\vec{u} = \frac{d\mathcal{P}}{d\tau} = \lim_{\Delta\tau\to 0} \frac{\mathcal{P}(\tau + \Delta\tau) - \mathcal{P}(\tau)}{\Delta\tau} \equiv \lim_{\Delta\tau\to 0} \frac{\Delta\vec{x}}{\Delta\tau} \,, \tag{23.2}$$

Eq. (23.1) implies that $\vec{u} \cdot \vec{u} = \mathbf{g}(\vec{u}, \vec{u}) = -1$.

The 4-velocity $\vec{u}$ is an example of a *timelike* vector; it has a negative inner product with itself (negative "squared length"). This shows up pictorially in the fact that $\vec{u}$ lies inside the *light cone* (the cone swept out by the trajectories of photons emitted from the tail of $\vec{u}$; see Fig. 23.1). Vectors $\vec{k}$ on the light cone (the tangents to the world lines of the photons) are *null* and so have vanishing squared lengths, $\vec{k} \cdot \vec{k} = \mathbf{g}(\vec{k}, \vec{k}) = 0$; and vectors $\vec{A}$ that lie outside the light cone are *spacelike* and have positive squared lengths, $\vec{A} \cdot \vec{A} > 0$.

An example of a physical law in 4-dimensional geometric language is the Lorentz force law

$$\frac{d\vec{p}}{d\tau} = q\mathbf{F}(\_, \vec{u}) \,, \tag{23.3}$$

where q is the particle's charge and both sides of this equation are vectors, i.e. first-rank tensors, i.e. tensors with just one slot. As we learned in Sec. 1.5, it is convenient to give names to slots. When we do so, we can rewrite the Lorentz force law as

$$\frac{dp^\alpha}{d\tau} = qF^{\alpha\beta}u_\beta \,. \tag{23.4}$$

Here α is the name of the slot of the vector $d\vec{p}/d\tau$, α and β are the names of the slots of $\mathbf{F}$, β is the name of the slot of $\mathbf{u}$, and the double use of β with one up and one down on the right side of the equation represents the insertion of $\vec{u}$ into the β slot of $\mathbf{F}$, whereby the two β slots disappear and we wind up with a vector whose slot is named α. As we learned in Sec. 1.5, this *slot-naming index notation* is isomorphic to the notation for components of vectors, tensors, and physical laws in some reference frame. However, no reference frames are

needed or involved when one formulates the laws of physics in geometric, frame-independent language as above.

Those readers who do not feel completely comfortable with these concepts, statements and notation should reread the relevant portions of Chap. 1.

23.2.2 Inertial Frames and Components of Vectors, Tensors and Physical Laws

In special relativity a key role is played by *inertial reference frames*. An inertial frame is an (imaginary) latticework of rods and clocks that moves through spacetime freely (inertially, without any force acting on it). The rods are orthogonal to each other and attached to inertial-guidance gyroscopes so they do not rotate. These rods are used to identify the spatial, Cartesian coordinates $(x^1, x^2, x^3) = (x, y, z)$ of an event $\mathcal{P}$ [which we also denote by lower case Latin indices $x^j(\mathcal{P})$ with j running over 1,2,3]. The latticework's clocks are ideal and are synchronized with each other via the Einstein light-pulse process (Sec. 1.2). They are used to identify the temporal coordinate $x^0 = t$ of an event $\mathcal{P}$; i.e. $x^0(\mathcal{P})$ is the time measured by that latticework clock whose world line passes through $\mathcal{P}$, at the moment of passage. The spacetime coordinates of $\mathcal{P}$ are denoted by lower case Greek indices x^α, with α running over 0,1,2,3. An inertial frame's spacetime coordinates $x^\alpha(\mathcal{P})$ are called *Lorentz coordinates* or *inertial coordinates*.

In the real universe, spacetime curvature is very small in regions well-removed from concentrations of matter, e.g. in intergalactic space; so special relativity is highly accurate there. In such a region, frames of reference (rod-clock latticeworks) that are non-accelerating and non-rotating with respect to cosmologically distant galaxies (and thence with respect to a local frame in which the cosmic microwave radiation looks isotropic) constitute good approximations to inertial reference frames.

Associated with an inertial frame's Lorentz coordinates are basis vectors $\vec{e}_\alpha$ that point along the frame's coordinate axes (and thus are orthogonal to each other) and have unit length (making them orthonormal). This *orthonormality* is embodied in the inner products

$$\boxed{\vec{e}_\alpha \cdot \vec{e}_\beta = \eta_{\alpha\beta} \ ,} \tag{23.5}$$

where by definition

$$\boxed{\eta_{00} = -1 \ , \quad \eta_{11} = \eta_{22} = \eta_{33} = +1 \ , \quad \eta_{\alpha\beta} = 0 \text{ if } \alpha \neq \beta \ .} \tag{23.6}$$

Here and throughout Part VI (as in Chap. 1), we set the speed of light to unity [i.e. we use the *geometrized units* discussed in Eqs. (1.3a) and (1.3b)], so spatial lengths (e.g. along the x axis) and time intervals (e.g. along the t axis) are measured in the same units, seconds or meters with $1 \text{ s} = 2.99792458 \times 10^8$ m.

In Sec. 1.5 we used the basis vectors of an inertial frame to build a component representation of tensor analysis. The fact that the inner products of timelike vectors with each other are negative, e.g. $\vec{e}_0 \cdot \vec{e}_0 = -1$, while those of spacelike vectors are positive, e.g. $\vec{e}_1 \cdot \vec{e}_1 = +1$, forced us to introduce two types of components: *covariant* (indices down) and *contravariant* (indices up). The covariant components of a tensor were computable by inserting the basis

vectors into the tensor's slots, $u_\alpha = \vec{u}(\vec{e}_\alpha) \equiv \vec{u} \cdot \vec{e}_\alpha$; $F_{\alpha\beta} = \mathbf{F}(\vec{e}_\alpha, \vec{e}_\beta)$. For example, in our Lorentz basis the covariant components of the metric are $g_{\alpha\beta} = \mathbf{g}(\vec{e}_\alpha, \vec{e}_\beta) = \vec{e}_\alpha \cdot \vec{e}_\beta = \eta_{\alpha\beta}$. The contravariant components of a tensor were related to the covariant components via "index lowering" with the aid of the metric, $F_{\alpha\beta} = g_{\alpha\mu} g_{\beta\nu} F^{\mu\nu}$, which simply said that one reverses the sign when lowering a time index and makes no change of sign when lowering a space index. This lowering rule implied that the contravariant components of the metric in a Lorentz basis are the same numerically as the covariant components, $g^{\alpha\beta} = \eta_{\alpha\beta}$ and that they can be used to raise indices (i.e. to perform the trivial sign flip for temporal indices) $F^{\mu\nu} = g^{\mu\alpha} g^{\nu\beta} F_{\alpha\beta}$. As we saw in Sec. 1.5, tensors can be expressed in terms of their contravariant components as $\vec{p} = p^\alpha \vec{e}_\alpha$, and $\mathbf{F} = F^{\alpha\beta} \vec{e}_\alpha \otimes \vec{e}_\beta$, where $\otimes$ represents the tensor product [Eq. (1.10a)].

We also learned in Chap. 1 that any frame independent geometric relation between tensors can be rewritten as a relation between those tensors' components in any chosen Lorentz frame. When one does so, the resulting component equation takes *precisely the same form* as the slot-naming-index-notation version of the geometric relation. For example, the component version of the Lorentz force law says $dp^\alpha/d\tau = qF^{\alpha\beta}u_\beta$, which is identical to Eq. (23.4). The only difference is the interpretation of the symbols. In the component equation $F^{\alpha\beta}$ are the components of $\mathbf{F}$ and the repeated β in $F^{\alpha\beta}u_\beta$ is to be summed from 0 to 3. In the geometric relation $F^{\alpha\beta}$ means $\mathbf{F}(\_\,,\_\,)$ with the first slot named α and the second β, and the repeated β in $F^{\alpha\beta}u_\beta$ implies the insertion of $\vec{u}$ into the second slot of $\mathbf{F}$ to produce a single-slotted tensor, i.e. a vector whose slot is named α.

As we saw in Sec. 1.6, a particle's 4-velocity $\vec{u}$ (defined originally without the aid of any reference frame; Fig. 23.1) has components, in any inertial frame, given by $u^0 = \gamma$, $u^j = \gamma v^j$ where $v^j = dx^j/dt$ is the particle's ordinary velocity and $\gamma \equiv 1/\sqrt{1 - \delta_{ij} v^i v^j}$. Similarly, the particle's energy $E \equiv p^0$ is $m\gamma$ and its spatial momentum is $p^j = m\gamma v^j$, i.e. in 3-dimensional geometric notation, $\mathbf{p} = m\gamma\mathbf{v}$. This is an example of the manner in which a choice of Lorentz frame produces a "3+1" split of the physics: a split of 4-dimensional spacetime into 3-dimensional space (with Cartesian coordinates x^j) plus 1-dimensional time $t = x^0$; a split of the particle's 4-momentum $\vec{p}$ into its 3-dimensional spatial momentum $\mathbf{p}$ and its 1-dimensional energy $E = p^0$; and similarly a split of the electromagnetic field tensor $\mathbf{F}$ into the 3-dimensional electric field $\mathbf{E}$ and 3-dimensional magnetic field $\mathbf{B}$; cf. Secs. 1.6 and 1.10.

The principle of relativity (all laws expressible as geometric relations between geometric objects in Minkowski spacetime), when translated into 3+1 language, says that, *when the laws of physics are expressed in terms of components in a specific Lorentz frame, the form of those laws must be independent of one's choice of frame.* The components of tensors in one Lorentz frame are related to those in another by a Lorentz transformation (Sec. 1.7), so the principle of relativity can be restated as saying that, when expressed in terms of Lorentz-frame components, *the laws of physics must be Lorentz-invariant* (unchanged by Lorentz transformations). This is the version of the principle of relativity that one meets in most elementary treatments of special relativity. However, as the above discussion shows, it is a mere shadow of the true principle of relativity—the shadow cast onto Lorentz frames when one performs a 3+1 split. The ultimate, fundamental version of the principle of relativity is the one that needs no frames at all for its expression: *All the laws of physics are expressible*

as geometric relations between geometric objects that reside in Minkowski spacetime.

If the above discussion is not completely clear, the reader should study the relevant portions of Chap. 1.

23.2.3 Light Speed, the Interval, and Spacetime Diagrams

One set of physical laws that must be the same in all inertial frames is Maxwell's equations. Let us discuss the implications of Maxwell's equations for the speed of light c, momentarily abandoning geometrized units and returning to mks/SI units. According to Maxwell, c can be determined by performing non-radiative laboratory experiments; it is not necessary to measure the time it takes light to travel along some path. For example, measure the electrostatic force between two charges; that force is $\propto \epsilon_0^{-1}$, the electric permitivity of free space. Then allow one of these charges to travel down a wire and by its motion generate a magnetic field. Let the other charge move through this field and measure the magnetic force on it; that force is $\propto \mu_0$, the magnetic permitivity of free space. The ratio of these two forces can be computed and is $\propto 1/\mu_0\epsilon_0 = c^2$. By combining the results of the two experiments, we therefore can deduce the speed of light c; this is completely analogous to deducing the speed of seismic waves through rock from a knowledge of the rock's density and elastic moduli, using elasticity theory (Chap. 11). The principle of relativity, in operational form, dictates that the results of the electric and magnetic experiments must be independent of the Lorentz frame in which one chooses to perform them; therefore, the speed of light is frame-independent—as we argued by a different route in Sec. 1.2. It is this frame independence that enables us to introduce geometrized units with $c = 1$.

Another example of frame independence (Lorentz invariance) is provided by the *interval between two events*. The components $g_{\alpha\beta} = \eta_{\alpha\beta}$ of the metric imply that, if $\Delta\vec{x}$ is the vector separating the two events and Δx^α are its components in some Lorentz coordinate system, then the squared length of $\Delta\vec{x}$ [also called the *interval* and denoted $(\Delta s)^2$] is given by

$$\boxed{(\Delta s)^2 \equiv \Delta\vec{x}\cdot\Delta\vec{x} = \mathbf{g}(\Delta\vec{x},\Delta\vec{x}) = g_{\alpha\beta}\Delta x^\alpha \Delta x^\beta = -(\Delta t)^2 + (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 \ .} \tag{23.7}$$

Since $\Delta\vec{x}$ is a geometric, frame-independent object, so must be the interval. This implies that the equation $(\Delta s)^2 = -(\Delta t)^2 + (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2$ by which one computes the interval between the two chosen events in one Lorentz frame must give the same numerical result when used in any other frame; i.e., this expression must be Lorentz invariant. This *invariance of the interval* is the starting point for most introductions to special relativity—and, indeed, we used it as a starting point in Sec. 1.2.

Spacetime diagrams will play a major role in our development of general relativity. Accordingly, it is important that the reader feel very comfortable with them. We recommend reviewing Fig. 1.12 and Ex. 1.14.

****

EXERCISES

Exercise 23.1 *Example: Invariance of a Null Interval*

You have measured the intervals between a number of adjacent events in spacetime and thereby have deduced the metric $\mathbf{g}$. Your friend claims that the metric is some other frame-independent tensor $\tilde{\mathbf{g}}$ that differs from $\mathbf{g}$. Suppose that your correct metric $\mathbf{g}$ and his wrong one $\tilde{\mathbf{g}}$ agree on the forms of the light cones in spacetime, i.e. they agree as to which intervals are null, which are spacelike and which are timelike; but they give different answers for the value of the interval in the spacelike and timelike cases, i.e. $\mathbf{g}(\Delta\vec{x}, \Delta\vec{x}) \neq \tilde{\mathbf{g}}(\Delta\vec{x}, \Delta\vec{x})$. Prove that $\tilde{\mathbf{g}}$ and $\mathbf{g}$ differ solely by a scalar multiplicative factor. [*Hint*: pick some Lorentz frame and perform computations there, then lift yourself back up to a frame-independent viewpoint.]

Exercise 23.2 *Problem: Causality*
If two events occur at the same spatial point but not simultaneously in one inertial frame, prove that the temporal order of these events is the same in all inertial frames. Prove also that in all other frames the temporal interval Δt between the two events is larger than in the first frame, and that there are no limits on the events' spatial or temporal separation in the other frames. Give *two* proofs of these results, one algebraic and the other via spacetime diagrams.

****

23.3 Differential Geometry in General Bases and in Curved Manifolds

The tensor-analysis formalism reviewed in the last section is inadequate for general relativity in several ways:

First, in general relativity we shall need to use bases $\vec{e}_\alpha$ that are not orthornormal, i.e. for which $\vec{e}_\alpha \cdot \vec{e}_\beta \neq \eta_{\alpha\beta}$. For example, near a spinning black hole there is much power in using a time basis vector $\vec{e}_t$ that is tied in a simple way to the metric's time-translation symmetry and a spatial basis vector $\vec{e}_\phi$ that is tied to its rotational symmetry. This time basis vector has an inner product with itself $\vec{e}_t \cdot \vec{e}_t = g_{tt}$ that is influenced by the slowing of time near the hole so $g_{tt} \neq -1$; and $\vec{e}_\phi$ is not orthogonal to $\vec{e}_t$, $\vec{e}_t \cdot \vec{e}_\phi = g_{t\phi} \neq 0$, as a result of the dragging of inertial frames by the hole's spin. In this section we shall generalize our formalism to treat such non-orthonormal bases.

Second, in the curved spacetime of general relativity (and in any other curved manifold, e.g. the two-dimensional surface of the earth) the definition of a vector as an arrow connecting two points is suspect, as it is not obvious on what route the arrow should travel nor that the linear algebra of tensor analysis should be valid for such arrows. In this section we shall refine the concept of a vector to deal with this problem, and in the process we shall find ourselves introducing the concept of a *tangent space* in which the linear algebra of tensors takes place—a different tangent space for tensors that live at different points in the manifold.

Third, once we have been forced to think of a tensor as residing in a specific tangent space at a specific point in the manifold, there arises the question of how one can transport tensors from the tangent space at one point to the tangent space at an adjacent point. Since

the notion of a gradient of a vector depends on comparing the vector at two different points and thus depends on the details of transport, we will have to rework the notion of a gradient and the gradient's connection coefficients; and since, in doing an integral, one must add contributions that live at different points in the manifold, we must also rework the notion of integration.

We shall tackle each of these three issues in turn in the following four subsections.

23.3.1 Non-Orthonormal Bases

Consider an n-dimensional manifold, e.g. 4-dimensional spacetime or 3-dimensional Euclidean space or the 2-dimensional surface of a sphere. At some point $\mathcal{P}$ in the manifold, introduce a set of basis vectors $\{\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_n\}$ and denote them generally as $\vec{e}_\alpha$. We seek to generalize the formalism of Sec. 23.2 in such a way that the index manipulation rules for components of tensors are unchanged. For example, we still want it to be true that covariant components of any tensor are computable by inserting the basis vectors into the tensor's slots, $F_{\alpha\beta} = \mathbf{F}(\vec{e}_\alpha, \vec{e}_\beta)$, and that the tensor itself can be reconstructed from its contravariant components as $\mathbf{F} = F^{\mu\nu}\vec{e}_\mu \otimes \vec{e}_\nu$, and that the two sets of components are computable from each other via raising and lowering with the metric components, $F_{\alpha\beta} = g_{\alpha\mu}g_{\beta\nu}F^{\mu\nu}$. The only thing we do not want to preserve is the orthonormal values of the metric components; i.e. we must allow the basis to be nonorthonormal and thus $\vec{e}_\alpha \cdot \vec{e}_\beta = g_{\alpha\beta}$ to have arbitrary values (except that the metric should be nondegenerate, so no linear combination of the $\vec{e}_\alpha$'s vanishes, which means that the matrix $||g_{\alpha\beta}||$ should have nonzero determinant).

We can easily achieve our goal by introducing a second set of basis vectors, denoted $\{\vec{e}^1, \vec{e}^2, \ldots, \vec{e}^n\}$, which is *dual* to our first set in the sense that

$$\boxed{\vec{e}^\mu \cdot \vec{e}_\beta \equiv \mathbf{g}(\vec{e}^\mu, \vec{e}_\beta) = \delta^\mu{}_\beta} \tag{23.8}$$

where $\delta^\alpha{}_\beta$ is the Kronecker delta. This duality relation actually constitutes a *definition* of the e^μ once the $\vec{e}_\alpha$ have been chosen. To see this, regard $\vec{e}^\mu$ as a tensor of rank one. This tensor is defined as soon as its value on each and every vector has been determined. Expression (23.8) gives the value $\vec{e}^\mu(\vec{e}_\beta) = \vec{e}^\mu \cdot \vec{e}_\beta$ of $\vec{e}^\mu$ on each of the four basis vectors $\vec{e}_\beta$; and since every other vector can be expanded in terms of the $\vec{e}_\beta$'s and $\vec{e}^\mu(\_)$ is a linear function, Eq. (23.8) thereby determines the value of $\vec{e}^\mu$ on every other vector.

The duality relation (23.8) says that $\vec{e}^1$ is always perpendicular to all the $\vec{e}_\alpha$ except $\vec{e}_1$; and its scalar product with $\vec{e}_1$ is unity—and similarly for the other basis vectors. This interpretation is illustrated for 3-dimensional Euclidean space in Fig. 23.2. In Minkowski spacetime, if $\vec{e}_\alpha$ are an orthonormal Lorentz basis, then duality dictates that $\vec{e}^0 = -\vec{e}_0$, and $\vec{e}^j = +\vec{e}_j$.

The duality relation (23.8) leads immediately to the same index-manipulation formalism as we have been using, if one defines the contravariant, covariant and mixed components of tensors in the obvious manner

$$\boxed{F^{\mu\nu} = \mathbf{F}(\vec{e}^\mu, \vec{e}^\nu)\ , \quad F_{\alpha\beta} = \mathbf{F}(\vec{e}_\alpha, \vec{e}_\beta)\ , \quad F^\mu{}_\beta = \mathbf{F}(\vec{e}^\mu, \vec{e}_\beta)\ ;} \tag{23.9}$$

see Ex. 23.4. Among the consequences of this duality are the following: (i)

$$\boxed{g^{\mu\beta}g_{\nu\beta} = \delta^\mu_\nu\ ,} \tag{23.10}$$

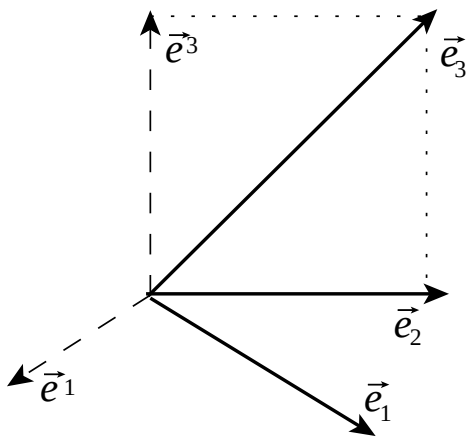


Fig. 23.2: Non-orthonormal basis vectors $\vec{e}_j$ in Euclidean 3-space and two members $\vec{e}^1$ and $\vec{e}^3$ of the dual basis. The vectors $\vec{e}_1$ and $\vec{e}_2$ lie in the horizontal plane, so $\vec{e}^3$ is orthogonal to that plane, i.e. it points vertically upward, and its inner product with $\vec{e}_3$ is unity. Similarly, the vectors $\vec{e}_2$ and $\vec{e}_3$ span a vertical plane, so $\vec{e}^1$ is orthogonal to that plane, i.e. it points horizontally, and its inner product with $\vec{e}_1$ is unity.

i.e., the matrix of contravariant components of the metric is inverse to that of the covariant components, $||g^{\mu\nu}|| = ||g_{\alpha\beta}||^{-1}$; this relation guarantees that when one raises indices on a tensor $F_{\alpha\beta}$ with $g^{\mu\alpha}$ and then lowers them back down with $g_{\nu\beta}$, one recovers one's original covariant components $F_{\alpha\beta}$ unaltered. (ii)

$$\boxed{\mathbf{F} = F^{\mu\nu}\vec{e}_\mu \otimes \vec{e}_\nu = F_{\alpha\beta}\vec{e}^\alpha \otimes \vec{e}^\beta = F^\mu{}_\beta \vec{e}_\mu \otimes \vec{e}^\beta \ ,} \tag{23.11}$$

i.e., one can reconstruct a tensor from its components by lining up the indices in a manner that accords with the rules of index manipulation. (iii)

$$\boxed{\mathbf{F}(\vec{p},\vec{q}) = F^{\alpha\beta}p_\alpha p_\beta \ ,} \tag{23.12}$$

i.e., the component versions of tensorial equations are identical in mathematical symbology to the slot-naming-index-notation versions.

Associated with any coordinate system $x^\alpha(\mathcal{P})$ there is a *coordinate basis* whose basis vectors are defined by

$$\boxed{\vec{e}_\alpha \equiv \frac{\partial \mathcal{P}}{\partial x^\alpha} \ .} \tag{23.13}$$

Since the derivative is taken holding the other coordinates fixed, the basis vector $\vec{e}_\alpha$ points along the α coordinate axis (the axis on which x^α changes and all the other coordinates are held fixed).

In an orthogonal curvilinear coordinate system, e.g. circular polar coordinates (ϖ, ϕ) in Euclidean 2-space, this coordinate basis is quite different from the coordinate system's orthonormal basis. For example, $\vec{e}_\phi = (\partial\mathcal{P}/\partial\phi)_\varpi$ is a very long vector at large radii and a very short vector at small radii [cf. Fig. 23.3]; the corresponding unit-length vector is $\vec{e}_{\hat{\phi}} = (1/\varpi)\vec{e}_\phi$. By contrast, $\vec{e}_\varpi = (\partial\mathcal{P}/\partial\varpi)_\phi$ already has unit length, so the corresponding

Box 23.1
Dual Bases in Other Contexts

Vector spaces appear in a wide variety of contexts in mathematics and physics, and wherever they appear it can be useful to introduce dual bases.

When a vector space does not posses a metric, the basis $\{\vec{e}^{\mu}\}$ lives in a different space from $\{\vec{e}_\alpha\}$, and the two spaces are said to be dual to each other. An important example occurs in manifolds that do not have metrics. There the vectors in the space spanned by $\{\vec{e}^{\mu}\}$ are often called a *one forms* and are represented pictorially as families of parallel surfaces; the vectors in the space spanned by $\{\vec{e}_\alpha\}$ are called *tangent vectors* and are represented pictorially as arrows; the one forms are linear functions of tangent vectors, and the result that a one form $\tilde{\beta}$ gives when a tangent vector $\vec{a}$, is inserted into its slot, $\tilde{\beta}(\vec{a})$, is the number of surfaces of $\tilde{\beta}$ pierced by the arrow $\vec{a}$; see, e.g., MTW. A metric produces a one-to-one mapping between the one forms and the tangent vectors. In this book we regard this mapping as equating each one form to a tangent vector and thereby as making the space of one forms and the space of tangent vectors be identical. This permits us to avoid ever speaking about one forms, except here in this box.

Quantum mechanics provides another example of dual spaces. The kets $|\psi\rangle$ are the tangent vectors and the bras $\langle\phi|$ are the one forms: linear *complex valued* functions of kets with the value that $\langle\phi|$ gives when $|\psi\rangle$ is inserted into its slot being the inner product $\langle\phi|\psi\rangle$.

orthonormal basis vector is simply $\vec{e}_{\hat{\varpi}} = \vec{e}_\varpi$. The metric components in the coordinate basis are readily seen to be $g_{\phi\phi} = \varpi^2$, $g_{\varpi\varpi} = 1$, $g_{\varpi\phi} = g_{\phi\varpi} = 0$ which is in accord with the equation for the squared distance (interval) between adjacent points $ds^2 = g_{ij}dx^i dx^j = d\varpi^2 + \varpi^2 d\phi^2$. The metric components in the orthonormal basis, of course, are $g_{\hat{i}\hat{j}} = \delta_{ij}$.

Henceforth, we shall use hats to identify orthonormal bases; bases whose indices do not have hats will typically (though not always) be coordinate bases.

In general, we can construct the basis $\{\vec{e}^{\mu}\}$ that is dual to the coordinate basis $\{\vec{e}_\alpha\} = \{\partial\mathcal{P}/\partial x^\alpha\}$ by taking the gradients of the coordinates, viewed as scalar fields $x^\alpha(\mathcal{P})$:

$$\boxed{\vec{e}^{\mu} = \vec{\nabla}x^\mu \ .} \tag{23.14}$$

It is straightforward to verify the duality relation (23.8) for these two bases:

$$\vec{e}^{\mu} \cdot \vec{e}_\alpha = \vec{e}_\alpha \cdot \vec{\nabla}x^\mu = \nabla_{\vec{e}_\alpha} x^\mu = \nabla_{\partial\mathcal{P}/\partial x^\alpha} x^\mu = \frac{\partial x^\mu}{\partial x^\alpha} = \delta^\mu_\alpha \ . \tag{23.15}$$

In any coordinate system, the expansion of the metric in terms of the dual basis, $\mathbf{g} = g_{\alpha\beta}\vec{e}^{\alpha} \otimes \vec{e}^{\beta} = g_{\alpha\beta}\vec{\nabla}x^\alpha \otimes \vec{\nabla}x^\beta$ is intimately related to the line element $ds^2 = g_{\alpha\beta}dx^\alpha dx^\beta$: Consider an infinitesimal vectorial displacement $d\vec{x} = dx^\alpha(\partial/\partial x^\alpha)$. Insert this displacement into the metric's two slots, to obtain the interval ds^2 along $d\vec{x}$. The result is $ds^2 = g_{\alpha\beta}\boldsymbol{\nabla}x^\alpha \otimes \boldsymbol{\nabla}x^\beta(d\vec{x}, d\vec{x}) = g_{\alpha\beta}(d\vec{x}\cdot\boldsymbol{\nabla}x^\alpha)(d\vec{x}\cdot\boldsymbol{\nabla}x^\beta) = g_{\alpha\beta}dx^\alpha dx^\beta$; i.e.

$$\boxed{ds^2 = g_{\alpha\beta}dx^\alpha dx^\beta \ .} \tag{23.16}$$

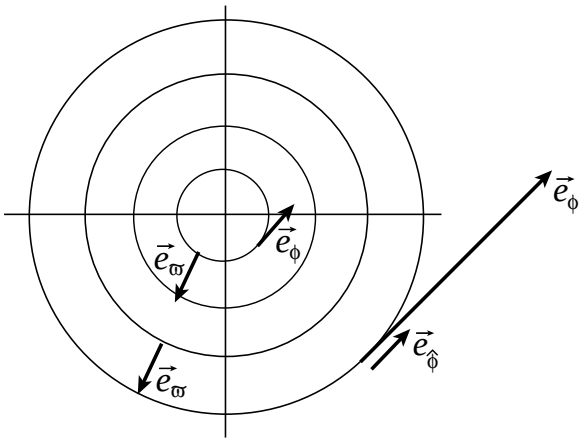


Fig. 23.3: A circular coordinate system $\{\varpi, \phi\}$ and its coordinate basis vectors $\vec{e}_\varpi = \partial\mathcal{P}/\partial\varpi$, $\vec{e}_\phi = \partial\mathcal{P}/\partial\phi$ at several locations in the coordinate system. Also shown is the orthonormal basis vector $\vec{e}_{\hat{\phi}}$.

Here the second equality follows from the definition of the tensor product $\otimes$, and the third from the fact that for any scalar field ψ, $d\vec{x} \cdot \boldsymbol{\nabla}\psi$ is the change $d\psi$ along $d\vec{x}$.

Any two bases $\{\vec{e}_\alpha\}$ and $\{\vec{e}_{\bar{\mu}}\}$ can be expanded in terms of each other:

$$\boxed{\vec{e}_\alpha = \vec{e}_{\bar{\mu}} L^{\bar{\mu}}{}_\alpha \ , \quad \vec{e}_{\bar{\mu}} = \vec{e}_\alpha L^\alpha{}_{\bar{\mu}} \ .} \tag{23.17}$$

(Note: by convention the first index on L is always placed up and the second is always placed down.) The quantities $||L^{\bar{\mu}}{}_\alpha||$ and $||L^\alpha{}_{\bar{\mu}}||$ are transformation matrices and since they operate in opposite directions, they must be the inverse of each other

$$\boxed{L^{\bar{\mu}}{}_\alpha L^\alpha{}_{\bar{\nu}} = \delta^{\bar{\mu}}{}_{\bar{\nu}} \ , \quad L^\alpha{}_{\bar{\mu}} L^{\bar{\mu}}{}_\beta = \delta^\alpha{}_\beta \ .} \tag{23.18}$$

These $||L^{\bar{\mu}}{}_\alpha||$ are the generalizations of Lorentz transformations to arbitrary bases; cf. Eqs. (1.47a), (1.47b). As in the Lorentz-transformation case, the transformation laws (23.17) for the basis vectors imply corresponding transformation laws for components of vectors and tensors—laws that entail lining up indices in the obvious manner; e.g.

$$\boxed{A_{\bar{\mu}} = L^\alpha{}_{\bar{\mu}} A_\alpha \ , \quad T^{\bar{\mu}\bar{\nu}}{}_{\bar{\rho}} = L^{\bar{\mu}}{}_\alpha L^{\bar{\nu}}{}_\beta L^\gamma{}_{\bar{\rho}} T^{\alpha\beta}{}_\gamma \ , \quad \text{and similarly in the opposite direction.}} \tag{23.19}$$

For coordinate bases, these $L^{\bar{\mu}}{}_\alpha$ are simply the partial derivatives of one set of coordinates with respect to the other

$$\boxed{L^{\bar{\mu}}{}_\alpha = \frac{\partial x^{\bar{\mu}}}{\partial x^\alpha} \ , \quad L^\alpha{}_{\bar{\mu}} = \frac{\partial x^\alpha}{\partial x^{\bar{\mu}}} \ ,} \tag{23.20}$$

as one can easily deduce via

$$\vec{e}_\alpha = \frac{\partial\mathcal{P}}{\partial x^\alpha} = \frac{\partial x^\mu}{\partial x^\alpha}\frac{\partial\mathcal{P}}{\partial x^\mu} = \vec{e}_\mu \frac{\partial x^\mu}{\partial x^\alpha} \ . \tag{23.21}$$

In many physics textbooks a tensor is *defined* as a set of components $F_{\alpha\beta}$ that obey the transformation laws

$$F_{\alpha\beta} = F_{\mu\nu}\frac{\partial x^\mu}{\partial x^\alpha}\frac{\partial x^\nu}{\partial x^\beta} \ . \tag{23.22}$$

This definition is in accord with Eqs. (23.19) and (23.20), though it hides the true and very simple nature of a tensor as a linear function of frame-independent vectors.

23.3.2 Vectors as Differential Operators; Tangent Space; Commutators

As was discussed above, the notion of a vector as an arrow connecting two points is problematic in a curved manifold, and must be refined. As a first step in the refinement, let us consider the tangent vector $\vec{A}$ to a curve $\mathcal{P}(\zeta)$ at some point $\mathcal{P}_o \equiv \mathcal{P}(\zeta = 0)$. We have defined that tangent vector by the limiting process

$$\boxed{\vec{A} \equiv \frac{d\mathcal{P}}{d\zeta} \equiv \lim_{\Delta\zeta\to 0}\frac{\mathcal{P}(\Delta\zeta) - \mathcal{P}(0)}{\Delta\zeta} \ ;} \tag{23.23}$$

cf. Eq. (23.2). In this definition the difference $\mathcal{P}(\zeta) - \mathcal{P}(0)$ means the tiny arrow reaching from $\mathcal{P}(0) \equiv \mathcal{P}_o$ to $\mathcal{P}(\Delta\zeta)$. In the limit as $\Delta\zeta$ becomes vanishingly small, these two points get arbitrarily close together; and in such an arbitrarily small region of the manifold, the effects of the manifold's curvature become arbitrarily small and negligible (just think of an arbitrarily tiny region on the surface of a sphere), so the notion of the arrow should become sensible. However, before the limit is completed, we are required to divide by $\Delta\zeta$, which makes our arbitrarily tiny arrow big again. What meaning can we give to this?

One way to think about it is to imagine embedding the curved manifold in a higher dimensional flat space (e.g., embed the surface of a sphere in a flat 3-dimensional Euclidean space as shown in Fig. 23.4). Then the tiny arrow $\mathcal{P}(\Delta\zeta) - \mathcal{P}(0)$ can be thought of equally well as lying on the sphere, or as lying in a surface that is tangent to the sphere and is flat, as measured in the flat embedding space. We can give meaning to $[\mathcal{P}(\Delta\zeta) - \mathcal{P}(0)]/\Delta\zeta$ if we regard this as a formula for lengthening an arrow-type vector in the flat tangent surface; correspondingly, we must regard the resulting tangent vector $\vec{A}$ as an arrow living in the tangent surface.

The (conceptual) flat tangent surface at the point $\mathcal{P}_o$ is called the *tangent space* to the curved manifold at that point. It has the same number of dimensions n as the manifold itself (two in the case of Fig. 23.4). Vectors at $\mathcal{P}_o$ are arrows residing in that point's tangent space, tensors at $\mathcal{P}_o$ are linear functions of these vectors, and all the linear algebra of vectors and tensors that reside at $\mathcal{P}_o$ occurs in this tangent space. For example, the inner product of two vectors $\vec{A}$ and $\vec{B}$ at $\mathcal{P}_o$ (two arrows living in the tangent space there) is computed via the standard relation $\vec{A}\cdot\vec{B} = \mathbf{g}(\vec{A},\vec{B})$ using the metric $\mathbf{g}$ that also resides in the tangent space.

This pictorial way of thinking about the tangent space and vectors and tensors that reside in it is far too heuristic to satisfy most mathematicians. Therefore, mathematicians have insisted on making it much more precise at the price of greater abstraction: *Mathematicians*

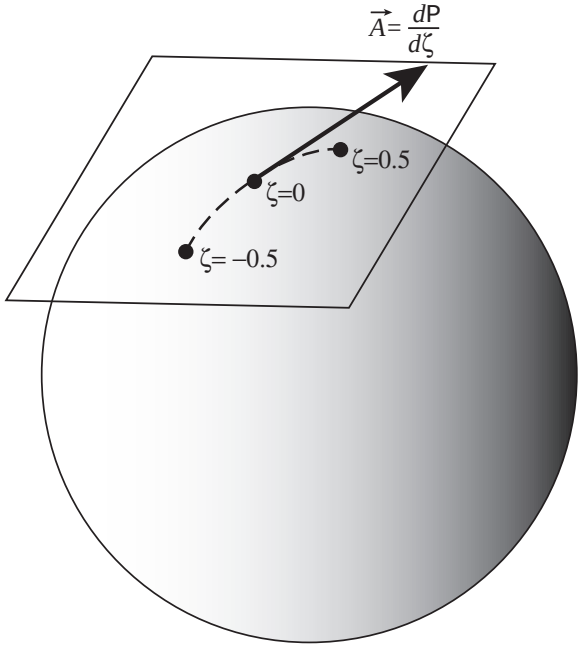


Fig. 23.4: A curve $\mathcal{P}(\zeta)$ on the surface of a sphere and the curve's tangent vector $\vec{A} = d\mathcal{P}/d\zeta$ at $\mathcal{P}(\zeta = 0) \equiv \mathcal{P}_o$. The tangent vector lives in the tangent space at $\mathcal{P}_o$, i.e. in the flat plane that is tangent to the sphere there as seen in the flat Euclidean 3-space in which the sphere's surface is embedded.

define the tangent vector to the curve $\mathcal{P}(\zeta)$ to be the derivative $d/d\zeta$ which differentiates scalar fields along the curve. This derivative operator is very well defined by the rules of ordinary differentiation; if $\psi(\mathcal{P})$ is a scalar field in the manifold, then $\psi[\mathcal{P}(\zeta)]$ is a function of the real variable ζ, and its derivative $(d/d\zeta)\psi[\mathcal{P}(\zeta)]$ evaluated at $\zeta = 0$ is the ordinary derivative of elementary calculus. Since the derivative operator $d/d\zeta$ differentiates in the manifold along the direction in which the curve is moving, it is often called the *directional derivative* along $\mathcal{P}(\zeta)$. Mathematicians notice that all the directional derivatives at a point $\mathcal{P}_o$ of the manifold form a vector space (they can be multiplied by scalars and added and subtracted to get new vectors), and so they define this vector space to be the tangent space at $\mathcal{P}_o$.

This mathematical procedure turns out to be isomorphic to the physicists' more heuristic way of thinking about the tangent space. In physicists' language, if one introduces a coordinate system in a region of the manifold containing $\mathcal{P}_o$ and constructs the corresponding coordinate basis $\vec{e}_\alpha = \partial\mathcal{P}/\partial x^\alpha$, then one can expand any vector in the tangent space as $\vec{A} = A^\alpha \partial\mathcal{P}/\partial x^\alpha$. One can also construct, in physicists' language, the directional derivative along $\vec{A}$; it is $\partial_{\vec{A}} \equiv A^\alpha \partial/\partial x^\alpha$. Evidently, the components A^α of the physicist's vector $\vec{A}$ (an arrow) are identical to the coefficients A^α in the coordinate-expansion of the directional derivative $\partial_{\vec{A}}$. There therefore is a one-to-one correspondence between the directional derivatives $\partial_{\vec{A}}$ at $\mathcal{P}_o$ and the vectors $\vec{A}$ there, and a complete isomorphism between the tangent-space manipulations that a mathematician will perform treating the directional derivatives as vectors, and those that a physicist will perform treating the arrows as vectors.

"Why not abandon the fuzzy concept of a vector as an arrow, and *redefine the vector* $\vec{A}$ *to be the same as the directional derivative* $\partial_{\vec{A}}$?" mathematicians have demanded of physicists. Slowly, over the past century, physicists have come to see the merit in this approach: (*i*) It does, indeed, make the concept of a vector more rigorous than before. (*ii*) It simplifies a number of other concepts in mathematical physics, e.g., the commutator of two vector fields; see below. (*iii*) It facilitates communication with mathematicians. With these motivations in mind, and because one always gains conceptual and computational power by having multiple viewpoints at one's finger tips (see, e.g., Feynman, 1966), we shall regard vectors henceforth *both* as arrows living in a tangent space and as directional derivatives. Correspondingly, we shall assert the equalities

$$\boxed{\frac{\partial \mathcal{P}}{\partial x^\alpha} = \frac{\partial}{\partial x^\alpha} \quad , \quad \vec{A} = \partial_{\vec{A}} \, ,} \tag{23.24}$$

and shall often expand vectors in a coordinate basis using the notation

$$\boxed{\vec{A} = A^\alpha \frac{\partial}{\partial x^\alpha} \, .} \tag{23.25}$$

This directional-derivative viewpoint on vectors makes natural the concept of the *commutator* of two vector fields $\vec{A}$ and $\vec{B}$: $[\vec{A}, \vec{B}]$ is the vector which, when viewed as a differential operator, is given by $[\partial_{\vec{A}}, \partial_{\vec{B}}]$—where the latter quantity is the same commutator as one meets elsewhere in physics, e.g. in quantum mechanics. Using this definition, we can compute the components of the commutator in a coordinate basis:

$$\boxed{[\vec{A}, \vec{B}] \equiv \left[A^\alpha \frac{\partial}{\partial x^\alpha}, B^\beta \frac{\partial}{\partial x^\beta} \right] = \left(A^\alpha \frac{\partial B^\beta}{\partial x^\alpha} - B^\alpha \frac{\partial A^\beta}{\partial x^\alpha} \right) \frac{\partial}{\partial x^\beta} \, .} \tag{23.26}$$

This is an operator equation where the final derivative is presumed to operate on a scalar field just as in quantum mechanics. From this equation we can read off the components of the commutator in any coordinate basis; they are $A^\alpha B^\beta{}_{,\alpha} - B^\alpha A^\beta{}_{,\alpha}$, where the comma denotes partial differentiation. Figure 23.5 uses this equation to deduce the geometric meaning of the commutator: it is the fifth leg needed to close a quadrilateral whose other four legs are constructed from the vector fields $\vec{A}$ and $\vec{B}$.

The commutator is useful as a tool for distinguishing between coordinate bases and non-coordinate bases (also called non-holonomic bases): In a coordinate basis, the basis vectors are just the coordinate system's partial derivatives, $\vec{e}_\alpha = \partial/\partial x^\alpha$, and since partial derivatives commute, it must be that $[\vec{e}_\alpha, \vec{e}_\beta] = 0$. Conversely (as Fig. 23.5 explains), if one has a basis with vanishing commutators $[\vec{e}_\alpha, \vec{e}_\beta] = 0$, then it is possible to construct a coordinate system for which this is the coordinate basis. In a non-coordinate basis, at least one of the commutators $[\vec{e}_\alpha, \vec{e}_\beta]$ will be nonzero.

23.3.3 Differentiation of Vectors and Tensors; Connection Coefficients

In a curved manifold, the differentiation of vectors and tensors is rather subtle. To elucidate the problem, let us recall how we defined such differentiation in Minkowski spacetime or

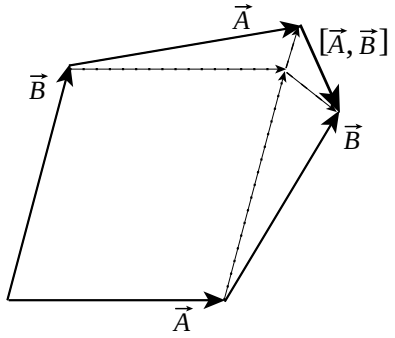


Fig. 23.5: The commutator $[\vec{A},\vec{B}]$ of two vector fields. In this diagram the vectors are assumed to be so small that the curvature of the manifold is negligible in the region of the diagram, so all the vectors can be drawn lying in the surface itself rather than in their respective tangent spaces. In evaluating the two terms in the commutator (23.26), a locally orthonormal coordinate basis is used, so $A^\alpha \partial B^\beta/\partial x^\alpha$ is the amount by which the vector $\vec{B}$ changes when one travels along $\vec{A}$ (i.e. it is the short dashed curve in the upper right), and $B^\alpha \partial A^\beta/\partial x^\alpha$ is the amount by which $\vec{A}$ changes when one travels along $\vec{B}$ (i.e. it is the other short dashed curve). According to Eq. (23.26), the difference of these two short-dashed curves is the commutator $[\vec{A},\vec{B}]$. As the diagram shows, this commutator closes the quadrilateral whose legs are $\vec{A}$ and $\vec{B}$. If the commutator vanishes, then there is no gap in the quadrilateral, which means that in the region covered by this diagram one can construct a coordinate system in which $\vec{A}$ and $\vec{B}$ are coordinate basis vectors.

Euclidean space (Sec. 1.9). Converting to the above notation, we began by defining the directional derivative of a tensor field $\mathbf{F}(\mathcal{P})$ along the tangent vector $\vec{A} = d/d\zeta$ to a curve $\mathcal{P}(\zeta)$:

$$\boxed{\nabla_{\vec{A}}\mathbf{F} \equiv \lim_{\Delta\zeta\to 0} \frac{\mathbf{F}[\mathcal{P}(\Delta\zeta)] - \mathbf{F}[\mathcal{P}(0)]}{\Delta\zeta}} \ . \tag{23.27}$$

This definition is problematic because $\mathbf{F}[\mathcal{P}(\Delta\zeta))]$ lives in a different tangent space than $\mathbf{F}[\mathcal{P}(0)]$. To make the definition meaningful, we must identify some *connection* between the two tangent spaces, when their points $\mathcal{P}(\Delta\zeta)$ and $\mathcal{P}(0)$ are arbitrarily close together. That connection is equivalent to identifying a rule for transporting $\mathbf{F}$ from one tangent space to the other.

In flat space or flat spacetime, and when $\mathbf{F}$ is a vector $\vec{F}$, that transport rule is obvious: keep $\vec{F}$ parallel to itself and keep its length fixed during the transport; in other words, keep constant its components in an orthonormal coordinate system (Cartesian coordinates in Euclidean space, Lorentz coordinates in Minkowski spacetime). This is called the *law of parallel transport*. For a tensor $\mathbf{F}$ the parallel transport law is the same: keep its components fixed in an orthonormal coordinate basis.

In curved spacetime there is no such thing as an orthonormal coordinate basis. Just as the curvature of the earth's surface prevents one from placing a Cartesian coordinate system on it, so the spacetime curvature prevents one from introducing Lorentz coordinates; see Chap. 24. However, in an arbitrarily small region on the earth's surface one can introduce coordinates that are arbitrarily close to Cartesian (as surveyors well know); the fractional deviations from Cartesian need be no larger than $\mathcal{O}(L^2/R^2)$, where L is the size of the

region and R is the earth's radius (see Sec.24.3). Correspondingly, in curved spacetime, in an arbitrarily small region one can introduce coordinates that are arbitrarily close to Lorentz, differing only by amounts quadratic in the size of the region. Such coordinates are sufficiently like their flat space counterparts that they can be used to define parallel transport in the curved manifolds: In Eq. (23.27) one must transport $\mathbf{F}$ from $\mathcal{P}(\Delta\zeta)$ to $\mathcal{P}(0)$, holding its components fixed in a locally orthonormal coordinate basis (parallel transport), and then take the difference in the tangent space at $\mathcal{P}_o = \mathcal{P}(0)$, divide by $\Delta\zeta$, and let $\Delta\zeta \to 0$. The result is a tensor at $\mathcal{P}_o$: the directional derivative $\nabla_{\vec{A}}\mathbf{F}$ of $\mathbf{F}$.

Having made the directional derivative meaningful, one can proceed as in Sec. 1.9, and define the gradient of $\mathbf{F}$ by $\nabla_{\vec{A}}\mathbf{F} = \vec{\nabla}\mathbf{F}(\_, \_, \vec{A})$ [i.e., put $\vec{A}$ in the last, differentiation, slot of $\vec{\nabla}\mathbf{F}$; Eq. (1.54b)].

As in Chap. 1, in any basis we denote the components of $\vec{\nabla}\mathbf{F}$ by $F_{\alpha\beta;\gamma}$; and as in Sec. 10.3 (elasticity theory), we can compute these components in any basis with the aid of that basis's *connection coefficients* (also called *Christoffel symbols*).

In Sec. 10.3 we restricted ourselves to an orthonormal basis in Euclidean space and thus had no need to distinguish between covariant and contravariant indices; all indices were written as subscripts. Now, with non-othonormal bases and in spacetime, we must distinguish covariant and contravariant indices. Accordingly, by analogy with Eq. (10.19), we define the connection coefficients $\Gamma^{\mu}{}_{\alpha\beta}$ as

$$\boxed{\nabla_\beta \vec{e}_\alpha = \Gamma^{\mu}{}_{\alpha\beta}\vec{e}_\mu \ ,} \tag{23.28}$$

where $\nabla_\beta \equiv \nabla_{\vec{e}_\beta}$. The duality between bases $\vec{e}^{\nu} \cdot \vec{e}_\alpha = \delta^\nu_\alpha$ then implies

$$\nabla_\beta \vec{e}^{\mu} = -\Gamma^{\mu}{}_{\alpha\beta}\vec{e}^{\alpha} \ . \tag{23.29}$$

Note the sign flip, which is required to keep $\nabla_\beta(\vec{e}^{\mu} \cdot \vec{e}_\alpha) = 0$, and note that the differentiation index always goes last. Duality also implies that Eqs. (23.28) and (23.29) can be rewritten as

$$\Gamma^{\mu}{}_{\alpha\beta} = \vec{e}^{\mu} \cdot \nabla_\beta \vec{e}_\alpha = -\vec{e}_\alpha \nabla_\beta \vec{e}^{\mu} \ . \tag{23.30}$$

With the aid of these connection coefficients, we can evaluate the components $A_{\alpha;\beta}$ of the gradient of a vector field in any basis. We just compute

$$\begin{aligned} A^{\mu}{}_{;\beta}\vec{e}_\mu &= \nabla_\beta \vec{A} = \nabla_\beta(\vec{A}^\mu \vec{e}_\mu) = (\nabla_\beta A^\mu)\vec{e}_\mu + \vec{A}^\mu \nabla_\beta \vec{e}_\mu \\ &= A^{\mu}{}_{,\beta}\vec{e}_\mu + A^\mu \Gamma^{\alpha}{}_{\mu\beta}\vec{e}_\alpha \\ &= (A^{\mu}{}_{,\beta} + A^\alpha \Gamma^{\mu}{}_{\alpha\beta})\vec{e}_\mu \ . \end{aligned} \tag{23.31}$$

In going from the first line to the second, we have used the notation

$$A^{\mu}{}_{,\beta} \equiv \partial_{\vec{e}_\beta} A^\mu \ ; \tag{23.32}$$

i.e. the comma denotes the result of letting a basis vector act as a differential operator on the component of the vector. In going from the second line of (23.31) to the third, we have renamed the summed-over index α μ and renamed μ α. By comparing the first and last expressions in Eq. (23.31), we conclude that

$$\boxed{A^{\mu}{}_{;\beta} = A^{\mu}{}_{,\beta} + A^\alpha \Gamma^{\mu}{}_{\alpha\beta} \ .} \tag{23.33}$$

The first term in this equation dscribes the changes in $\vec{A}$ associated with changes of its components; the second term corrects for artificial changes of components that are induced by turning and length changes of the basis vectors.

By a similar computation, we conclude that in any basis the covariant components of the gradient are

$$\boxed{A_{\alpha;\beta} = A_{\alpha,\beta} - \Gamma^{\mu}{}_{\alpha\beta} A_{\mu} \ ,} \tag{23.34}$$

where again $A_{\alpha,\beta} \equiv \partial_\beta A_\alpha$. Notice that when the index being "corrected" is down [Eq. (23.34)], the connection coefficient has a minus sign; when it is up [Eq. (23.33)], the connection coefficient has a plus sign. This is in accord with the signs in Eqs. (23.29)–(23.30).

These considerations should make obvious the following equations for the components of the gradient of a tensor:

$$F^{\alpha\beta}{}_{;\gamma} = F^{\alpha\beta}{}_{,\gamma} + \Gamma^{\alpha}{}_{\mu\gamma} F^{\mu\beta} + \Gamma^{\beta}{}_{\mu\gamma} F^{\alpha\mu} \ , \quad F_{\alpha\beta;\gamma} = F_{\alpha\beta,\gamma} - \Gamma^{\mu}{}_{\alpha\gamma} F_{\mu\beta} - \Gamma^{\mu}{}_{\beta\gamma} F_{\alpha\mu} \ . \tag{23.35}$$

Notice that each index of $\mathbf{F}$ must be corrected, the correction has a sign dictated by whether the index is up or down, the differentiation index always goes last on the Γ, and all other indices can be deduced by requiring that the free indices in each term be the same and all other indices be summed.

If we have been given a basis, then how can we compute the connection coefficients? We can try to do so by drawing pictures and examining how the basis vectors change from point to point—a method that is fruitful in spherical and cylindrical coordinates in Euclidean space (Sec. 10.3). However, in other situations this method is fraught with peril, so we need a firm mathematical prescription. It turns out that the following prescription works; see below for a proof:

(*i*) Evaluate the *commutation coefficients* $c_{\alpha\beta}{}^{\rho}$ of the basis, which are defined by the two equivalent relations

$$\boxed{[\vec{e}_\alpha, \vec{e}_\beta] \equiv c_{\alpha\beta}{}^{\rho} \vec{e}_\rho \quad , \qquad c_{\alpha\beta}{}^{\rho} \equiv \vec{e}^{\rho} \cdot [\vec{e}_\alpha, \vec{e}_\beta] \ .} \tag{23.36}$$

[Note that in a coordinate basis the commutation coefficients will vanish. *Warning*: commutation coefficients also appear in the theory of Lie Groups; there it is conventional to use a different ordering of indices than here, $c_{\alpha\beta}{}^{\rho}{}_{\rm here} = c^{\rho}_{\alpha\beta_{\rm Lie\ groups}}$.] (*ii*) Lower the last index on the commutation coefficients using the metric components in the basis:

$$\boxed{c_{\alpha\beta\gamma} \equiv c_{\alpha\beta}{}^{\rho} g_{\rho\gamma} \ .} \tag{23.37}$$

(*iii*) Compute the *covariant Christoffel symbols*

$$\boxed{\Gamma_{\alpha\beta\gamma} \equiv \frac{1}{2}(g_{\alpha\beta,\gamma} + g_{\alpha\gamma,\beta} - g_{\beta\gamma,\alpha} + c_{\alpha\beta\gamma} + c_{\alpha\gamma\beta} - c_{\beta\gamma\alpha}) \ .} \tag{23.38}$$

Here the commas denote differentiation with respect to the basis vectors as though the connection coefficients were scalar fields [Eq. (23.32)]. Notice that the pattern of indices is the same on the g's and on the c's. It is a peculiar pattern—one of the few aspects of index gymnastics that cannot be reconstructed by merely lining up indices. In a coordinate basis the c's will vanish and $\Gamma_{\alpha\beta\gamma}$ will be symmetric in its last two indices; in an orthonormal basis

$g_{\mu\nu}$ are constant so the g's will vanish and $\Gamma_{\alpha\beta\gamma}$ will be antisymmetric in its first two indices; and in a Cartesian or Lorentz coordinate basis, which is both coordinate and orthonormal, both the c's and the g's will vanish, so $\Gamma_{\alpha\beta\gamma}$ will vanish. (*iv*) Raise the first index on the covariant Christoffel symbols to obtain the connection coefficients, which are also sometimes called the *mixed Christoffel symbols*

$$\boxed{\Gamma^{\mu}{}_{\beta\gamma} = g^{\mu\alpha}\Gamma_{\alpha\beta\gamma} \ .} \tag{23.39}$$

The gradient operator $\vec{\nabla}$ is an example of a geometric object that is not a tensor. The connection coefficients can be regarded as the components of $\vec{\nabla}$; and because $\vec{\nabla}$ is not a tensor, these components $\Gamma^{\alpha}{}_{\beta\gamma}$ do not obey the tensorial transformation law (23.19) when switching from one basis to another. Their transformation law is far more complicated and is very rarely used. Normally one computes them from scratch in the new basis, using the above prescription or some other, equivalent prescription (cf. Chap. 14 of MTW). For most curved spacetimes that one meets in general relativity, these computations are long and tedious and therefore are normally carried out on computers using symbolic manipulations software such as Macsyma, or GRTensor (running under Maple or Mathematica), or Mathtensor (under Mathematica). Such software is easily found on the Internet using a search engine.

The above prescription for computing the connection coefficients follows from two key properties of the gradient $\vec{\nabla}$: *First*, The gradient of the metric tensor vanishes,

$$\boxed{\vec{\nabla}\mathbf{g} = 0 \ .} \tag{23.40}$$

This can be seen by introducing a locally orthornormal coordinate basis at the arbitrary point $\mathcal{P}$ where the gradient is to be evaluated. In such a basis, the effects of curvature show up only at quadratic order in distance away from $\mathcal{P}$, which means that the coordinate bases $\vec{e}_\alpha \equiv \partial/\partial x^\alpha$ behave, at first order in distance, just like those of an orthonormal coordinate system in flat space. Since $\nabla_\beta \vec{e}_\alpha$ involves only first derivatives and it vanishes in an orthonormal coordinate system in flat space, it must also vanish here—which means that the connection coefficients vanish at $\mathcal{P}$ in this basis. Therefore, the components of $\vec{\nabla}\mathbf{g}$ at $\mathcal{P}$ are $g_{\alpha\beta;\gamma} = g_{\alpha\beta,\gamma} = \partial g_{\alpha\beta}/\partial x^\gamma$, which vanishes since the components of $\mathbf{g}$ in this basis are all 0 or ± 1 plus corrections second order in distance from $\mathcal{P}$. This vanishing of the components of $\vec{\nabla}\mathbf{g}$ in our special basis guarantees that $\vec{\nabla}\mathbf{g}$ itself vanishes at $\mathcal{P}$; and since $\mathcal{P}$ was an arbitrary point, $\vec{\nabla}\mathbf{g}$ must vanish everywhere and always.

Second, for any two vector fields $\vec{A}$ and $\vec{B}$, the gradient is related to the commutator by

$$\boxed{\nabla_{\vec{A}}\vec{B} - \nabla_{\vec{B}}\vec{A} = [\vec{A}, \vec{B}] \ .} \tag{23.41}$$

This relation, like $\vec{\nabla}\mathbf{g} = 0$, is most easily derived by introducing a locally orthonormal coordinate basis at the point $\mathcal{P}$ where one wishes to check its validity. Since $\Gamma^{\mu}{}_{\alpha\beta} = 0$ at $\mathcal{P}$ in that basis, the components of $\nabla_{\vec{A}}\vec{B} - \nabla_{\vec{B}}\vec{A}$ are $B^{\alpha}{}_{;\beta}A^{\beta} - A^{\alpha}{}_{;\beta}B^{\beta} = B^{\alpha}{}_{,\beta}A^{\beta} - A^{\alpha}{}_{,\beta}B^{\beta}$ [cf. Eq. (23.33)]. But these components are identical to those of the commutator $[\vec{A}, \vec{B}]$ [Eq. (23.26)]. Since the components of these two vectors [the left and right sides of (23.41)] are identical at $\mathcal{P}$ in this special basis, the vectors must be identical, and since the point $\mathcal{P}$ was arbitrary, they must always be identical.

Turn, now, to the derivation of our prescription for computing the connection coefficients in an arbitrary basis. By virtue of the relation $\Gamma^{\mu}{}_{\beta\gamma} = g^{\mu\alpha}\Gamma_{\alpha\beta\gamma}$ [Eq. (23.39)] and its inverse

$$\Gamma_{\alpha\beta\gamma} = g_{\alpha\mu}\Gamma^{\mu}{}_{\beta\gamma} \ , \tag{23.42}$$

a knowledge of $\Gamma_{\alpha\beta\gamma}$ is equivalent to a knowledge of $\Gamma^{\mu}{}_{\beta\gamma}$. Thus, our task reduces to deriving expression (23.38) for $\Gamma_{\alpha\beta\gamma}$, in which the $c_{\alpha\beta\gamma}$ are defined by equations (23.36) and (23.37). As a first step in the derivation, notice that the constancy of the metric tensor, $\vec{\nabla}\mathbf{g} = 0$, when expressed in component notation using Eq. (23.35), and when combined with Eq. (23.42), becomes $0 = g_{\alpha\beta;\gamma} = g_{\alpha\beta,\gamma} - \Gamma_{\beta\alpha\gamma} - \Gamma_{\alpha\beta\gamma}$; i.e.,

$$\Gamma_{\alpha\beta\gamma} + \Gamma_{\beta\alpha\gamma} = g_{\alpha\beta,\gamma} \ . \tag{23.43}$$

This determines the part of $\Gamma_{\alpha\beta\gamma}$ that is symmetric in the first two indices. The commutator of the basis vectors determines the part antisymmetric in the last two indices: From

$$c_{\alpha\beta}{}^{\mu}\vec{e}_{\mu} = [\vec{e}_{\alpha},\ \vec{e}_{\beta}] = \nabla_{\alpha}\vec{e}_{\beta} - \nabla_{\beta}\vec{e}_{\alpha} = (\Gamma^{\mu}{}_{\beta\alpha} - \Gamma^{\mu}{}_{\alpha\beta})\vec{e}_{\mu} \tag{23.44}$$

(where the first equality is the definition (23.36) of the commutation coefficient, the second is expression (23.41) for the commutator in terms of the gradient, and the third follows from the definition (23.28) of the connection coefficient), we infer, by equating the components and lowering the μ index, that

$$\Gamma_{\gamma\beta\alpha} - \Gamma_{\gamma\alpha\beta} = c_{\alpha\beta\gamma} \ . \tag{23.45}$$

By combining equations (23.43) and (23.45) and performing some rather tricky algebra (cf. Ex. 8.15 of MTW), we obtain the computational rule (23.38).

23.3.4 Integration

Our desire to use general bases and work in curved space gives rise to two new issues in the definition of integrals.

First, the volume elements used in integration involve the Levi-Civita tensor [Eqs. (1.59b), (1.73), (1.76)], so we need to know the components of the Levi-Civita tensor in a general basis. It turns out [see, e.g., Ex. 8.3 of MTW] that the covariant components differ from those in an orthonormal basis by a factor $\sqrt{|g|}$ and the contravariant by $1/\sqrt{|g|}$, where

$$\boxed{g \equiv \det ||g_{\alpha\beta}||} \tag{23.46}$$

is the determinant of the matrix whose entries are the covariant components of the metric. More specifically, let us denote by $[\alpha\beta\ldots\nu]$ the value of $\epsilon_{\alpha\beta\ldots\nu}$ in an orthonormal basis of our n-dimensional space [Eq. (1.59b)]:

$$\begin{aligned}
[12\ldots N] &= +1 \ , \\
[\alpha\beta\ldots\nu] &= +1 \quad \text{if } \alpha, \beta, \ldots, \nu \text{ is an even permutation of } 1, 2, \ldots, N \\
&= -1 \quad \text{if } \alpha, \beta, \ldots, \nu \text{ is an odd permutation of } 1, 2, \ldots, N \\
&= 0 \quad \text{if } \alpha, \beta, \ldots, \nu \text{ are not all different.}
\end{aligned} \tag{23.47}$$

(In spacetime the indices must run from 0 to 3 rather than 1 to $n = 4$). Then in a general right-handed basis the components of the Levi-Civita tensor are

$$\boxed{\epsilon_{\alpha\beta\ldots\nu} = \sqrt{|g|}\,[\alpha\beta\ldots\nu]\ , \quad \epsilon^{\alpha\beta\ldots\nu} = \pm\frac{1}{\sqrt{|g|}}\,[\alpha\beta\ldots\nu]\ ,} \tag{23.48}$$

where the $\pm$ is plus in Euclidean space and minus in spacetime. In a left-handed basis the sign is reversed.

As an example of these formulas, consider a spherical polar coordinate system (r, θ, ϕ) in three-dimensional Euclidean space, and use the three infinitesimal vectors $dx^j(\partial/\partial x^j)$ to construct the volume element $d\Sigma$ [cf. Eq. (1.69b)]:

$$d\Sigma = \boldsymbol{\epsilon}\left(dr\frac{\partial}{\partial r}, d\theta\frac{\partial}{\partial\theta}, d\phi\frac{\partial}{\partial\phi}\right) = \epsilon_{r\theta\phi}drd\theta d\phi = \sqrt{g}drd\theta d\phi = r^2\sin\theta drd\theta d\phi\ . \tag{23.49}$$

Here the second equality follows from linearity of $\boldsymbol{\epsilon}$ and the formula for computing its components by inserting basis vectors into its slots; the third equality follows from our formula (23.48) for the components, and the fourth equality entails the determinant of the metric coefficients, which in spherical coordinates are $g_{rr} = 1$, $g_{\theta\theta} = r^2$, $g_{\phi\phi} = r^2\sin^2\theta$, all other g_{jk} vanish, so $g = r^4\sin^2\theta$. The resulting volume element $r^2\sin\theta drd\theta d\phi$ should be familiar and obvious.

The *second* new integration issue that we must face is the fact that integrals such as

$$\int_{\partial\mathcal{V}} T^{\alpha\beta}d\Sigma_\beta \tag{23.50}$$

[cf. Eqs. (1.76), (1.77)] involve constructing a vector $T^{\alpha\beta}d\Sigma_\beta$ in each infinitesimal region $d\Sigma_\beta$ of the surface of integration, and then adding up the contributions from all the infinitesimal regions. A major difficulty arises from the fact that each contribution lives in a different tangent space. To add them together, we must first transport them all to the same tangent space at some single location in the manifold. How is that transport to be performed? The obvious answer is "by the same parallel transport technique that we used in defining the gradient." However, when defining the gradient we only needed to perform the parallel transport over an infinitesimal distance, and now we must perform it over long distances. As we shall see in Chap. 24, when the manifold is curved, long-distance parallel transport gives a result that depends on the route of the transport, and in general there is no way to identify any preferred route. As a result, integrals such as (23.50) are ill-defined in a curved manifold. The only integrals that are well defined in a curved manifold are those such as $\int_{\partial\mathcal{V}} S^\alpha d\Sigma_\alpha$ whose infinitesimal contributions $S^\alpha d\Sigma_\alpha$ are scalars, i.e. integrals whose value is a scalar. This fact will have profound consequences in curved spacetime for the laws of energy, momentum, and angular momentum conservation.

****

EXERCISES

Exercise 23.3 *Problem: Practice with Frame-Independent Tensors*
Let $\mathbf{A}, \mathbf{B}$ be second rank tensors.

(a) Show that $\mathbf{A} + \mathbf{B}$ is also a second rank tensor.

(b) Show that $\mathbf{A} \otimes \mathbf{B}$ is a fourth rank tensor.

(c) Show that the contraction of $\mathbf{A} \otimes \mathbf{B}$ on its first and fourth slots is a second rank tensor. (If necessary, consult Chap. 1 for a discussion of contraction).

(d) Write the following quantitites in slot-naming index notation: the tensor $\mathbf{A} \otimes \mathbf{B}$; the simultaneous contraction of this tensor on its first and fourth slots and on its second and third slots.

Exercise 23.4 *Derivation: Index Manipulation Rules from Duality*
For an arbitrary basis $\{\vec{e}_\alpha\}$ and its dual basis $\{\vec{e}^{\bar{\mu}}\}$, use (i) the duality relation (23.8), the definition (23.19) of components of a tensor and the relation $\vec{A} \cdot \vec{B} = \mathbf{g}(\vec{A}, \vec{B})$ between the metric and the inner product to deduce the following results:

(a) The relations

$$\vec{e}^{\mu} = g^{\mu\alpha}\vec{e}_\alpha \ , \quad \vec{e}_\alpha = g_{\alpha\mu}\vec{e}^{\mu} \ . \tag{23.51}$$

(b) The fact that indices on the components of tensors can be raised and lowered using the components of the metric, e.g.

$$F^{\mu\nu} = g^{\mu\alpha}F_\alpha{}^\nu \ , \quad p_\alpha = g_{\alpha\beta}p^\beta \ . \tag{23.52}$$

(c) The fact that a tensor can be reconstructed from its components in the manner of Eq. (23.11).

Exercise 23.5 *Practice: Transformation Matrices for Circular Polar Bases*
Consider the circular coordinate system $\{\varpi, \phi\}$ and its coordinate bases and orthonormal bases as discussed in Fig. 23.3 and the associated text. These coordinates are related to Cartesian coordinates $\{x, y\}$ by the usual relations $x = \varpi\cos\phi$, $y = \varpi\sin\phi$.

(a) Evaluate the components ($L^x{}_\varpi$ etc.) of the transformation matrix that links the two coordinate bases $\{\vec{e}_x, \vec{e}_y\}$ and $\{\vec{e}_\varpi, \vec{e}_\phi\}$. Also evaluate the components ($L^\varpi{}_x$ etc.) of the inverse transformation matrix.

(b) Evaluate, similarly, the components of the transformation matrix and its inverse linking the bases $\{\vec{e}_x, \vec{e}_y\}$ and $\{\vec{e}_{\hat{\varpi}}, \vec{e}_{\hat{\phi}}\}$.

(c) Consider the vector $\vec{A} \equiv \vec{e}_x + 2\vec{e}_y$. What are its components in the other two bases?

Exercise 23.6 *Practice: Commutation and Connection Coefficients for Circular Polar Bases*
As in the previous exercise, consider the circular coordinates $\{\varpi, \phi\}$ of Fig. 23.3 and their associated bases.

(a) Evaluate the commutation coefficients $c_{\alpha\beta}{}^{\rho}$ for the coordinate basis $\{\vec{e}_{\varpi}, \vec{e}_{\phi}\}$, and also for the orthonormal basis $\{\vec{e}_{\hat{\varpi}}, \vec{e}_{\hat{\phi}}\}$.

(b) Compute by hand the connection coefficients for the coordinate basis and also for the orthonormal basis, using Eqs. (23.36)–(23.39). [Note: the answer for the orthonormal basis was worked out by a different method in our study of elasticity theory; Eq. (10.21).]

(c) Repeat this computation using symbolic manipulation software on a computer.

Exercise 23.7 *Practice: Connection Coefficients for Spherical Polar Coordinates*

(a) Consider spherical polar coordinates in 3-dimensional space and verify that the non-zero connection coefficients assuming an orthonormal basis are given by Eq. (10.22).

(b) Repeat the exercise assuming a coordinate basis with

$$\mathbf{e}_r \equiv \frac{\partial}{\partial r}, \qquad \mathbf{e}_\theta \equiv \frac{\partial}{\partial \theta}, \qquad \mathbf{e}_\phi \equiv \frac{\partial}{\partial \phi} \; . \tag{23.53}$$

(c) Repeat both computations using symbolic manipulation software on a computer.

Exercise 23.8 *Practice: Index Gymnastics — Geometric Optics*
In the geometric optics approximation (Chap. 6), for electromagnetic waves in Lorenz gauge, one can write the 4-vector potential in the form $\vec{A} = \vec{\mathcal{A}} e^{i\varphi}$, where $\vec{\mathcal{A}}$ is a slowly varying amplitude and φ is a rapidly varying phase. By the techniques of Chap. 6, one can deduce that the wave vector, defined by $\vec{k} \equiv \boldsymbol{\nabla}\varphi$, is null: $\vec{k} \cdot \vec{k} = 0$.

(a) Rewrite all of the equations in the above paragraph in slot-naming index notation.

(b) Using index manipulations, show that the wave vector $\vec{k}$ (which is a vector field because the wave's phase φ is a vector field) satisfies the geodesic equation, $\nabla_{\vec{k}}\vec{k} = 0$. The geodesics, to which $\vec{k}$ is the tangent vector, are the rays discussed in Chap. 6, along which the waves propagate.

Exercise 23.9 *Practice: Index Gymnastics — Irreducible Tensorial Parts of the Gradient of a 4-Velocity Field*
In our study of elasticity theory, we introduced the concept of the irreducible tensorial parts of a second-rank tensor in Euclidean space (Box. 10.2). Consider a fluid flowing through spacetime, with a 4-velocity $\vec{u}(\mathcal{P})$. The fluid's gradient $\boldsymbol{\nabla}\vec{u}$ ($u_{\alpha;\beta}$ in slot-naming index notation) is a second-rank tensor in spacetime. With the aid of the 4-velocity itself, we can break it down into irreducible tensorial parts as follows:

$$u_{\alpha;\beta} = -a_\alpha u_\beta + \frac{1}{3}\theta P_{\alpha\beta} + \sigma_{\alpha\beta} + \omega_{\alpha\beta} \; . \tag{23.54}$$

Here: (i) $P_{\alpha\beta}$ is defined by

$$P_{\alpha\beta} \equiv g_{\alpha\beta} + u_\alpha u_\beta \; , \tag{23.55}$$

(ii) $\sigma_{\alpha\beta}$ is symmetric and trace-free and is orthogonal to the 4-velocity, and (iii) $\omega_{\alpha\beta}$ is antisymmetric and is orthogonal to the 4-velocity.

(a) In quantum mechanics one deals with "projection operators" $\hat{P}$, which satisfy the equation $\hat{P}^2 = \hat{P}$. Show that $P_{\alpha\beta}$ is a projection tensor, in the sense that $P_{\alpha\beta}P^{\beta}{}_{\gamma} = P_{\alpha\gamma}$.

(b) This suggests that $P_{\alpha\beta}$ may project vectors into some subspace of 4-dimensional spacetime. Indeed it does: Show that for any vector A^α, $P_{\alpha\beta}A^\beta$ is orthogonal to $\vec{u}$; and if A^α is already perpendicular to $\vec{u}$, then $P_{\alpha\beta}A^\beta = A_\alpha$, i.e. the projection leaves the vector unchanged. Thus, $P_{\alpha\beta}$ projects vectors into the 3-space orthogonal to $\vec{u}$.

(c) What are the components of $P_{\alpha\beta}$ in the fluid's local rest frame, i.e. in an orthonormal basis where $\vec{u} = \vec{e}_{\hat{0}}$?

(d) Show that the rate of change of $\vec{u}$ along itself, $\nabla_{\vec{u}}\vec{u}$ (i.e., the fluid 4-acceleration) is equal to the vector $\vec{a}$ that appears in the decomposition (23.54). Show, further, that $\vec{a}\cdot\vec{u} = 0$.

(e) Show that the divergence of the 4-velocity, $\boldsymbol{\nabla}\cdot\vec{u}$, is equal to the scalar field θ that appears in the decomposition (23.54).

(f) The quantities $\sigma_{\alpha\beta}$ and $\omega_{\alpha\beta}$ are the relativistic versions of the fluid's shear and rotation tensors. Derive equations for these tensors in terms of $u_{\alpha;\beta}$ and $P_{\mu\nu}$.

(g) Show that, as viewed in a Lorentz reference frame where the fluid is moving with speed small compared to the speed of light, to first-order in the fluid's ordinary velocity $v^j = dx^j/dt$, the following are true: (i) $u^0 = 1$, $u^j = v^j$; (ii) θ is the nonrelativistic expansion of the fluid, $\theta = \boldsymbol{\nabla}\cdot\mathbf{v} \equiv v^j{}_{,j}$ [Eq. (12.63)]; (iii) σ_{jk} is the fluid's nonrelativistic shear [Eq. (12.63)]; (iv) ω_{jk} is the fluid's nonrelativist rotation tensor [denoted r_{jk} in Eq. (12.63)].

Exercise 23.10 *Practice: Integration — Gauss's Theorem*
In 3-dimensional Euclidean space the Maxwell equation $\boldsymbol{\nabla}\cdot\mathbf{E} = \rho_e/\epsilon_0$ can be combined with Gauss's theorem to show that the electric flux through the surface $\partial\mathcal{V}$ of a sphere is equal to the charge in the sphere's interior $\mathcal{V}$ divided by ϵ_0:

$$\int_{\partial\mathcal{V}} \mathbf{E}\cdot d\boldsymbol{\Sigma} = \int_{\mathcal{V}} (\rho_e/\epsilon_0) d\Sigma \ . \tag{23.56}$$

Introduce spherical polar coordinates so the sphere's surface is at some radius $r = R$. Consider a surface element on the sphere's surface with vectorial legs $d\phi\partial/\partial\phi$ and $d\theta\partial/\partial\theta$. Evaluate the components $d\Sigma_j$ of the surface integration element $d\boldsymbol{\Sigma} = \epsilon(..., d\theta\partial/\partial\theta, d\phi\partial/\partial\phi)$. Similarly, evaluate $d\Sigma$ in terms of vectorial legs in the sphere's interior. Then use these results for $d\Sigma_j$ and $d\Sigma$ to convert Eq. (23.56) into an explicit form in terms of integrals over r, θ, ϕ. The final answer should be obvious, but the above steps in deriving it are informative.

****

23.4 The Stress-Energy Tensor Revisited

In Sec. 1.12 we defined the *stress-energy tensor* $\mathbf{T}$ of any matter or field as a symmetric, second-rank tensor that describes the flow of 4-momentum through spacetime. More specifically, the total 4-momentum $\mathbf{P}$ that flows through some small 3-volume $\mathbf{\Sigma}$, going from the negative side of $\mathbf{\Sigma}$ to its positive side, is

$$\boxed{\mathbf{T}(\ldots,\vec{\Sigma}) = (\text{total 4-momentum } \vec{P} \text{ that flows through } \vec{\Sigma}); \quad \text{i.e., } T^{\alpha\beta}\Sigma_\beta = P^\alpha} \tag{23.57}$$

[Eq. (1.91)]. Of course, this stress-energy tensor depends on the location $\mathcal{P}$ of the 3-volume in spacetime; i.e., it is a tensor field $\mathbf{T}(\mathcal{P})$.

From this geometric, frame-independent definition of the stress-energy tensor, we were able to read off the physical meaning of its components in any inertial reference frame [Eqs. (1.92)]: T^{00} is the total energy density, including rest mass-energy; $T^{j0} = T^{0j}$ is the j-component of momentum density, or equivalently the j-component of energy flux; and T^{jk} are the components of the stress tensor, or equivalently of the momentum flux.

We gained some insight into the stress-energy tensor in the context of kinetic theory in Secs. 2.4.2 and 2.5.3, and we briefly introduced the stress-energy tensor for a perfect fluid in Eq. (1.99b). Because perfect fluids will play a very important role in this book's applications of general relativity to relativistic stars (Chap. 25) and cosmology (Chap. 27), we shall now explore the perfect-fluid stress-energy tensor in some depth, and shall see how it is related to the Newtonian description of perfect fluids, which we studied in Part IV.

Recall [Eq. (1.99a)] that in the local rest frame of a perfect fluid, there is no energy flux or momentum density, $T^{j0} = T^{0j} = 0$, but there is a total energy density (including rest mass) ρ and an isotropic pressure P:

$$T^{00} = \rho\ , \quad T^{jk} = P\delta^{jk}\ . \tag{23.58}$$

From this special form of $T^{\alpha\beta}$ in the local rest frame, one can derive Eq. (1.99b) for the stress-energy tensor in terms of the 4-velocity $\vec{u}$ of the local rest frame (i.e., of the fluid itself), the metric tensor of spacetime $\mathbf{g}$, and the rest-frame energy density ρ and pressure P:

$$\boxed{T^{\alpha\beta} = (\rho + P)u^\alpha u^\beta + Pg^{\alpha\beta}\ ; \quad \text{i.e., } \mathbf{T} = (\rho+P)\vec{u}\otimes\vec{u} + P\mathbf{g}\ ;} \tag{23.59}$$

see Ex. 23.11, below. This expression for the stress-energy tensor of a perfect fluid is an example of a geometric, frame-independent description of physics.

It is instructive to evaluate the nonrelativistic limit of this perfect-fluid stress-energy tensor and verify that it has the form we used in our study of nonrelativistic, inviscid fluid mechanics (Table 12.1 on page 24 of Chap. 12, with vanishing gravitational potential $\Phi = 0$). In the nonrelativistic limit the fluid is nearly at rest in the chosen Lorentz reference frame. It moves with ordinary velocity $\mathbf{v} = d\mathbf{x}/dt$ that is small compared to the speed of light, so the temporal part of its 4-velocity $u^0 = 1/\sqrt{1-v^2}$ and spatial part $\mathbf{u} = u^0\mathbf{v}$ can be approximated as

$$u^0 \simeq 1 + \frac{1}{2}v^2\ , \quad \mathbf{u} \simeq \left(1 + \frac{1}{2}v^2\right)\mathbf{v}\ . \tag{23.60}$$

In the fluid's rest frame, in special relativity, it has a rest mass density ρ_o [defined in Eq. (1.83)], an internal energy per unit rest mass u (not to be confused with the 4-velocity), and a total density of mass-energy

$$\rho = \rho_o(1+u) \ . \tag{23.61}$$

Now, in our chosen Lorentz frame the volume of each fluid element is Lorentz contracted by the factor $\sqrt{1-v^2}$ and therefore the rest mass density is increased from ρ_o to $\rho_o/\sqrt{1-v^2} = \rho_o u^0$; and correspondingly the rest-mass flux is $\rho_o u^0 \mathbf{v} = \rho_o \mathbf{u}$ [Eq. 1.83)], and the law of rest-mass conservation is $\partial(\rho_o u^0)/\partial t + \partial(\rho_o u^j)/\partial x^j = 0$, i.e. $\vec{\nabla}\cdot(\rho_o \vec{u}) = 0$. When taking the Newtonian limit, we should identify the Newtonian mass ρ_N with the low-velocity limit of this rest mass density:

$$\rho_N = \rho_o u^0 \simeq \rho_o \left(1 + \frac{1}{2}v^2\right) \ . \tag{23.62}$$

The nonrelativistic limit regards the specific internal energy u, the kinetic energy per unit mass $\frac{1}{2}v^2$, and the ratio of pressure to rest mass density P/ρ_o as of the same order of smallness

$$u \sim \frac{1}{2}v^2 \sim \frac{P}{\rho_o} \ll 1 \ , \tag{23.63}$$

and it expresses the momentum density T^{j0} accurate to first order in $v \equiv |\mathbf{v}|$, the momentum flux (stress) T^{jk} accurate to second order in v, the energy density T^{00} accurate to second order in v, and the energy flux T^{0j} accurate to third order in v. To these accuracies, the perfect-fluid stress-energy tensor (23.59) takes the following form:

$$\begin{aligned} T^{j0} &= \rho_N v^j \ , & T^{jk} &= P g^{jk} + \rho_N v^j v^k \ , \\ T^{00} &= \rho_N + \frac{1}{2}\rho_N v^2 + \rho_N u \ , & T^{0j} &= \rho_N v^j + \left(\frac{1}{2}v^2 + u + \frac{P}{\rho_N}\right)\rho_N v^j \ ; \end{aligned} \tag{23.64}$$

see Ex. 23.11(c). These are precisely the same as the momentum density, momentum flux, energy density, and energy flux that we used in our study of nonrelativistic, inviscid fluid mechanics (Chap. 12), aside from the notational change from there to here $\rho \to \rho_N$, and aside from including the rest mass-energy $\rho_N = \rho_N c^2$ in T_{00} here but not there, and including the rest-mass-energy flux $\rho_N v^j$ in T^{0j} here but not there.

Just as the nonrelativistic equations of fluid mechanics (Euler equation and energy conservation) are derivable by combining the nonrelativistic $T^{\alpha\beta}$ of Eq. (23.64) with the nonrelativistic laws of momentum and energy conservation, so also the relativistic equations of fluid mechanics are derivable by combining the relativistic version (23.59) of $T^{\alpha\beta}$ with the equation of 4-momentum conservation $\vec{\nabla}\cdot\mathbf{T} = 0$. (We shall give such a derivation and shall examine the resulting fluid mechanics equations in the context of general relativity in Chap. 24.) This, together with the fact that the relativistic $\mathbf{T}$ reduces to the nonrelativistic $T^{\alpha\beta}$ in the nonrelativistic limit, guarantees that the special relativistic equations of inviscid fluid mechanics will reduce to the nonrelativistic equations in the nonrelativistic limit.

A second important example of a stress-energy tensor is that for the electromagnetic field. We shall explore it in Ex. 23.13 below.

For a point particle which moves through spacetime along a world line $\mathcal{P}(\zeta)$ (where ζ is the affine parameter such that the particle's 4-momentum is $\vec{p} = d/d\zeta$), the stress-energy

tensor will vanish everywhere except on the world line itself. Correspondingly, $\mathbf{T}$ must be expressed in terms of a Dirac delta function. The relevant delta function is a scalar function of two points in spacetime, $\delta(\mathcal{Q},\mathcal{P})$ with the property that when one integrates over the point $\mathcal{P}$, using the 4-dimensional volume element $d\Sigma$ (which in any inertial frame just reduces to $d\Sigma = dtdxdydz$), one obtains

$$\boxed{\int_{\mathcal{V}} f(\mathcal{P})\delta(\mathcal{Q},\mathcal{P})d\Sigma = f(\mathcal{Q})\ .} \tag{23.65}$$

Here $f(\mathcal{P})$ is an arbitrary scalar field and the region $\mathcal{V}$ of 4-dimensional integration must include the point $\mathcal{Q}$. One can verify that in terms of Lorentz coordinates this delta function can be expressed as

$$\boxed{\delta(\mathcal{Q},\mathcal{P}) = \delta(t_{\mathcal{Q}} - t_{\mathcal{P}})\delta(x_{\mathcal{Q}} - x_{\mathcal{P}})\delta(y_{\mathcal{Q}} - y_{\mathcal{P}})\delta(z_{\mathcal{Q}} - z_{\mathcal{P}})\ ,} \tag{23.66}$$

where the deltas on the right-hand side are ordinary one-dimensional Dirac delta functions.

In terms of the spacetime delta function $\delta(\mathcal{Q},\mathcal{P})$ the stress-energy tensor of a point particle takes the form

$$\boxed{\mathbf{T}(\mathcal{Q}) = \int_{-\infty}^{+\infty} \vec{p}(\zeta)\otimes\vec{p}(\zeta)\delta\left(\mathcal{Q},\mathcal{P}(\zeta)\right)d\zeta\ ,} \tag{23.67}$$

where the integral is along the world line $\mathcal{P}(\zeta)$ of the particle. It is a straightforward but sophisticated exercise [Ex. 23.14] to verify that the integral of this stress-energy tensor over any 3-surface $\mathcal{S}$ that slices through the particle's world line just once, at an event $\mathcal{P}(\zeta_o)$, is equal to the particle's 4-momentum at the intersection point:

$$\int_{\mathcal{S}} T^{\alpha\beta}(\mathcal{Q})d\Sigma_\beta = p^\alpha(\zeta_o)\ . \tag{23.68}$$

****

EXERCISES

Exercise 23.11 *Derivation: Stress-Energy Tensor for a Perfect Fluid*

(a) Derive the frame-independent expression (23.59) for the perfect fluid stress-energy tensor from its rest-frame components (23.58).

(b) Read Ex. 1.28, and work part (b)—i.e., show that for a perfect fluid the inertial mass per unit volume is isotropic and is equal to $(\rho + P)\delta^{ij}$ when thought of as a tensor, or simply $\rho + P$ when thought of as a scalar.

(c) Show that in the nonrelativistic limit the components of the perfect fluid stress-energy tensor (23.59) take on the forms (23.64), and verify that these agree with the densities and fluxes of energy and momentum that are used in nonrelativistic fluid mechanics (e.g., Table 12.1 on page 24 of Chap. 12).

(d) Show that it is the contribution of the pressure P to the relativistic density of inertial mass that causes the term $(P/\rho_N)\rho_N\mathbf{v} = P\mathbf{v}$ to appear in the nonrelativistic energy flux.

Exercise 23.12 *Problem: Electromagnetic Field Tensor*
As we saw in Sec. 1.10, in 4-dimensional spacetime the electromagnetic field is described by a second-rank tensor $\mathbf{F}(\ldots,\ldots)$ which is antisymmetric on its two slots, $F^{\alpha\beta} = -F^{\beta\alpha}$; and the 4-force (rate of change of 4-momentum) that it exerts on a particle with rest mass m, charge q, proper time τ, 4-velocity $\vec{u} = d/d\tau$, and 4-momentum $\vec{p}$ is

$$\frac{d\vec{p}}{d\tau} = \nabla_{\vec{u}}\vec{p} = q\mathbf{F}(\ldots,\vec{u}) \ ; \quad \text{i.e.,} \ \frac{dp^\alpha}{d\tau} = qF^{\alpha\beta}u_\beta \ . \tag{23.69}$$

Here the second form of the equation, valid in a Lorentz frame, follows from the component form of $\nabla_{\vec{u}}\vec{p}$: $p^\alpha{}_{;\mu}u^\mu = p^\alpha{}_{,\mu}u^\mu = dp^\alpha/d\tau$.

(a) By comparing this with the Lorentz force law for a low-velocity particle, $d\mathbf{p}/dt = q(\mathbf{E} + \mathbf{v}\times\mathbf{B})$, show that the components of the electromagnetic field tensor in a Lorentz reference frame are

$$\boxed{||\, F^{\alpha\beta}\,|| = \left|\left|\begin{array}{cccc} 0 & E^x & E^y & E^z \\ -E^x & 0 & B^z & -B^y \\ -E^y & -B^z & 0 & B^x \\ -E^z & B^y & -B^x & 0 \end{array}\right|\right| \ ;} \tag{23.70}$$

$$\text{i.e.,} \ F^{0i} = -F^{i0} = E^i \ , \quad F^{ij} = -F^{ji} = \epsilon^{ij}{}_k B^k \ , \tag{23.71}$$

where E^j and B^j are the components of the 3-vector electric and magnetic fields that reside in the 3-space of the Lorentz frame.

(b) Define $*\mathbf{F} \equiv$("dual" of $\mathbf{F}$) by

$$\boxed{*F_{\mu\nu} = \frac{1}{2}\epsilon_{\mu\nu\alpha\beta}F^{\alpha\beta} \ ,} \tag{23.72}$$

where $\boldsymbol{\epsilon}$ is the Levi-Civita tensor (Sec. 23.3.4). What are the components of $*\mathbf{F}$ in a Lorentz frame in terms of that frame's electric and magnetic fields [analog of Eq. (23.70)]?

(c) Let $\vec{u}$ be the 4-velocity of some observer. Show that the 4-vectors $\mathbf{F}(\ldots,\vec{u}) \equiv \vec{E}_{\vec{u}}$ and $-*\mathbf{F}(\ldots,\vec{u}) \equiv \vec{B}_{\vec{u}}$ lie in the 3-space of that observer's local rest frame (i.e., they are orthogonal to the observer's 4-velocity), and are equal to the electric and magnetic fields of that 3-space, i.e., the electric and magnetic fields measured by that observer.

(d) There are only two independent scalars constructable from the electromagnetic field tensor: $F^{\mu\nu}F_{\mu\nu}$ and $*F^{\mu\nu}F_{\mu\nu}$. Show that, when expressed in terms of the electric and magnetic fields measured by any observer (i.e., of any Lorentz reference frame), these take the form

$$\boxed{F^{\mu\nu}F_{\mu\nu} = 2(\mathbf{B}^2 - \mathbf{E}^2) \ , \quad *F^{\mu\nu}F_{\mu\nu} = 4\mathbf{B}\cdot\mathbf{E} \ .} \tag{23.73}$$

Exercise 23.13 *Problem: Electromagnetic Stress-energy Tensor*
Expressed in geometric, frame-independent language, the Maxwell equations take the form

$$\boxed{F^{\alpha\beta}{}_{;\beta} = 4\pi J^\alpha \; ; \quad *F^{\alpha\beta}{}_{;\beta} = 0 \;\text{ or equivalently }\; F_{\alpha\beta;\gamma} + F_{\beta\gamma;\alpha} + F_{\gamma\alpha;\beta} = 0 \; .} \tag{23.74}$$

Here J^α is the density-current 4-vector, whose components in a specific Lorentz frame have the physical meanings

$$J^0 = (\text{charge density}) \; , \quad J^i = (i\text{-component of current density}) \; . \tag{23.75}$$

The stress-energy tensor for the electromagnetic field has the form

$$\boxed{T^{\mu\nu} = \frac{1}{4\pi}(F^{\mu\alpha}F^{\nu}{}_{\alpha} - \frac{1}{4}g^{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}) \; .} \tag{23.76}$$

(a) Show that in any Lorentz reference frame the electromagnetic energy density T^{00}, energy flux T^{0j}, momentum density T^{j0}, and stress T^{jk} have the following forms when expressed in terms of the electric and magnetic fields measured in that frame:

$$\begin{aligned} T^{00} &= \frac{\mathbf{E}^2 + \mathbf{B}^2}{8\pi} \; , \quad T^{i0} = T^{0i} = \frac{\epsilon^i{}_{jk}E^jB^k}{4\pi} \; , \\ T^{jk} &= \frac{1}{8\pi}\left[(\mathbf{E}^2 + \mathbf{B}^2)g^{jk} - 2(E^jE^k + B^jB^k)\right] \; . \end{aligned} \tag{23.77}$$

Show that, expressed in index-free notation, the energy flux has the standard Poynting-vector form $\mathbf{E}\times\mathbf{B}/4\pi$, and the stress tensor consists of a pressure $P_\perp = \mathbf{E}^2/8\pi$ orthogonal to $\mathbf{E}$, a pressure $P_\perp = \mathbf{B}^2/8\pi$ orthogonal to $\mathbf{B}$, a tension $-P_{||} = \mathbf{E}^2/8\pi$ along $\mathbf{E}$, and a tension $-P_{||} = \mathbf{B}^2/8\pi$ along $\mathbf{B}$.

(b) Show that the divergence of the stress-energy tensor (23.76) is given by

$$T^{\mu\nu}{}_{;\nu} = \frac{1}{4\pi}(F^{\mu\alpha}{}_{;\nu}F^{\nu}{}_{\alpha} + F^{\mu\alpha}F^{\nu}{}_{\alpha;\nu} - \frac{1}{2}F_{\alpha\beta}{}^{;\mu}F^{\alpha\beta}) \; . \tag{23.78}$$

(c) Combine this with the Maxwell equations to show that

$$\boldsymbol{\nabla}\cdot\mathbf{T} = -\mathbf{F}(\ldots,\mathbf{J}) \; ; \quad \text{i.e., } T^{\alpha\beta}{}_{;\beta} = -F^{\alpha\beta}J_\beta \; . \tag{23.79}$$

(d) Show that in a Lorentz reference frame the time and space components of this equation reduce to

$$\frac{\partial}{\partial t}T^{00} + \frac{\partial}{\partial x^j}T^{0j} = -E^jJ_j \equiv -(\text{rate of Joule heating}) \; , \tag{23.80}$$

$$\left(\frac{\partial}{\partial t}T^{k0} + \frac{\partial}{\partial x^j}T^{kj}\right)\mathbf{e}_k = -(J^0\mathbf{E} + \mathbf{J}\times\mathbf{B}) = -\begin{pmatrix}\text{Lorentz force}\\ \text{per unit volume}\end{pmatrix} \; . \tag{23.81}$$

Explain why these relations guarantee that, although the electromagnetic stress-energy tensor is not divergence-free, the total stress-energy tensor (electromagnetic plus that of the medium or fields that produce the charge-current 4-vector $\vec{J}$) is divergence-free; i.e., the total 4-momentum is conserved.

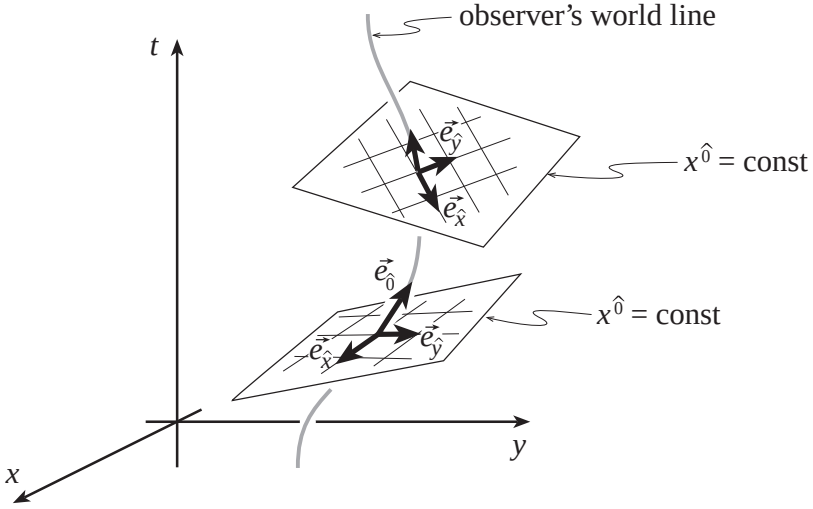


Fig. 23.6: The proper reference frame of an accelerated observer. The spatial basis vectors $\vec{e}_{\hat{x}}$, $\vec{e}_{\hat{y}}$, and $\vec{e}_{\hat{z}}$ are orthogonal to the observer's world line and rotate, relative to local gyroscopes, as they move along the world line. The flat 3-planes spanned by these basis vectors are surfaces of constant coordinate time $x^{\hat{0}} \equiv$ (proper time as measured by the observer's clock at the event where the 3-plane intersects the observer's world line); in other words, they are the observer's "3-space". In each of these flat 3-planes the spatial coordinates $\hat{x}$, $\hat{y}$, $\hat{z}$ are Cartesian, with $\partial/\partial\hat{x} = \vec{e}_{\hat{x}}$, $\partial/\partial\hat{y} = \vec{e}_{\hat{y}}$, $\partial/\partial\hat{z} = \vec{e}_{\hat{z}}$.

Exercise 23.14 *Derivation: Stress-Energy Tensor for a Point Particle*
Derive Eq. (23.68).

****

23.5 The Proper Reference Frame of an Accelerated Observer [MTW pp. 163–176, 327–332]

Physics experiments and astronomical measurements almost always use apparatus that accelerates and rotates. For example, if the apparatus is in an earth-bound laboratory and is attached to the laboratory floor and walls, then it accelerates upward (relative to freely falling particles) with the negative of the "acceleration of gravity", and it rotates (relative to inertial gyroscopes) because of the rotation of the earth. It is useful in studying such apparatus to regard it as attached to an accelerating, rotating reference frame. As preparation for studying such reference frames in the presence of gravity, we here shall study them in flat spacetime.

Consider an observer who moves along an accelerated world line through flat spacetime (Fig. 23.6) so she has a nonzero 4-acceleration

$$\vec{a} = \vec{\nabla}_{\vec{u}}\vec{u} \; . \tag{23.82}$$

Have that observer construct, in the vicinity of her world line, a coordinate system $\{x^{\hat{\alpha}}\}$ (called her *proper reference frame*) with these properties: (i) The spatial origin is centered on her world line at all times, i.e., her world line is given by $x^{\hat{j}} = 0$. (ii) Along her world line the time coordinate $x^{\hat{0}}$ is the same as the proper time ticked by an ideal clock that she carries. (iii) In the immediate vicinity of her world line the spatial coordinates $x^{\hat{j}}$ measure physical distance along the axes of a little Cartesian latticework that she carries. These properties dictate that in the immediate vicinity of her world line the metric has the form $ds^2 = \eta_{\hat{\alpha}\hat{\beta}} dx^{\hat{\alpha}} dx^{\hat{\beta}}$; in other words, all along her world line the coordinate basis vectors are orthonormal:

$$g_{\hat{\alpha}\hat{\beta}} = \frac{\partial}{\partial x^{\hat{\alpha}}} \cdot \frac{\partial}{\partial x^{\hat{\beta}}} = \eta_{\hat{\alpha}\hat{\beta}} \quad \text{at } x^{\hat{j}} = 0 \; . \tag{23.83}$$

Properties (i) and (ii) dictate, moreover, that along the observer's world line the basis vector $\vec{e}_{\hat{0}} \equiv \partial/\partial x^{\hat{0}}$ differentiates with respect to her proper time, and thus is identically equal to her 4-velocity $\vec{U}$,

$$\vec{e}_{\hat{0}} = \frac{\partial}{\partial x^{\hat{0}}} = \vec{U} \; . \tag{23.84}$$

There remains freedom as to how the observer's latticework is oriented, spatially: The observer can lock it to a gyroscopic inertial-guidance system, in which case we shall say that it is "nonrotating", or she can rotate it relative to such gyroscopes. We shall assume that the latticework rotates. Its angular velocity as measured by the observer (by comparing the latticework's orientation with inertial-guidance gyroscopes) is a 3-dimensional, spatial vector $\mathbf{\Omega}$; and as viewed geometrically, it is a 4-vector $\vec{\Omega}$ whose components in the observer's reference frame are $\Omega^{\hat{j}} \neq 0$ and $\Omega^{\hat{0}} = 0$, i.e., it is a 4-vector that is orthogonal to the observer's 4-velocity, $\vec{\Omega} \cdot \vec{U} = 0$; i.e., it is a 4-vector that lies in the observer's 3-space. Similarly, the latticework's acceleration as measured by an accelerometer attached to it is a 3-dimensional spatial vector $\mathbf{a}$ which can be thought of as a 4-vector with components in the observer's frame $a^{\hat{0}} = 0$, $a^{\hat{j}} = (\hat{j}$-component of the measured $\mathbf{a})$. This 4-vector, in fact, is the observer's 4-acceleration, as one can verify by computing the 4-acceleration in an inertial frame in which the observer is momentarily at rest.

Geometrically the coordinates of the proper reference frame are constructed as follows: (i) Begin with the basis vectors $\vec{e}_{\hat{\alpha}}$ along the observer's world line (Fig. 23.6)—basis vectors that satisfy equations (23.83) and (23.84), and that rotate with angular velocity $\vec{\Omega}$ relative to gyroscopes. Through the observer's world line at time $x^{\hat{0}}$ construct the flat 3-plane spanned by the spatial basis vectors $\vec{e}_{\hat{j}}$. Because $\vec{e}_{\hat{j}} \cdot \vec{e}_{\hat{0}} = 0$, this 3-plane is orthogonal to the world line. All events in this 3-plane are given the same value of coordinate time $x^{\hat{0}}$ as the event where it intersects the world line; thus the 3-plane is a surface of constant coordinate time $x^{\hat{0}}$. The spatial coordinates in this flat 3-plane are ordinary, Cartesian coordinates $x^{\hat{j}}$ with $\vec{e}_{\hat{j}} = \partial/\partial x^{\hat{j}}$.

It is instructive to examine the coordinate transformation between these proper-reference-frame coordinates $x^{\hat{\alpha}}$ and the coordinates x^{μ} of an inertial reference frame. We shall pick a very special inertial frame for this purpose: Choose an event on the observer's world line, near which the coordinate transformation is to be constructed; adjust the origin of her proper time so this event is $x^{\hat{0}} = 0$ (and of course $x^{\hat{j}} = 0$); and choose the inertial frame

to be one which, arbitrarily near this event, coincides with the observer's proper reference frame. Then, if we were doing Newtonian physics, the coordinate transformation from the proper reference frame to the inertial frame would have the form (accurate through terms quadratic in $x^{\hat{\alpha}}$)

$$x^i = x^{\hat{i}} + \frac{1}{2}a^{\hat{i}}(x^{\hat{0}})^2 + \epsilon^{\hat{i}}{}_{\hat{j}\hat{k}}\Omega^{\hat{j}}x^{\hat{k}}x^{\hat{0}} \ , \quad x^0 = x^{\hat{0}} \ . \tag{23.85}$$

Here the term $\frac{1}{2}a^{\hat{j}}(x^{\hat{0}})^2$ is the standard expression for the vectorial displacement produced, after time $x^{\hat{0}}$ by the acceleration $a^{\hat{j}}$; and the term $\epsilon^{\hat{i}}{}_{\hat{j}\hat{k}}\Omega^{\hat{j}}x^{\hat{k}}x^{\hat{0}}$ is the standard expression for the displacement produced by the rotation $\Omega^{\hat{j}}$ during a short time $x^{\hat{0}}$. In relativity theory there is only one departure from these familiar expressions (up through quadratic order): after time $x^{\hat{0}}$ the acceleration has produced a velocity $v^{\hat{j}} = a^{\hat{j}}x^{\hat{0}}$ of the proper reference frame relative to the inertial frame; and correspondingly there is a Lorentz-boost correction to the transformation of time: $x^0 = x^{\hat{0}} + v^{\hat{j}}x^{\hat{j}} = x^{\hat{0}}(1 + a_{\hat{j}}x^{\hat{j}})$ [cf. Eq. (1.49c)], accurate only to quadratic order. Thus, the full transformation to quadratic order is

$$\begin{aligned} x^i &= x^{\hat{i}} + \frac{1}{2}a^{\hat{i}}(x^{\hat{0}})^2 + \epsilon^{\hat{i}}{}_{\hat{j}\hat{k}}\Omega^{\hat{j}}x^{\hat{k}}x^{\hat{0}} \ , \\ x^0 &= x^{\hat{0}}(1 + a_{\hat{j}}x^{\hat{j}}) \ . \end{aligned} \tag{23.86}$$

From this transformation and the form of the metric, $ds^2 = -(dx^0)^2 + \delta_{ij}dx^i dx^j$ in the inertial frame, we easily can evaluate the form of the metric, accurate to linear order in $\mathbf{x}$, in the proper reference frame:

$$\boxed{ds^2 = -(1 + 2\mathbf{a}\cdot\mathbf{x})(dx^{\hat{0}})^2 + 2(\mathbf{\Omega}\times\mathbf{x})\cdot d\mathbf{x}\, dx^{\hat{0}} + \delta_{jk}dx^j dx^k} \tag{23.87}$$

[Ex. 23.15(a)]. Here the notation is that of 3-dimensional vector analysis, with $\mathbf{x}$ the 3-vector whose components are $x^{\hat{j}}$, $d\mathbf{x}$ that with components $dx^{\hat{j}}$, $\mathbf{a}$ that with components $a^{\hat{j}}$, and $\mathbf{\Omega}$ that with components $\Omega^{\hat{j}}$.

Because the transformation (23.86) was constructed near an arbitrary event on the observer's world line, the metric (23.87) is valid near any and every event on its world line; i.e., it is valid all along the world line. It, in fact, is the leading order in an expansion in powers of the spatial separation $x^{\hat{j}}$ from the world line. For higher order terms in this expansion see, e.g., Ni and Zimmermann (1978).

Notice that precisely on the observer's world line, the metric coefficients $g_{\hat{\alpha}\hat{\beta}}$ [the coefficients of $dx^{\hat{\alpha}}dx^{\hat{\beta}}$ in Eq. (23.87)] are $g_{\hat{\alpha}\hat{\beta}} = \eta_{\hat{\alpha}\hat{\beta}}$, in accord with equation (23.83). However, as one moves farther and farther away from the observer's world line, the effects of the acceleration $a^{\hat{j}}$ and rotation $\Omega^{\hat{j}}$ cause the metric coefficients to deviate more and more strongly from $\eta_{\hat{\alpha}\hat{\beta}}$.

From the metric coefficients of (23.87) one can compute the connection coefficients $\Gamma^{\hat{\alpha}}{}_{\hat{\beta}\hat{\gamma}}$ on the observer's world line; and from these connection coefficients one can infer the rates of change of the basis vectors along the world line,

$$\nabla_{\vec{U}}\vec{e}_{\hat{\alpha}} = \nabla_{\hat{0}}\vec{e}_{\hat{\alpha}} = \Gamma^{\hat{\mu}}{}_{\hat{\alpha}\hat{0}}\vec{e}_{\hat{\mu}} \ . \tag{23.88}$$

The result is (cf. Ex. 23.15)

$$\nabla_{\vec{U}}\hat{\vec{e}}_{\hat{0}} \equiv \nabla_{\vec{U}}\vec{U} = \vec{a}\ , \tag{23.89}$$

$$\nabla_{\vec{U}}\vec{e}_{\hat{j}} = (\vec{a}\cdot\vec{e}_{\hat{j}})\vec{U} + \boldsymbol{\epsilon}(\vec{U},\vec{\Omega},\vec{e}_{\hat{j}},\ \ldots)\ . \tag{23.90}$$

Equation (23.90) is a special case of a general "law of transport" for vectors that are orthogonal to the observer's world line and that the observer thus sees as purely spatial: For the spin vector $\vec{S}$ of an inertial-guidance gyroscope (one which the observer carries with herself, applying the forces that make it accelerate precisely at its center of mass so they do not also make it precess), the transport law is (23.90) with $\vec{e}_{\hat{j}}$ replaced by $\vec{S}$ and with $\vec{\Omega}=0$:

$$\boxed{\nabla_{\vec{U}}\vec{S} = \vec{U}(\vec{a}\cdot\vec{S})\ .} \tag{23.91}$$

The term on the right-hand side of this transport law is required to keep the spin vector always orthogonal to the observer's 4-velocity, $\nabla_{\vec{U}}(\vec{S}\cdot\vec{U}) = 0$. For any other vector $\vec{A}$, which rotates relative to inertial-guidance gyroscopes, the transport law has in addition to this "keep-it-orthogonal-to $\vec{U}$" term, also a second term which is the 4-vector form of $d\mathbf{A}/dt = \boldsymbol{\Omega}\times\mathbf{A}$:

$$\nabla_{\vec{U}}\vec{A} = \vec{U}(\vec{a}\cdot\vec{A}) + \boldsymbol{\epsilon}(\vec{U},\ \vec{\Omega},\ \vec{A},\ \ldots\ )\ . \tag{23.92}$$

Equation (23.90) is this general transport law with $\vec{A}$ replaced by $\vec{e}_{\hat{j}}$.

Consider a particle which moves freely through the neighborhood of an accelerated observer. As seen in an inertial reference frame, the particle moves through spacetime on a straight line, also called a *geodesic* of flat spacetime. Correspondingly, a geometric, frame-independent version of its *geodesic law of motion* is

$$\boxed{\nabla_{\vec{u}}\vec{u} = 0\ ;} \tag{23.93}$$

i.e., it parallel transports its 4-velocity $\vec{u}$ along itself. It is instructive to examine the component form of this "geodesic equation" in the proper reference frame of the observer. Since the components of $\vec{u}$ in this frame are $u^{\alpha} = dx^{\alpha}/d\tau$, where τ is the particle's proper time (not the observer's proper time), the components $u^{\hat{\alpha}}{}_{;\hat{\mu}}u^{\hat{\mu}} = 0$ of the geodesic equation (23.93) are

$$u^{\hat{\alpha}}{}_{,\hat{\mu}}u^{\hat{\mu}} + \Gamma^{\hat{\alpha}}{}_{\hat{\mu}\hat{\nu}}u^{\hat{\mu}}u^{\hat{\nu}} = 0\ ; \tag{23.94}$$

or equivalently

$$\boxed{\frac{d^2x^{\hat{\alpha}}}{d\tau^2} + \Gamma^{\hat{\alpha}}{}_{\hat{\mu}\hat{\nu}}\frac{dx^{\hat{\mu}}}{d\tau}\frac{dx^{\hat{\nu}}}{d\tau} = 0\ .} \tag{23.95}$$

Suppose for simplicity that the particle is moving slowly relative to the observer, so its ordinary velocity $v^{\hat{j}} = dx^{\hat{j}}/dx^{\hat{0}}$ is very nearly equal to $u^{\hat{j}} = dx^{\hat{j}}/d\tau$ and is very small compared to unity (the speed of light), and $u^{\hat{0}} = dx^{\hat{0}}/d\tau$ is very nearly unity. Then to first order in the ordinary velocity $v^{\hat{j}}$, the spatial part of the geodesic equation (23.95) becomes

$$\frac{d^2x^{\hat{i}}}{(dx^{\hat{0}})^2} = -\Gamma^{\hat{i}}{}_{\hat{0}\hat{0}} - (\Gamma^{\hat{i}}{}_{\hat{j}\hat{0}} + \Gamma^{\hat{i}}{}_{\hat{0}\hat{j}})v^{\hat{j}}\ . \tag{23.96}$$

By computing the connection coefficients from the metric coefficients of (23.87) [Ex. 23.15], we bring this low-velocity geodesic law of motion into the form

$$\frac{d^2x^{\hat{i}}}{(dx^{\hat{0}})^2} = -a^{\hat{i}} - 2\epsilon^{\hat{i}}{}_{\hat{j}\hat{k}}\Omega^{\hat{j}}v^{\hat{k}} \ , \quad \text{i.e.,} \ \frac{d^2\mathbf{x}}{(dx^{\hat{0}})^2} = -\mathbf{a} - 2\mathbf{\Omega}\times\mathbf{v} \ . \tag{23.97}$$

This is the standard nonrelativistic form of the law of motion for a free particle as seen in a rotating, accelerating reference frame: the first term on the right-hand side is the inertial acceleration due to the failure of the frame to fall freely, and the second term is the Coriolis acceleration due to the frame's rotation. There would also be a centrifugal acceleration if we had kept terms higher order in distance away from the observer's world line, but it has been lost due to our linearizing the metric (23.87) in that distance.

This analysis shows how the elegant formalism of tensor analysis gives rise to familiar physics. In the next few chapters we will see it give rise to less familiar, general relativistic phenomena.

****

EXERCISES

Exercise 23.15 *Example: Proper Reference Frame*

(a) Show that the coordinate transformation (23.86) brings the metric $ds^2 = \eta_{\alpha\beta}dx^\alpha dx^\beta$ into the form (23.87), accurate to linear order in separation $x^{\hat{j}}$ from the origin of coordinates.

(b) Compute the connection coefficients for the coordinate basis of (23.87) at an arbitrary event on the observer's world line. Do so first by hand calculations, and then verify your results using symbolic-manipulation software on a computer.

(c) From those connection coefficients show that the rate of change of the basis vectors $\mathbf{e}_{\hat{\alpha}}$ along the observer's world line is given by (23.89), (23.90).

(d) From the connection coefficients show that the low-velocity limit (23.96) of the geodesic equation is given by (23.97).

Exercise 23.16 *Problem: Uniformly Accelerated Observer*

As a special example of an accelerated observer, consider one whose world line, expressed in terms of a Lorentz coordinate system (t, x, y, z), is

$$t = \frac{1}{a}\sinh(a\tau) \ , \quad x = \frac{1}{a}\cosh(a\tau) \ , \quad y = 0 \ , \quad z = 0 \ . \tag{23.98}$$

Here a is a constant with dimensions of 1/(length) or equivalently (length)/(time)^2, and τ is a parameter that varies along the accelerated world line.

(a) Show that τ is the observer's proper time, and evaluate the observer's 4-acceleration $\vec{a}$, and show that $|\vec{a}| = a$ where a is the constant in (23.98), so the observer feels constant, time-independent accleration in his proper reference frame.

(b) The basis vectors $\mathbf{e}_{\hat{0}}$ and $\mathbf{e}_{\hat{1}}$ of the observer's proper reference frame lie in the t, x plane in spacetime, $\mathbf{e}_{\hat{2}}$ points along the y-axis, and $\mathbf{e}_{\hat{3}}$ points along the z axis. Draw a spacetime diagram, on it draw the observer's world line, and at several points along it draw the basis vectors $\mathbf{e}_{\hat{\mu}}$. What are $\mathbf{e}_{\hat{\mu}}$ in terms of the Lorentz coordinate basis vectors $\partial/\partial x^\alpha$?

(c) What is the angular velocity $\vec{\Omega}$ of the proper reference frame?

(d) Express the coordinates $x^{\hat{\mu}}$ of the observer's proper reference frame in terms of the Lorentz coordinates (t, x, y, z) accurate to first order in distance away from the observer's world line and accurate for all proper times τ. Show that under this coordinate transformation the Lorentz-frame components of the metric, $g_{\alpha\beta} = \eta_{\alpha\beta}$, are transformed into the components given by Eq. (23.87).

Exercise 23.17 *Challenge: Thomas Precession*
As is well known in quantum mechanics, the spin-orbit contribution to the Hamiltonian for an electron in an atom is

$$H_{SO} = \frac{-e}{2m_e^2c^2r}\frac{d\phi}{dr}\mathbf{L}\cdot\mathbf{S} \tag{23.99}$$

where ϕ is the electrostatic potential and $\mathbf{L}, \mathbf{S}$ are the electron's angular momentum and spin respectively. This is one half the naive value and the difference, known as the *Thomas precession*, is a purely special relativistic kinematic effect. Using the language of this chapter, explain from first principles how the Thomas precession arises.

****

Bibliographic Note

For a very readable presentation of most of this chapter's material, from much the same point of view, see Chap. 20 of Hartle (2003). For an equally elementary introduction from a somewhat different viewpoint, see Chaps. 1–4 of Schutz (1980). A far more detailed and somewhat more sophisticated introduction, largely but not entirely from our viewpoint, will be found in Chaps. 1–6 of Misner, Thorne and Wheeler (1973). More sophisticated treatments from rather different viewpoints than ours are given in Chaps. 1 and 2 and Sec. 3.1 of Wald (1984), and in Chaps. 1 and 2 of Carroll (2004). A treasure trove of exercises on this material, with solutions, will be found in Chaps. 6, 7, and 8 of Lightman, Press, Price and Teukolsky (1975).

Box 23.2
Important Concepts in Chapter 23

- Most important concepts from Chap. 1
  - Principle of Relativity, Sec. 23.2.1
  - Metric defined in terms of interval, Sec. 23.2.1
  - Inertial frames, Sec. 23.2.2
  - Interval and spacetime diagrams, Sec. 23.2.3
- Differential geometry in general bases, Sec. 23.3
  - Dual bases, $\{\vec{e}_\alpha\}$, $\{\vec{e}^\mu\}$ with $\vec{e}_\alpha \cdot \vec{e}^\mu = \delta^\mu_\alpha$, Sec. 23.3.1
  - Covariant, contravariant and mixed components of a tensor, Sec. 23.3.1
  - Changes of bases and corresponding transformation of components of tensors, Sec. 23.3.1
  - Coordinate bases, Sec. 23.3.1
  - Orthonormal bases, Sec. 23.2.2
  - Vector as a differential operator (directional derivative), Sec. 23.3.2
  - Tangent space, Sec. 23.3.2
  - Commutator of vector fields, Sec. 23.3.2
  - Parallel transport of vectors, Sec. 23.3.3
  - Connection coefficients, how to compute them, and how to use them in computing components of the gradients of tensor fields, Sec. 23.3.3
  - Christoffel symbols (connection coefficients in a coordinate basis), Sec. 23.3.3
  - Levi-Civita tensor and its components, Sec. 23.3.4
  - Volume elements for integration, Sec. 23.3.4
- Stress-energy tensor, Sec. 23.4
  - Definition, Sec. 23.4
  - For perfect fluid, Sec. 23.4
  - For point particle, Sec. 23.4
  - For electromagnetic field, Ex. 23.13
- Proper reference frame of an accelerated observer and metric in it, Sec. 23.5
  - Transport law for inertial-guidance gyroscope, Sec. 23.5
  - Geodesic law of motion, Sec. 23.5

Bibliography

Carroll, S. M., 2004. *Spacetime and Geometry: An Introduction to General Relativity*, San Francisco: Addison Wesley.

Feynman, R. P., 1966. *The Character of Physical Law*, Cambridge MA: MIT Press.

Hartle, J. B., 2003. *Gravity: An Introduction to Einstein's General Relativity*, San Francisco: Addison-Wesley.

Lightman, A. P., Press, W. H., Price, R. H. & Teukolsky, S. A. 1975. *Problem Book In Relativity and Gravitation*, Princeton: Princeton University Press.

MTW: Misner, Charles W., Thorne, Kip S., and Wheeler, John A. 1973. *Gravitation*, San Francisco:W. H. Freeman.

Ni, W.-T., and Zimmermann, M., 1978. "Inertial and gravitational effects in the proper reference frame of an accelerated, rotating observer," *Physical Review D*, **17**, 1473.

Schutz, B. 1980. *Geometrical Methods of Mathematical Physics*, Cambridge: Cambridge University Press.

Wald, R. M. 1984. *General Relativity*, Chicago: University of Chicago Press.

Contents

24 Fundamental Concepts of General Relativity **1**

24.1 Overview 1

24.2 Local Lorentz Frames, the Principle of Relativity, and Einstein's Equivalence Principle 5

24.3 The Spacetime Metric, and Gravity as a Curvature of Spacetime 6

24.4 Free-fall Motion and Geodesics of Spacetime 9

24.5 Relative Acceleration, Tidal Gravity, and Spacetime Curvature 14

24.5.1 Newtonian Description of Tidal Gravity 14

24.5.2 Relativistic Description of Tidal Gravity 16

24.5.3 Comparison of Newtonian and Relativistic Descriptions 17

24.6 Properties of the Riemann Curvature Tensor 19

24.7 Curvature Coupling Delicacies in the Equivalence Principle, and Some Non-gravitational Laws of Physics in Curved Spacetime[1] 23

24.8 The Einstein Field Equation[2] 26

24.9 Weak Gravitational Fields 29

24.9.1 Newtonian Limit of General Relativity 30

24.9.2 Linearized Theory 31

24.9.3 Gravitational Field Outside a Stationary, Linearized Source 33

24.9.4 Conservation Laws for Mass, Momentum and Angular Momentum . . 35

[1] See MTW Chap. 16.

[2] See MTW Chap. 17.

Chapter 24

Fundamental Concepts of General Relativity

Version 0624.1.K.pdf, 02 May 2007
Please send comments, suggestions, and errata via email to kip@tapir.caltech.edu or on paper to Kip Thorne, 130-33 Caltech, Pasadena CA 91125

Box 24.1
Reader's Guide

- This chapter relies significantly on
  - The special relativity portions of Chap. 1.
  - Chapter 23, on the transition from special relativity to general relativity.
- This chapter is a foundation for the applications of general relativity theory in Chaps. 25–27.

24.1 Overview

Newton's theory of gravity is logically incompatible with the special theory of relativity: Newtonian gravity presumes the existence of a universal, frame-independent 3-dimensional space in which lives the Newtonian potential Φ, and a universal, frame-independent time t with respect to which the propagation of Φ is instantaneous. Special relativity, by contrast, insists that the concepts of time and of 3-dimensional space are frame-dependent, so that instantaneous propagation of Φ in one frame would mean non-instantaneous propagation in another.

The most straightforward way to remedy this incompatibility is to retain the assumption that gravity is described by a scalar field Φ, but modify Newton's instantaneous, action-at-

a-distance field equation

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\Phi = 4\pi G\rho \tag{24.1}$$

(where G is Newton's gravitation constant and ρ is the mass density) to read

$$\vec{\nabla}^2\Phi \;\equiv\; g^{\alpha\beta}\Phi_{;\alpha\beta} = -4\pi G T^{\mu}{}_{\mu}\ , \tag{24.2}$$

where $\vec{\nabla}^2$ is the squared gradient, or d'Alembertian in Minkowski spacetime and $T^{\mu}{}_{\mu}$ is the trace (contraction on its slots) of the stress-energy tensor. This modified field equation at first sight is attractive and satisfactory (but see Ex. 24.1, below): (*i*) It satisfies Einstein's Principle of Relativity in that it is expressed as a geometric, frame-independent relationship between geometric objects; and (*ii*) in any Lorentz frame it takes the form [with factors of $c =$ (speed of light) restored]

$$\left(-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}+\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\Phi = \frac{4\pi G}{c^2}(T^{00}-T^{xx}-T^{yy}-T^{zz})\ , \tag{24.3}$$

which, in the kinds of situation contemplated by Newton [energy density predominantly due to rest mass, $T^{00}\cong\rho c^2$; stress negligible compared to rest mass-energy, $|T^{jk}|\ll\rho c^2$; and $1/c\times$ (time rate of change of Φ) negligible compared to spatial gradient of Φ], reduces to the Newtonian field equation (24.1).

Not surprisingly, most theoretical physicists in the decade following Einstein's formulation of special relativity (1905–1915) presumed that gravity would be correctly describable, within the framework of special relativity, by this type of modification of Newton's theory, or something resembling it. For a brief historical account see Chap. 13 of Pais (1982). To Einstein, by contrast, it seemed clear as early as 1907 that the correct description of gravity should involve a generalization of special relativity rather than an incorporation into special relativity: Since an observer in a local, freely falling reference frame near the earth should not feel any gravitational acceleration at all, local freely falling frames (*local inertial frames*) should in some sense be the domain of special relativity, and gravity should somehow be described by the *relative acceleration* of such frames.

Although the seeds of this idea were in Einstein's mind as early as 1907 (see the discussion of the equivalence principle in Einstein, 1907), it required eight years for him to bring them to fruition. A first crucial step, which took half the eight years, was for Einstein to conquer his initial aversion to Minkowski's (1908) geometric formulation of special relativity, and to realize that a curvature of Minkowski's 4-dimensional spacetime is the key to understanding the relative acceleration of freely falling frames. The second crucial step was to master the mathematics of differential geometry, which describes spacetime curvature, and using that mathematics to formulate a logically self-consistent theory of gravity. This second step took an additional four years and culminated in Einstein's (1915, 1916) general theory of relativity. For a historical account of Einstein's eight-year struggle toward general relativity see, e.g., Part IV of Pais (1982); and for selected quotations from Einstein's technical papers during this eight-year period, which tell the story of his struggle in his own words, see Sec. 17.7 of MTW.

It is remarkable that Einstein was led, not by experiment, but by philosophical and aesthetic arguments, to reject the incorporation of gravity into special relativity [Eqs. (24.2) and (24.3) above], and insist instead on describing gravity by curved spacetime. Only after the full formulation of his general relativity did experiments begin to confirm that he was right and that the advocates of special-relativistic gravity were wrong, and only more than 50 years after general relativity was formulated did the experimental evidence become extensive and strong. For detailed discussions see, e.g., Will (1981, 1986), and Part 9 of MTW.

The mathematical tools, the diagrams, and the phrases by which we describe general relativity have changed somewhat in the seventy years since Einstein formulated his theory; and, indeed, we can even assert that we understand the theory more deeply than did Einstein. However, the basic ideas are unchanged; and general relativity's claim to be the most elegant and aesthetic of physical theories has been reinforced and strengthened by our growing insights.

General relativity is not merely a theory of gravity. Like special relativity before it, the general theory is a framework within which to formulate all the laws of physics, classical and quantum—but now with gravity included. However, there is one remaining, crucial, gaping hole in this framework: It is incapable of functioning, indeed it fails completely, when conditions become so extreme that space and time themselves must be quantized. In those extreme conditions general relativity must be married in some deep, as-yet-ill-understood way, with quantum theory, to produce an all-inclusive quantum theory of gravity—a theory which, one may hope, will be a "theory of everything." To this we shall return, briefly, in Chaps. 25 and 27.

In this chapter we present, in modern language, the foundations of general relativity. Our presentation will rely heavily on the concepts, viewpoint, and formalism developed in Chaps. 1 and 23.

We shall begin in Sec. 24.2 with a discussion of three concepts that are crucial to Einstein's viewpoint on gravity: a local Lorentz frame (the closest thing there is, in the presence of gravity, to special relativity's "global" Lorentz frame), the extension of the principle of relativity to deal with gravitational situations, and Einstein's equivalence principle by which one can "lift" laws of physics out of the flat spacetime of special relativity and into the curved spacetime of general relativity. In Sec. 24.3 we shall see how gravity prevents the meshing of local Lorentz frames to form global Lorentz frames, and shall infer from this that spacetime must be curved. In Sec. 24.4 we shall lift into curved spacetime the law of motion for free test particles, and in Sec. 24.5 we shall see how spacetime curvature pushes two freely moving test particles apart and shall use this phenomenon to make contact between spacetime curvature and the Newtonian "tidal gravitational field" (gradient of the Newtonian gravitational acceleration). In Sec. 24.6 we shall study a number of mathematical and geometric properties of the tensor field that embodies spacetime curvature: the Riemann tensor. In Sec. 24.7 we shall examine "curvature coupling delicacies" which plague the lifting of laws of physics from flat spacetime to curved spacetime. In Sec. 24.8 we shall meet the Einstein field equation, which describes the manner in which spacetime curvature is produced by the total stress-energy tensor of all matter and nongravitational fields. In Ex. 24.12 we shall examine in some detail how Newton's laws of gravity arise as a weak-gravity limit of

general relativity. Finally, in Sec. 24.9 we shall examine the conservation laws for energy, momentum, and angular momentum of gravitating bodies that live in "asymptotically flat" regions of spacetime.

****

EXERCISES

Exercise 24.1 *Example: A Special Relativistic, Scalar-Field Theory of Gravity*
Equation (24.2) is the field equation for a special relativistic theory of gravity with gravitational potential Φ. To complete the theory one must describe the forces that the field Φ produces on matter.

(a) One conceivable choice for the force on a test particle of rest mass m is the following generalization of the familiar Newtonian expression:

$$\nabla_{\vec{u}}\vec{p} = -m\vec{\nabla}\Phi \ ; \quad \text{i.e.,} \ \ \frac{dp^\alpha}{d\tau} = -m\Phi^{,\alpha} \ \ \text{in a Lorentz frame,} \tag{24.4}$$

where τ is proper time along the particle's world line, $\vec{p}$ is the particle's 4-momentum, $\vec{u}$ is its 4-velocity, and $\vec{\nabla}\Phi$ is the spacetime gradient of the gravitational potential. Show that this equation of motion reduces, in a Lorentz frame and for low particle velocities, to the standard Newtonian equation of motion. Show, however, that this equation of motion is flawed in that the gravitational field will alter the particle's rest mass—in violation of extensive experimental evidence that the rest mass of an elementary particle is unique and conserved.

(b) Show that the above equation of motion, when modified to read

$$\nabla_{\vec{u}}\vec{p} = -(\mathbf{g} + \vec{u}\otimes\vec{u})\cdot m\vec{\nabla}\Phi \ ;$$
$$\text{i.e.,} \ \ \frac{dp^\alpha}{d\tau} = -(g^{\alpha\beta} + u^\alpha u^\beta)m\Phi_{,\beta} \ \ \text{in a Lorentz frame,} \tag{24.5}$$

preserves the particle's rest mass. In this equation of motion $\vec{u}$ is the particle's 4-velocity, and $\mathbf{g} + \vec{u}\otimes\vec{u}$ projects $\vec{\nabla}\Phi$ into the "3-space" orthogonal to the particle's world line; cf. Fig. 23.6.

(c) Show, by treating a zero-rest-mass particle as the limit of a particle of finite rest mass ($\vec{p} = m\vec{u}$ and $\zeta = \tau/m$ finite as τ and m go to zero), that the above theory predicts that in any Lorentz reference frame $p^\alpha e^\Phi$ (with $\alpha = 0, 1, 2, 3$) are constant along the zero-rest-mass particle's world line. Explain why this prediction implies that there will be no deflection of light around the limb of the sun, which conflicts severely with experiments that were done *after* Einstein formulated his general theory of relativity. (There was no way, experimentally, to rule out the above theory in the epoch, ca. 1914, when Einstein was doing battle with his colleagues over whether gravity should be treated within the framework of special relativity or should be treated as a geometric extension of special relativity.)

****

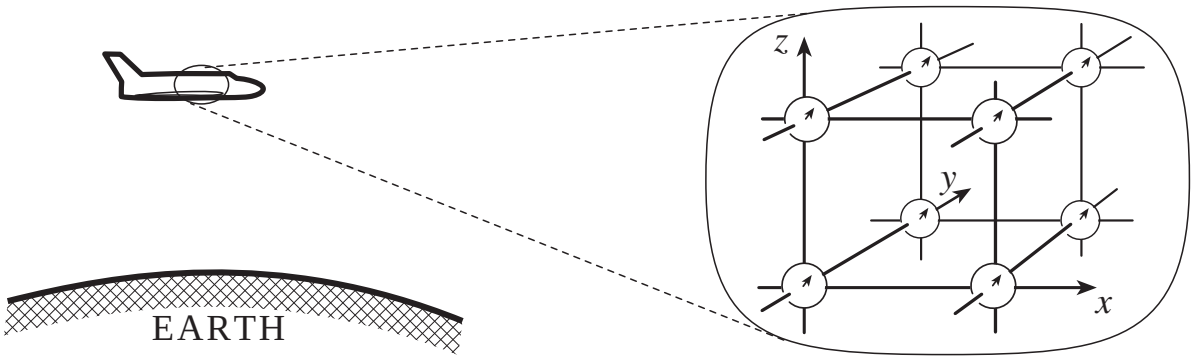


Fig. 24.1: A local inertial frame (local Lorentz frame) inside a space shuttle that is orbiting the earth.

24.2 Local Lorentz Frames, the Principle of Relativity, and Einstein's Equivalence Principle

One of Einstein's greatest insights was to recognize that special relativity is valid not globally, but only locally, inside locally freely falling (inertial) reference frames. Figure 24.1 shows a specific example of a *local inertial frame*: The interior of a space shuttle in earth orbit, where an astronaut has set up a freely falling (from his viewpoint "freely floating") latticework of rods and clocks. This latticework is constructed by all the rules appropriate to a special relativistic, inertial (Lorentz) reference frame (Sec. 1.2): (*i*) the latticework moves freely through spacetime so no forces act on it, and its rods are attached to gyroscopes so they do not rotate; (*ii*) the measuring rods are orthogonal to each other, with their intervals of length uniform compared, e.g., to the wavelength of light (orthonormal lattice); (*iii*) the clocks are densely packed in the lattice, they tick uniformly relative to ideal atomic standards (they are ideal clocks), and they are synchronized by the Einstein light-pulse process. However, there is one crucial change from special relativity: The latticework must be *small enough* that one can neglect the effects of inhomogeneities of gravity (which general relativity will associate with spacetime curvature; and which, for example, would cause two freely floating particles, one nearer the earth than the other, to gradually move apart even though initially they are at rest with respect to each other). The necessity for smallness is embodied in the word "local" of "local inertial frame", and we shall quantify it with ever greater precision as we move on through this chapter.

We shall use the phrases *local Lorentz frame* and *local inertial frame* interchangeably to describe the above type of synchronized, orthonormal latticework; and the spacetime coordinates t, x, y, z that the latticework provides (in the manner of Sec. 1.2) we shall call, interchangeably, *local Lorentz coordinates* and *local inertial coordinates*.

Since, in the presence of gravity, inertial reference frames must be restricted to be local, the inertial-frame version of the *principle of relativity* must similarly be restricted to say: *All the* **local**, *nongravitational laws of physics are the same in every* **local** *inertial frame, everywhere and everywhen in the universe*. Here, by "local" laws we mean those laws, classical or quantum, which can be expressed entirely in terms of quantities confined to (measurable within) a local inertial frame; and the exclusion of gravitational laws from this

version of the principle of relativity is necessary because gravity is to be described by a curvature of spacetime which (by definition, see below) cannot show up in a local inertial frame. This version of the principle of relativity can be described in operational terms using precisely the same language as for the special relativistic version (Secs. 23.2.1 and 23.2.2): If two different observers, in two different local Lorentz frames, in different (or the same) regions of the universe and epochs of the universe, are given identical written instructions for a self-contained physics experiment (an experiment that can be performed within the confines of the local Lorentz frame), then their two experiments must yield the same results, to within their experimental accuracies.

It is worth emphasizing that the principle of relativity is asserted to hold everywhere and everywhen in the universe: the local laws of physics must have the same form in the early universe, a fraction of a second after the big bang, as they have on earth today, and as they have at the center of the sun or inside a black hole.

It is reasonable to expect that *the specific forms that the local, nongravitational laws of physics take in general relativistic local Lorentz frames are the same as they take in the (global) Lorentz frames of special relativity*. The assertion that this is so is a modern version of *Einstein's equivalence principle*. In the next section we will use this principle to deduce some properties of the general relativistic spacetime metric; and in Sec. 24.7 we will use it to deduce the explicit forms of some of the nongravitational laws of physics in curved spacetime.

24.3 The Spacetime Metric, and Gravity as a Curvature of Spacetime

The Einstein equivalence principle guarantees that nongravitational physics within a local Lorentz frame can be described using a spacetime metric $\mathbf{g}$, which gives for the invariant interval between neighboring events with separation vector $\vec{\xi} = \Delta x^\alpha \partial/\partial x^\alpha$, the standard special relativistic expression

$$\vec{\xi}^2 = g_{\alpha\beta}\xi^\alpha\xi^\beta = (\Delta s)^2 = -(\Delta t)^2 + (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 \ . \tag{24.6}$$

Correspondingly, in a local Lorentz frame the components of the spacetime metric take on their standard special-relativity values

$$g_{\alpha\beta} = \eta_{\alpha\beta} \equiv \{-1 \text{ if } \alpha = \beta = 0 \ , \ \ +1 \text{ if } \alpha = \beta = (x, \text{ or } y, \text{ or } z), \ \ 0 \text{ otherwise}\} \ . \tag{24.7}$$

Turn, now, to a first look at the gravity-induced constraints on the size of a local Lorentz frame: Above the earth set up, initially, a family of local Lorentz frames scattered over the entire region from two earth radii out to four earth radii, with all the frames initially at rest with respect to the earth [Fig. 24.2(a)]. From experience—or, if you prefer, from Newton's theory of gravity which after all is quite accurate near earth—we know that as time passes these frames will all fall toward the earth. If (as a pedagogical aid) we drill holes through the earth to let the frames continue falling after reaching the earth's surface, the frames will all pass through the earth's center and fly out the earth's opposite side.

Obviously, two adjacent frames, which initially were at rest with respect to each other, acquire a relative velocity during their fall, which causes them to interpenetrate and pass

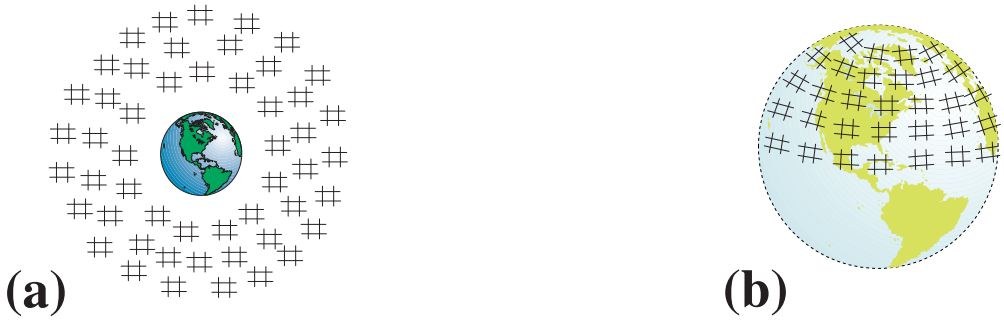


Fig. 24.2: (a) A family of local Lorentz frames, all momentarily at rest above the earth's surface. (b) A family of local, 2-dimensional Euclidean coordinate systems on the earth's surface. The nonmeshing of Lorentz frames in (a) is analogous to the nonmeshing of Euclidean coordinates in (b) and motivates attributing gravity to a curvature of spacetime.

through each other as they cross the earth's center. Gravity is the cause of this relative velocity.

If these two adjacent frames could be meshed to form a larger Lorentz frame, then as time passes they would always remain at rest relative to each other. Thus, a meshing to form a larger Lorentz frame is impossible. The gravity-induced relative velocity prevents it. In brief: *Gravity prevents the meshing of local Lorentz frames to form global Lorentz frames*.

This situation is closely analogous to the nonmeshing of local, 2-dimensional, Euclidean coordinate systems on the surface of the earth [Figure 24.2(b)]: The curvature of the earth prevents a Euclidean mesh—thereby giving grief to map makers and surveyors. This analogy suggested to Einstein, in 1912, a powerful new viewpoint on gravity: Just as the curvature of space prevents the meshing of local Euclidean coordinates on the earth's surface, so it must be that a curvature of spacetime prevents the meshing of local Lorentz frames in the spacetime above the earth—or anywhere else in spacetime, for that matter. And since it is already known that gravity is the cause of the nonmeshing of Lorentz frames, it must be that *gravity is a manifestation of spacetime curvature*.

To make this idea more quantitative, consider, as a pedagogical tool, the 2-dimensional metric of the earth's surface expressed in terms of a spherical polar coordinate system and in "line-element" form:

$$ds^2 = R^2 d\theta^2 + R^2 \sin^2\theta d\phi^2 \ . \tag{24.8}$$

Here R is the radius of the earth, or equivalently the "radius of curvature" of the earth's surface. This line element, rewritten in terms of the alternative coordinates

$$x \equiv R\phi \ , \quad y \equiv R\left(\frac{\pi}{2} - \theta\right) \ , \tag{24.9}$$

has the form

$$ds^2 = \cos^2(y/R)dx^2 + dy^2 = dx^2 + dy^2 + \mathrm{O}(y^2/R^2)dx^2 \ , \tag{24.10}$$

where $\mathrm{O}(y^2/R^2)$ means "terms of order y^2/R^2" or smaller. Notice that the metric coefficients have the standard Euclidean form $g_{jk} = \delta_{jk}$ all along the equator ($y = 0$); but as one moves away from the equator, they begin to differ from Euclidean by fractional amounts of

$O(y^2/R^2) = O[y^2/(\text{radius of curvature of earth})^2]$. Thus, local Euclidean coordinates can be meshed and remain Euclidean all along the equator—or along any other great circle—, but the earth's curvature forces the coordinates to cease being Euclidean when one moves off the chosen great circle, thereby causing the metric coefficients to differ from δ_{jk} by amounts $\Delta g_{jk} = O[(\text{distance from great circle})^2 \ / \ (\text{radius of curvature})^2]$.

Turn next to a specific example of curved spacetime: that of a "$k = 0$ Friedmann model" for our expanding universe (to be studied in depth in Chap. 27 below). In spherical coordinates (η, χ, θ, ϕ), the 4-dimensional metric of this curved spacetime, described as a line element, takes the form

$$ds^2 = a^2(\eta)[-d\eta^2 + d\chi^2 + \chi^2(d\theta^2 + \sin^2\theta d\phi^2)] \ , \tag{24.11}$$

where a, the "expansion factor of the universe," is a monotonic increasing function of the "time" coordinate η. This line element, rewritten near $\chi = 0$ in terms of the alternative coordinates

$$t = \int_0^\eta a d\eta \ + \ \frac{1}{2}\chi^2\frac{da}{d\eta} \ , \quad x = a\chi\sin\theta\cos\phi \ , \quad y = a\chi\sin\theta\sin\phi \ , \quad z = a\chi\cos\theta \ , \tag{24.12}$$

takes the form [cf. Ex. 24.2]

$$ds^2 = \eta_{\alpha\beta}dx^\alpha dx^\beta + O\left(\frac{x^2+y^2+z^2}{\mathcal{R}^2}\right)dx^\alpha dx^\beta \ , \tag{24.13}$$

where $\mathcal{R}$ is a quantity which, by analogy with the radius of curvature R of the earth's surface, can be identified as a "radius of curvature" of spacetime:

$$\frac{1}{\mathcal{R}^2} = O\left(\frac{\dot{a}^2}{a^2} + \frac{\ddot{a}}{a}\right) \quad \text{where } \dot{a} \equiv \left(\frac{da}{dt}\right)_{x=y=z=0}, \quad \ddot{a} \equiv \left(\frac{d^2a}{dt^2}\right)_{x=y=z=0} . \tag{24.14}$$

From the form of the metric coefficients in Eq. (24.14) we see that all along the world line $x = y = z = 0$ the coordinates are precisely Lorentz, but as one moves away from that world line they cease to be Lorentz, and the metric coefficients begin to differ from $\eta_{\alpha\beta}$ by amounts $\Delta g_{\alpha\beta} = O[(\text{distance from the chosen world line})^2/(\text{radius of curvature of spacetime})^2]$. This is completely analogous to our equatorial Euclidean coordinates on the earth's surface. The curvature of the earth's surface prevented our local Euclidean coordinates from remaining Euclidean as we moved away from the equator; here the curvature of spacetime prevents our local Lorentz coordinates from remaining Lorentz as we move away from our chosen world line.

Notice that our chosen world line is that of the spatial origin of our local Lorentz coordinates. Thus, we can think of those coordinates as provided by a spatially tiny latticework of rods and clocks, like that of Figure 24.1; and the latticework remains locally Lorentz for all time (as measured by its own clocks), but it ceases to be locally Lorentz when one moves a finite spatial distance (in its own frame) away from its spatial origin. (This is analogous to the local Euclidean coordinates on the Earth's equator: they remain Euclidean all along the equator [Eq. (24.10)], going all around the world, but they deviate from Euclidean when one moves away from the equator.)

This behavior is generic. One can show that, if any freely falling observer, anywhere in spacetime, sets up a little latticework of rods and clocks in accord with our standard rules and keeps the latticework's spatial origin on his or her free-fall world line, then the coordinates provided by the latticework will be locally Lorentz, with metric coefficients

$$\boxed{g_{\alpha\beta} = \begin{Bmatrix} \eta_{\alpha\beta} + \mathrm{O}\left(\frac{\delta_{jk}x^j x^k}{\mathcal{R}^2}\right) \\ \eta_{\alpha\beta} \quad \text{at spatial origin} \end{Bmatrix} \text{in a local Lorentz frame,}} \tag{24.15}$$

where $\mathcal{R}$ is the radius of curvature of spacetime. Notice that because the deviations of the metric from $\eta_{\alpha\beta}$ are second order in the distance from the spatial origin, the first derivatives of the metric coefficients are of first order, $g_{\alpha\beta,\mu} = \mathrm{O}(x^j/\mathcal{R}^2)$. This, plus the vanishing of the commutation coefficients in our coordinate basis, implies that the connection coefficients of the local Lorentz frame's coordinate basis are

$$\boxed{\Gamma^{\alpha}{}_{\beta\gamma} = \begin{Bmatrix} \mathrm{O}\left(\frac{\sqrt{\delta_{jk}x^j x^k}}{\mathcal{R}^2}\right) \\ 0 \quad \text{at spatial origin} \end{Bmatrix} \text{in a local Lorentz frame.}} \tag{24.16}$$

It is instructive to compare Eq. (24.15) for the metric in the local Lorentz frame of a freely falling observer in curved spacetime with Eq. (23.87) for the metric in the proper reference frame of an accelerated observer in flat spacetime. Whereas the spacetime curvature in (24.15) produces corrections to $g_{\alpha\beta} = \eta_{\alpha\beta}$ of *second* order in distance from the world line, the acceleration and spatial rotation of the reference frame in (23.87) produces corrections of *first* order. This remains true when one studies accelerated observers in curved spacetime (Chap. 25). In their proper reference frames the metric coefficients $g_{\alpha\beta}$ will contain both the first-order terms of (23.87) due to acceleration and rotation, and the second-order terms of (24.15) due to spacetime curvature.

****

EXERCISES

Exercise 24.2 *Derivation: Local Lorentz Frame in Friedman Universe*
By inserting the coordinate transformation (24.12) into the Friedman metric (24.11), derive the metric (24.13), (24.14) for a local Lorentz frame.

****

24.4 Free-fall Motion and Geodesics of Spacetime

In order to make more precise the concept of spacetime curvature, we will need to study quantitatively the relative acceleration of neighboring, freely falling particles.[1] Before we

[1] See MTW pp. 244–247, 312–324

can carry out such a study, however, we must understand quantitatively the motion of a single freely falling particle in curved spacetime. That is the objective of this section.

In a global Lorentz frame of flat, special relativistic spacetime a free particle moves along a straight world line, i.e., a world line with the form

$$(t, x, y, z) = (t_o, x_o, y_o, z_o) + (p^0, p^x, p^y, p^z)\zeta \ ; \quad \text{i.e., } x^\alpha = x^\alpha_o + p^\alpha \zeta \ . \tag{24.17}$$

Here p^α are the Lorentz-frame components of the particle's 4-momentum; ζ is the affine parameter such that $\vec{p} = d/d\zeta$, i.e., $p^\alpha = dx^\alpha/d\zeta$ [Eq. (1.18) *ff*]; and x^α_o are the coordinates of the particle when its affine parameter is $\zeta = 0$. The straight-line motion (24.17) can be described equally well by the statement that the Lorentz-frame components p^α of the particle's 4-momentum are constant, i.e., are independent of ζ

$$\frac{dp^\alpha}{d\zeta} = 0 \ . \tag{24.18}$$

Even nicer is the frame-independent description, which says that as the particle moves it parallel-transports its tangent vector $\vec{p}$ along its world line

$$\boxed{\nabla_{\vec{p}}\vec{p} = 0 \ , \quad \text{or, equivalently } p^\alpha{}_{;\beta} p^\beta = 0 \ .} \tag{24.19}$$

For a particle of nonzero rest mass m, which has $\vec{p} = m\vec{u}$ and $\zeta = \tau/m$ with $\vec{u} = d/d\tau$ its 4-velocity and τ its proper time, Eq. (24.19) is equivalent to $\nabla_{\vec{u}}\vec{u} = 0$. This is the form of the particle's law of motion discussed in Eq. (23.93).

This description of the motion is readily carried over into curved spacetime using the equivalence principle: Let $\mathcal{P}(\zeta)$ be the world line of a freely moving particle in curved spacetime. At a specific event $\mathcal{P}_o = \mathcal{P}(\zeta_o)$ on that world line introduce a local Lorentz frame (so the frame's spatial origin, like the particle, passes through $\mathcal{P}_o$ as time progresses). Then the equivalence principle tells us that the particle's law of motion must be the same in this local Lorentz frame as it is in the global Lorentz frame of special relativity:

$$\left(\frac{dp^\alpha}{d\zeta}\right)_{\zeta=\zeta_o} = 0 \ . \tag{24.20}$$

More powerful than this local-Lorentz-frame description of the motion is a description that is frame-independent. We can easily deduce such a description from Eq. (24.20). Since the connection coefficients vanish at the origin of the local Lorentz frame where (24.20) is being evaluated [cf. Eq. (24.16)], (24.20) can be written equally well, in our local Lorentz frame, as

$$0 = \left(\frac{dp^\alpha}{d\zeta} + \Gamma^\alpha{}_{\beta\gamma} p^\beta \frac{dx^\gamma}{d\zeta}\right)_{\zeta=\zeta_o} = \left((p^\alpha{}_{,\gamma} + \Gamma^\alpha{}_{\beta\gamma} p^\beta)\frac{dx^\gamma}{d\zeta}\right)_{\zeta=\zeta_o} = (p^\alpha{}_{;\gamma} p^\gamma)_{\zeta=\zeta_o} \ . \tag{24.21}$$

Thus, as the particle passes through the spatial origin of our local Lorentz coordinate system, the components of the directional derivative of its 4-momentum along itself vanish. Now, if two 4-vectors have components that are equal in one basis, their components are guaranteed [by the tensorial transformation law (23.19)] to be equal in all bases, and correspondingly

the two vectors, viewed as frame-independent, geometric objects, must be equal. Thus, since Eq. (24.21) says that the components of the 4-vector $\nabla_{\vec{p}}\vec{p}$ and the zero vector are equal in our chosen local Lorentz frame, it must be true that

$$\boxed{\nabla_{\vec{p}}\vec{p} = 0 \ .} \tag{24.22}$$

at the moment when the particle passes through the point $\mathcal{P}_o = \mathcal{P}(\zeta_o)$. Moreover, since $\mathcal{P}_o$ is an arbitrary point (event) along the particle's world line, it must be that (24.22) is a geometric, frame-independent *equation of motion* for the particle, valid everywhere along its world line. Notice that this geometric, frame-independent equation of motion $\nabla_{\vec{p}}\vec{p} = 0$ in curved spacetime is precisely the same as that [Eq. (24.19)] for flat spacetime. We shall generalize this conclusion to other laws of physics in Sec. 24.7 below.

Our equation of motion (24.22) for a freely moving point particle says, in words, that the particle *parallel transports* its 4-momentum along its world line. In any curved manifold, not just in spacetime, the relation $\vec{\nabla}_{\vec{p}}\vec{p} = 0$ is called the *geodesic equation*, and the curve to which $\vec{p}$ is the tangent vector is called a *geodesic*. On the surface of a sphere such as the earth, the geodesics are the great circles; they are the unique curves along which local Euclidean coordinates can be meshed, keeping one of the two Euclidean coordinates constant along the curve [cf. Eq. (24.10)], and they are the trajectories generated by an airplane's inertial guidance system, which tries to fly the plane along the straightest trajectory it can. Similarly, in spacetime the trajectories of freely falling particles are geodesics; they are the unique curves along which local Lorentz coordinates can be meshed, keeping the three spatial coordinates constant along the curve and letting the time vary, thereby producing a local Lorentz reference frame [Eqs. (24.15) and (24.16)], and they are also the spacetime trajectories along which inertial guidance systems will guide a spacecraft.

The geodesic equation guarantees that the square of the 4-momentum will be conserved along the particle's world line; in slot-naming index notation,

$$(g_{\alpha\beta}p^\alpha p^\beta)_{;\gamma}p^\gamma = 2g_{\alpha\beta}p^\alpha p^\beta{}_{;\gamma}p^\gamma = 0 \ . \tag{24.23}$$

(Here the standard rule for differentiating products has been used; this rule follows from the definition (23.27) of the frame-independent directional derivative of a tensor; it also can be deduced in a local Lorentz frame where $\Gamma^\alpha{}_{\mu\nu} = 0$ so each gradient with a ";" reduces to a partial derivative with a ",".) Also in Eq. (24.23) the term involving the gradient of the metric has been discarded since it vanishes [Eq. (23.40)], and the two terms involving derivatives of p^α and p^β, being equal, have been combined. In index-free notation the frame-independent relation (24.23) says

$$\nabla_{\vec{p}}(\vec{p}\cdot\vec{p}) = 2\vec{p}\cdot\nabla_{\vec{p}}\vec{p} = 0 \ . \tag{24.24}$$

This is a pleasing result, since the square of the 4-momentum is the negative of the particle's squared rest mass, $\vec{p}\cdot\vec{p} = -m^2$, which surely should be conserved along the particle's free-fall world line! Note that, as in flat spacetime, so also in curved, for a particle of finite rest mass the free-fall trajectory (the geodesic world line) is timelike, $\vec{p}\cdot\vec{p} = -m^2 < 0$, while for a zero-rest-mass particle it is null, $\vec{p}\cdot\vec{p} = 0$. Spacetime also supports spacelike geodesics, i.e.,

curves with tangent vectors $\vec{p}$ that satisfy the geodesic equation (24.22) and are spacelike, $\vec{p}\cdot\vec{p} > 0$. Such curves can be thought of as the world lines of freely falling "tachyons," i.e., faster-than-light particles—though it seems unlikely that such particles really exist in Nature. Note that the constancy of $\vec{p}\cdot\vec{p}$ along a geodesic implies that a geodesic can never change its character: if initially timelike, it will always remain timelike; if initially null, it will remain null; if initially spacelike, it will remain spacelike.

When studying the motion of a particle with finite rest mass, one often uses as the tangent vector to the geodesic the particle's 4-velocity $\vec{u} = \vec{p}/m$ rather than the 4-momentum, and correspondingly one uses as the parameter along the geodesic the particle's proper time $\tau = m\zeta$ rather than ζ (recall: $\vec{u} = d/d\tau$; $\vec{p} = d/d\zeta$). In this case the geodesic equation becomes

$$\boxed{\nabla_{\vec{u}}\vec{u} = 0\ ;} \tag{24.25}$$

cf. Eq. (23.93). Similarly, for spacelike geodesics, one often uses as the tangent vector $\vec{u} = d/ds$, where s is proper distance (square root of the invariant interval) along the geodesic; and the geodesic equation then assumes the same form (24.25) as for a timelike geodesic.

The geodesic world line of a freely moving particle has three very important properties:

(i) When written in a coordinate basis, the geodesic equation $\nabla_{\vec{p}}\vec{p} = 0$ becomes the following differential equation for the particle's world line $x^\alpha(\zeta)$ in the coordinate system [Ex. 24.3]

$$\boxed{\frac{d^2x^\alpha}{d\zeta^2} = -\Gamma^\alpha{}_{\mu\nu}\frac{dx^\mu}{d\zeta}\frac{dx^\nu}{d\zeta}\ .} \tag{24.26}$$

Here $\Gamma^\alpha{}_{\mu\nu}$ are the connection coefficients of the coordinate system's coordinate basis. [Equation (23.95) was a special case of this.] Note that these are four coupled equations ($\alpha = 0, 1, 2, 3$) for the four coordinates x^α as functions of affine parameter ζ along the geodesic. If the initial position, x^α at $\zeta = 0$, and initial tangent vector (particle momentum), $p^\alpha = dx^\alpha/d\zeta$ at $\zeta = 0$, are specified, then these four equations will determine uniquely the coordinates $x^\alpha(\zeta)$ as a function of ζ along the geodesic.

(ii) Consider a spacetime that possesses a symmetry, which is embodied in the fact that the metric coefficients in some coordinate system are independent of one of the coordinates x^A. Associated with that symmetry there will be a conserved quantity $p_A \equiv \vec{p}\cdot\partial/\partial x^A$ associated with free-particle motion. Exercise 24.4 derives this result and develops a familiar example.

(iii) Among all timelike curves linking two events $\mathcal{P}_0$ and $\mathcal{P}_1$ in spacetime, those whose proper time lapse (timelike length) is stationary under small variations of the curve are timelike geodesics; see Ex. 24.5. In other words, timelike geodesics are the curves that satisfy the action principle (24.30) below. Now, one can always send a photon from $\mathcal{P}_0$ to $\mathcal{P}_1$ by bouncing it off a set of strategically located mirrors, and that photon path is the limit of a timelike curve as the curve becomes null. Therefore, there exist timelike curves from $\mathcal{P}_0$ to $\mathcal{P}_1$ with vanishingly small length, so the geodesics cannot minimize the proper time lapse. This means that the curve of *maximal* proper time lapse (length) is a geodesic, and that any other geodesics will have a length that is a "saddle point" (stationary under variations of the path but not a maximum or a minimum).

****

EXERCISES

Exercise 24.3 *Derivation: Geodesic equation in an arbitrary coordinate system.*
Show that in an arbitrary coordinate system $x^\alpha(\mathcal{P})$ the geodesic equation (24.22) takes the form (24.26).

Exercise 24.4 *Derivation: Constant of Geodesic Motion in a Spacetime with Symmetry*

(a) Suppose that in some coordinate system the metric coefficients are independent of some specific coordinate x^A: $g_{\alpha\beta,A} = 0$. [Example: in spherical polar coordinates t, r, θ, ϕ in flat spacetime $g_{\alpha\beta,\phi} = 0$, so we could set $x^A = \phi$.] Show that

$$p_A \equiv \vec{p} \cdot \frac{\partial}{\partial x^A} \tag{24.27}$$

is a constant of the motion for a freely moving particle [$p_\phi =$ (conserved z-component of angular momentum) in above, spherically symmetric example]. [*Hint*: Show that the geodesic equation can be written in the form

$$\frac{dp_\alpha}{d\zeta} - \Gamma_{\mu\alpha\nu} p^\mu p^\nu = 0 \ , \tag{24.28}$$

where $\Gamma_{\mu\alpha\nu}$ is the covariant Christoffel symbol of Eqs. (23.38), (23.39).] Note the analogy of the constant of the motion p_A with Hamiltonian mechanics: there, if the Hamiltonian is independent of x^A then the generalized momentum p_A is conserved; here, if the metric coefficients are independent of x^A, then the covariant component p_A of the momentum is conserved. For an elucidation of the connection between these two conservation laws, see the Hamiltonian formulation of geodesic motion in Exercise 25.2 of MTW.

(b) As an example, consider a particle moving freely through a time-independent, Newtonian gravitational field. In Ex. 24.12 below we shall learn that such a gravitational field can be described in the language of general relativity by the spacetime metric

$$ds^2 = -(1 + 2\Phi)dt^2 + (\delta_{jk} + h_{jk})dx^j dx^k \ , \tag{24.29}$$

where $\Phi(x,\ y,\ z)$ is the time-independent Newtonian potential and h_{jk} are contributions to the metric that are independent of the time coordinate t and have magnitude of order $|\Phi|$. That the gravitational field is weak means $|\Phi| \ll 1$ (or, in cgs units, $|\Phi/c^2| \ll 1$). The coordinates being used are Lorentz, aside from tiny corrections of order $|\Phi|$; and, as this exercise and Ex. 24.12 show, they coincide with the coordinates of the Newtonian theory of gravity. Suppose that the particle has a velocity $v^j \equiv dx^j/dt$ through this coordinate system that is less than or of order $|\Phi|^{\frac{1}{2}}$ and thus small compared to the speed of light. Because the metric is independent of the time coordinate t, the component p_t of the particle's 4-momentum must be conserved

along its world line. Since, throughout physics, the conserved quantity associated with time-translation invariance is always the energy, we expect that p_t, when evaluated accurate to first order in $|\Phi|$, must be equal to the particle's conserved Newtonian energy, $E = m\Phi + \frac{1}{2}mv^j v^k \delta_{jk}$, aside from some multiplicative and additive constants. Show that this, indeed, is true, and evaluate the constants.

Exercise 24.5 *Problem: Action principle for geodesic motion*
Show, by introducing a specific but arbitrary coordinate system, that among all timelike world lines that a particle could take to get from event $\mathcal{P}_0$ to $\mathcal{P}_1$, the one or ones whose proper time lapse is stationary under small variations of path are the free-fall geodesics. In other words, an action principle for a timelike geodesic $\mathcal{P}(\lambda)$ [i.e., $x^\alpha(\lambda)$ in any coordinate system x^α] is

$$\boxed{\delta \int_{\mathcal{P}_0}^{\mathcal{P}_1} d\tau = \int_0^1 \left(g_{\alpha\beta} \frac{dx^\alpha}{d\lambda} \frac{dx^\beta}{d\lambda} \right)^{\frac{1}{2}} d\lambda = 0 \, ,} \tag{24.30}$$

where λ is an arbitrary parameter which, by definition, ranges from 0 at $\mathcal{P}_0$ to 1 at $\mathcal{P}_1$. [Note: unless, after the variation, you choose the arbitrary parameter λ to be "affine" ($\lambda = a\tau + b$ where a and b are constants), your equation for $d^2x^\alpha/d\lambda^2$ will not look quite like (24.26).]

****

24.5 Relative Acceleration, Tidal Gravity, and Spacetime Curvature

Now that we understand the motion of an individual freely falling particle in curved spacetime, we are ready to study the effects of gravity on the relative motions of such particles.[2] Before doing so in general relativity, let us recall the Newtonian discussion of the same problem:

24.5.1 Newtonian Description of Tidal Gravity

Consider, as shown in Fig. 24.3(a), two point particles, A and B, falling freely through 3-dimensional Euclidean space under the action of an external Newtonian potential Φ (i.e., a potential generated by other masses, not by the particles themselves). At Newtonian time $t = 0$ the particles are separated by only a small distance and are moving with the same velocity $\mathbf{v}_A = \mathbf{v}_B$. As time passes, however, the two particles, being at slightly different locations in space, experience slightly different gravitational potentials Φ and gravitational accelerations $\mathbf{g} = -\boldsymbol{\nabla}\Phi$ and thence develop slightly different velocities, $\mathbf{v}_A \neq \mathbf{v}_B$. To quantify this, denote by $\boldsymbol{\xi}$ the vector separation of the two particles in Euclidean 3-space. The components of $\boldsymbol{\xi}$ on any Euclidean basis [e.g., that of Fig. 24.3(a)] are $\xi^j = x^j_B - x^j_A$, where x^j_I is the coordinate location of particle I. Correspondingly, the rate of change of ξ^j

[2] See MTW pp. 29–37, 218–224, 265–275

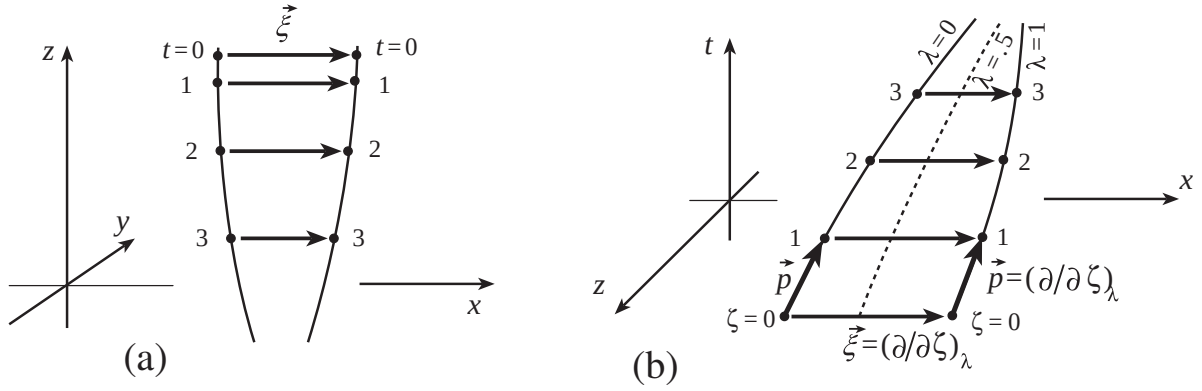


Fig. 24.3: The effects of tidal gravity on the relative motions of two freely falling particles. Diagram (a) depicts this in a Euclidean 3-space diagram using Newton's theory of gravity. Diagram (b) depicts it in a spacetime diagram using Einstein's theory of gravity, general relativity.

with respect to Newtonian time is $d\xi^j/dt = v_B^j - v_A^j$; i.e., the relative velocity of the two particles is the difference of their two velocities. The second time derivative of the relative separation, i.e., the relative acceleration of the two particles, is thus given by

$$\frac{d^2\xi^j}{dt^2} = \frac{d^2x_B^j}{dt^2} - \frac{d^2x_A^j}{dt^2} = -\left(\frac{\partial\Phi}{\partial x^j}\right)_B + \left(\frac{\partial\Phi}{\partial x^j}\right)_A = -\frac{\partial^2\Phi}{\partial x^j \partial x^k}\xi^k \ , \tag{24.31}$$

accurate to first order in the separation ξ^k. This equation gives the components of the relative acceleration in an arbitrary Euclidean basis. Rewritten in geometric, basis-independent language this equation says

$$\boxed{\frac{d^2\boldsymbol{\xi}}{dt^2} = -\mathcal{E}(\ldots, \boldsymbol{\xi}) \ ; \quad \text{i.e., } \frac{d^2\xi^j}{dt^2} = -\mathcal{E}^j{}_k\xi^k \ ,} \tag{24.32}$$

where $\mathcal{E}$ is a symmetric, second-rank tensor, called the *Newtonian tidal gravitational field*:

$$\boxed{\mathcal{E} = \boldsymbol{\nabla}\boldsymbol{\nabla}\Phi = -\boldsymbol{\nabla}\mathbf{g} \ ; \quad \text{i.e., } \mathcal{E}_{jk} = \frac{\partial^2\Phi}{\partial x^j \partial x^k} \ \text{ in Euclidean coordinates.}} \tag{24.33}$$

The name "tidal gravitational field" comes from the fact that this is the field which, generated by the moon and the sun, produces the tides on the earth's oceans. Note that, since this field is the gradient of the Newtonian gravitational acceleration $\mathbf{g}$, it is a quantitative measure of the inhomogeneities of Newtonian gravity.

Equation (24.31) shows quantitatively how the tidal gravitational field produces the relative acceleration of our two particles. As a specific application, one can use it to compute, in Newtonian theory, the relative accelerations and thence relative motions of two neighboring local Lorentz frames as they fall toward and through the center of the earth [Fig. 24.2(a) and associated discussion].

24.5.2 Relativistic Description of Tidal Gravity

Turn attention, now, to the general relativistic description of the relative motions of two free particles. As shown in Fig. 24.3(b), the particles, labeled A and B, move along geodesic world lines with affine parameters ζ and 4-momentum tangent vectors $\vec{p} = d/d\zeta$. The origins of ζ along the two world lines can be chosen however we wish, so long as events with the same ζ on the two world lines, $\mathcal{P}_A(\zeta)$ and $\mathcal{P}_B(\zeta)$ are close enough to each other that we can perform power-series expansions in their separation, $\vec{\xi}(\zeta) = \mathcal{P}_B(\zeta) - \mathcal{P}_A(\zeta)$, and keep only the leading terms. As in our Newtonian analysis, we require that the two particles initially have vanishing relative velocity, $\nabla_{\vec{p}}\vec{\xi} = 0$, and we shall compute the tidal-gravity-induced relative acceleration $\nabla_{\vec{p}}\nabla_{\vec{p}}\vec{\xi}$.

As a tool in our calculation, we shall introduce into spacetime a two-dimensional surface which contains our two geodesics A and B, and also contains an infinity of other geodesics in between and alongside them; and on that surface we shall introduce two coordinates, ζ =(affine parameter along each geodesic) and λ=(a parameter that labels the geodesics); see Fig. 24.3(b). Geodesic A will carry the label $\lambda = 0$; geodesic B will be $\lambda = 1$; $\vec{\xi} \equiv (\partial/\partial\lambda)_{\zeta=\text{const}}$ will be a vector field which, evaluated on geodesic A (i.e., at $\lambda = 0$), is equal to the separation vector we wish to study; and the vector field $\vec{p} = (\partial/\partial\zeta)_{\lambda=\text{const}}$ will be a vector field which, evaluated on any geodesic (A, B, or other curve of constant λ), is equal to the 4-momentum of the particle which moves along that geodesic. Our identification of $(\partial/\partial\lambda)_{\zeta=\text{const}}(\lambda = 0)$ with the separation vector $\vec{\xi}$ between geodesics A and B is the leading term in a power series expansion; it is here that we require, for good accuracy, that the geodesics be close together and be so parametrized that $\mathcal{P}_A(\zeta)$ is close to $\mathcal{P}_B(\zeta)$.

Our objective is to compute the relative acceleration of particles B and A, $\nabla_{\vec{p}}\nabla_{\vec{p}}\vec{\xi}$ evaluated at $\lambda = 0$. The quantity $\nabla_{\vec{p}}\vec{\xi}$, which we wish to differentiate a second time in that computation, is one of the terms in the following expression for the commutator of the vector fields $\vec{p}$ and $\vec{\xi}$ [Eq. (23.41)]:

$$\boxed{[\vec{p}, \vec{\xi}] = \nabla_{\vec{p}}\vec{\xi} - \nabla_{\vec{\xi}}\vec{p}\,.} \tag{24.34}$$

Because $\vec{p} = (\partial/\partial\zeta)_\lambda$ and $\vec{\xi} = (\partial/\partial\lambda)_\zeta$, these two vector fields commute, and Eq. (24.34) tells us that $\nabla_{\vec{p}}\vec{\xi} = \nabla_{\vec{\xi}}\vec{p}$. Correspondingly, the relative acceleration of our two particles can be expressed as

$$\nabla_{\vec{p}}\nabla_{\vec{p}}\vec{\xi} = \nabla_{\vec{p}}\nabla_{\vec{\xi}}\vec{p} = (\nabla_{\vec{p}}\nabla_{\vec{\xi}} - \nabla_{\vec{\xi}}\nabla_{\vec{p}})\vec{p}\,. \tag{24.35}$$

Here the second equality results from adding on, for use below, a term that vanishes because $\nabla_{\vec{p}}\vec{p} = 0$ (geodesic equation).

This first part of our calculation was performed efficiently using index-free notation. The next step will be easier if we introduce indices as names for slots. Then expression (24.35) takes the form

$$(\xi^\alpha{}_{;\beta}p^\beta)_{;\gamma}p^\gamma = (p^\alpha{}_{;\gamma}\xi^\gamma)_{;\delta}p^\delta - (p^\alpha{}_{;\gamma}p^\gamma)_{;\delta}\xi^\delta\,, \tag{24.36}$$

which can be evaluated by using the rule for differentiating products and then renaming indices and collecting terms; the result is

$$(\xi^\alpha{}_{;\beta}p^\beta)_{;\gamma}p^\gamma = (p^\alpha{}_{;\gamma\delta} - p^\alpha{}_{;\delta\gamma})\xi^\gamma p^\delta + p^\alpha{}_{;\gamma}(\xi^\gamma{}_{;\delta}p^\delta - p^\gamma{}_{;\delta}\xi^\delta)\,. \tag{24.37}$$

The second term in this expression vanishes, since it is just the commutator of $\vec{\xi}$ and $\vec{p}$ [Eq. (24.34)] written in slot-naming index notation, and as we noted above, $\vec{\xi}$ and $\vec{p}$ commute. The remaining first term,

$$(\xi^\alpha{}_{;\beta}p^\beta)_{;\gamma}p^\gamma = (p^\alpha{}_{;\gamma\delta} - p^\alpha{}_{;\delta\gamma})\xi^\gamma p^\delta \ , \tag{24.38}$$

reveals that *the relative acceleration of the two particles is caused by noncommutation of the two slots of a double gradient* (slots here named γ and δ). In the flat spacetime of special relativity the two slots would commute and there would be no relative acceleration. Spacetime curvature prevents them from commuting and thereby causes the relative acceleration.

Now, one can show that $p^\alpha{}_{;\gamma\delta} - p^\alpha{}_{;\delta\gamma}$ is linear in p^α; see Ex. 24.6. Therefore, there must exist a fourth rank tensor field $\mathbf{R}(\_,\_,\_,\_)$ such that

$$p^\alpha{}_{;\gamma\delta} - p^\alpha{}_{;\delta\gamma} = -R^\alpha{}_{\beta\gamma\delta}p^\beta \tag{24.39}$$

for any vector field $\vec{p}(\mathcal{P})$. The tensor $\mathbf{R}$ can be regarded as responsible for the failure of gradients to commute, so it must be some aspect of spacetime curvature. It is called the *Riemann curvature tensor*.

Inserting Eq. (24.39) into Eq. (24.38) and writing the result in both slot-naming index notation and abstract notation, we obtain

$$\boxed{(\xi^\alpha{}_{;\beta}p^\beta)_{;\gamma}p^\gamma = -R^\alpha{}_{\beta\gamma\delta}p^\beta\xi^\gamma p^\delta \ , \quad \nabla_{\vec{p}}\nabla_{\vec{p}}\vec{\xi} = -\mathbf{R}(\ldots,\vec{p},\vec{\xi},\vec{p}) \ .} \tag{24.40}$$

This is the *equation of relative acceleration* for freely moving test particles. It is also called the *equation of geodesic deviation*, because it describes the manner in which spacetime curvature $\mathbf{R}$ forces geodesics that are initially parallel (the world lines of freely moving particles with zero initial relative velocity) to deviate from each other; cf. Fig. 24.3(b).

24.5.3 Comparison of Newtonian and Relativistic Descriptions

It is instructive to compare this relativistic description of the relative acceleration of freely moving particles with the Newtonian description. For this purpose we shall consider a region of spacetime, such as our solar system, in which the Newtonian description of gravity is highly accurate; and there we shall study the relative acceleration of two free particles from the viewpoint of a local Lorentz frame in which the particles are both initially at rest.

In the Newtonian description, the transformation from a Newtonian universal reference frame (e.g., that of the center of mass of the solar system) to the chosen local Lorentz frame is achieved by introducing new Euclidean coordinates that are uniformly accelerated relative to the old ones, with just the right uniform acceleration to annul the gravitational acceleration at the center of the local Lorentz frame. This transformation adds a spatially homogeneous constant to the Newtonian acceleration $\mathbf{g} = -\boldsymbol{\nabla}\Phi$ but leaves unchanged the tidal field $E = \boldsymbol{\nabla}\boldsymbol{\nabla}\Phi$. Correspondingly, the Newtonian equation of relative acceleration in the local Lorentz frame retains its standard Newtonian form, $d^2\xi^j/dt^2 = -\mathcal{E}^j{}_k\xi^k$ [Eq. (24.32)], with the components of the tidal field computable equally well in the original universal reference frame, or in the local Lorentz frame, from the standard relation $\mathcal{E}^j{}_k = \mathcal{E}_{jk} = \partial^2\Phi/\partial x^j\partial x^k$.

As an aid in making contact between the relativistic and the Newtonian descriptions, we shall convert over from using the 4-momentum $\vec{p}$ as the tangent vector and ζ as the parameter along the particles' world lines to using the 4-velocity $\vec{u} = \vec{p}/m$ and the proper time $\tau = m\zeta$; this conversion brings the relativistic equation of relative acceleration (24.40) into the form

$$\nabla_{\vec{u}} \nabla_{\vec{u}} \vec{\xi} = -\mathbf{R}(\ldots, \vec{u}, \vec{\xi}, \vec{u}) \ . \tag{24.41}$$

Because the particles are (momentarily) at rest near the origin of the local Lorentz frame, their 4-velocities are $\vec{u} \equiv d/d\tau = \partial/\partial t$, which implies that the components of their 4-velocities are $u^0 = 1$, $u^j = 0$, and their proper times τ are equal to coordinate time t, which in turn coincides with the time t of the Newtonian analysis: $\tau = t$. In the relativistic analysis, as in the Newtonian, the separation vector $\vec{\xi}$ will have only spatial components, $\xi^0 = 0$ and $\xi^j \neq 0$. [If this were not so, we could make it so by a readjustment of the origin of proper time for particle B; cf. Fig. 24.3(b).] These facts, together with the vanishing of all the connection coefficients and derivatives of them ($\Gamma^j{}_{k0,0}$) that appear in $(\xi^j{}_{;\beta} u^\beta)_{;\gamma} u^\gamma$ at the origin of the local Lorentz frame [cf. Eqs. (24.15) and (24.16)], imply that the local Lorentz components of the equation of relative acceleration (24.41) take the form

$$\frac{d^2 \xi^j}{dt^2} = -R^j{}_{0k0} \xi^k \ . \tag{24.42}$$

By comparing this with the Newtonian equation of relative acceleration (24.32) we infer that, *in the Newtonian limit, in the local rest frame of the two particles,*

$$\boxed{R^j{}_{0k0} = \mathcal{E}_{jk} = \frac{\partial^2 \Phi}{\partial x^j \partial x^k} \ .} \tag{24.43}$$

Thus, *the Riemann curvature tensor is the relativistic generalization of the Newtonian tidal field*. This conclusion and the above equations make quantitative the statement that *gravity is a manifestation of spacetime curvature*.

Outside a spherical body with weak (Newtonian) gravity, such as the Earth, the Newtonian potential is $\Phi = -GM/r$, where G is Newton's gravitation constant, M is the body's mass and r is the distance from its center. If we introduce Cartesian coordinates with origin at the body's center and with z-axis through the point at which the Riemann tensor is to be measured, then Φ in these coordinates is $\Phi = -GM/(z^2 + x^2 + y^2)^{\frac{1}{2}}$, and on the z-axis the only nonzero $R^j{}_{0k0}$, as computed from Eq. (24.43), are

$$\boxed{R^z{}_{0z0} = \frac{-2GM}{r^3}, \quad R^x{}_{0x0} = R^y{}_{0y0} = \frac{+GM}{r^3} \ .} \tag{24.44}$$

Correspondingly, for two particles separated from each other in the radial (z) direction, the relative acceleration (24.42) is $d^2\xi^j/dt^2 = +(2GM/r^3)\xi^j$; i.e., the particles are pulled apart by the body's tidal gravitational field. Similarly, for two particles separated from each other in a horizontal direction (in the x–y plane), $d^2\xi^j/dt^2 = -(GM/r^3)\xi^j$; i.e., the particles are pushed together by the body's tidal gravitational field. There thus is a radial tidal stretch and a lateral tidal squeeze; and the lateral squeeze has half the strength of the radial stretch

but occurs in two laterial dimensions compared to the one radial dimension. These stretch and squeeze, produced by the sun and moon, are responsible for the tides on the earth's oceans.

****

EXERCISES

Exercise 24.6 *Derivation: Linearity of Commutator of Double Gradient*

(a) Let a and b be scalar fields with arbitrary but smooth dependence on location in spacetime, and $\vec{A}$ and $\vec{B}$ be tensor fields. Show that

$$(aA^\alpha + bB^\alpha)_{;\gamma\delta} - (aA^\alpha + bB^\alpha)_{;\delta\gamma} = a(A^\alpha{}_{;\gamma\delta} - A^\alpha{}_{;\delta\gamma}) + b(B^\alpha{}_{;\gamma\delta} - B^\alpha{}_{;\delta\gamma}) \ . \quad (24.45)$$

[*Hint*: The double gradient of a scalar field commutes, as one can easily see in a local Lorentz frame.]

(b) Use Eq. (24.45) to show that (i) the commutator of the double gradient is independent of how the differentiated vector field varies from point to point, and depends only on the value of the field at the location where the commutator is evaluated, and (ii) the commutator is linear in that value. Thereby conclude that there must exist a fourth rank tensor field $\mathbf{R}$ such that Eq. (24.39) is true for any vector field $\vec{p}$.

****

24.6 Properties of the Riemann Curvature Tensor

We now pause, in our study of the foundations of general relativity, to examine a few properties of the Riemann curvature tensor $\mathbf{R}$.[3]

We begin, as a tool for deriving other things, by evaluating the components of the Riemann tensor at the spatial origin of a local Lorentz frame; i.e. at a point where $\Gamma^\alpha{}_{\beta\gamma}$ vanishes but its derivatives do not. For any vector field $\vec{p}$ a straightforward computation reveals

$$p^\alpha{}_{;\gamma\delta} - p^\alpha{}_{;\delta\gamma} = (\Gamma^\alpha{}_{\beta\gamma,\delta} - \Gamma^\alpha{}_{\beta\delta,\gamma})p^\beta \ . \quad (24.46)$$

By comparing with Eq. (24.39), we can read off the local-Lorentz components of Riemann:

$$R^\alpha{}_{\beta\gamma\delta} = \Gamma^\alpha{}_{\beta\delta,\gamma} - \Gamma^\alpha{}_{\beta\gamma,\delta} \quad \text{at spatial origin of a local Lorentz frame.} \quad (24.47)$$

From this expression we infer that, at a spatial distance $\sqrt{\delta_{ij}x^i x^j}$ from the origin of a local Lorentz frame, the connection coefficients and the metric have magnitudes

$$\Gamma^\alpha{}_{\beta\gamma} = \mathrm{O}(R^\alpha{}_{\mu\nu\lambda}\sqrt{\delta_{ij}x^i x^j}) \ , \quad g_{\alpha\beta} - \eta_{\alpha\beta} = \mathrm{O}(R^\mu{}_{\nu\lambda\rho}\delta_{ij}x^i x^j) \ , \quad \text{in a local Lorentz frame.} \quad (24.48)$$

[3] See MTW pp. 273–288, 324–327.

Comparison with Eqs. (24.15) and (24.16) shows that the radius of curvature of spacetime (a concept defined only semiquantitatively) is of order the inverse square root of the components of the Riemann tensor in a local Lorentz frame:

$$\boxed{\mathcal{R} = \mathrm{O}\left(\frac{1}{|R^{\alpha}{}_{\beta\gamma\delta}|^{\frac{1}{2}}}\right) \quad \text{in a local Lorentz frame.}} \tag{24.49}$$

By comparison with Eq. (24.44), we see that at radius r outside a weakly gravitating body of mass M, the radius of curvature of spacetime is

$$\boxed{\mathcal{R} \sim \left(\frac{r^3}{GM}\right)^{\frac{1}{2}} = \left(\frac{c^2 r^3}{GM}\right)^{\frac{1}{2}} ,} \tag{24.50}$$

where the factor c in the second expression makes the formula valid in conventional units. For further discussion see Ex. 24.7.

From the components (24.47) of the Riemann tensor in a local Lorentz frame, together with the vanishing of the connection coefficients at the origin and the standard expressions (10.18), (10.19) for the connection coefficients in terms of the metric components, one easily can show that

$$\boxed{R_{\alpha\beta\gamma\delta} = \frac{1}{2}(g_{\alpha\delta,\beta\gamma} + g_{\beta\gamma,\alpha\delta} - g_{\alpha\gamma,\beta\delta} - g_{\beta\delta,\alpha\gamma}) \quad \text{in a local Lorentz frame.}} \tag{24.51}$$

From these expressions, plus the commutation of partial derivatives $g_{\alpha\gamma,\beta\delta} = g_{\alpha\gamma,\delta\beta}$ and the symmetry of the metric one easily can show that in a local Lorentz frame the components of the Riemann tensor have the following symmetries:

$$\boxed{R_{\alpha\beta\gamma\delta} = -R_{\beta\alpha\gamma\delta} \ , \quad R_{\alpha\beta\gamma\delta} = -R_{\alpha\beta\delta\gamma} \ , \quad R_{\alpha\beta\gamma\delta} = +R_{\gamma\delta\alpha\beta}} \tag{24.52}$$

(antisymmetry in first pair of indices, antisymmetry in second pair of indices, and symmetry under interchange of the pairs). When one computes the value of the tensor on four vectors, $\mathbf{R}(\vec{A}, \vec{B}, \vec{C}, \vec{D})$ using component calculations in this frame, one trivially sees that these symmetries produce corresponding symmetries under interchange of the vectors inserted into the slots, and thence under interchange of the slots themselves. This is always the case: any symmetry that the components of a tensor exhibit in a special basis will induce the same symmetry on the slots of the geometric, frame-independent tensor. The resulting symmetries for $\mathbf{R}$ are given by Eq. (24.52) with the "Escher mind-flip" [Sec. 1.5.3] in which the indices switch from naming components in a special frame to naming slots: *The Riemann tensor is antisymmetric under interchange of its first two slots, antisymmetric under interchange of the last two, and symmetric under interchange of the two pairs.*

One additional symmetry can be verified, by calculation in the local Lorentz frame [i.e., from Eq. (24.51)]:

$$\boxed{R_{\alpha\beta\gamma\delta} + R_{\alpha\gamma\delta\beta} + R_{\alpha\delta\beta\gamma} = 0 \ .} \tag{24.53}$$

(Note that this cyclic symmetry is the same as occurs in the Maxwell equations (23.74) or (1.66), and also the same as occurs in the commutator identities $\left[\vec{B},[\vec{C},\vec{D}\,]\right]+\left[\vec{C},[\vec{D},\vec{B}\,]\right]+\left[\vec{D},[\vec{B},\vec{C}\,]\right]=0$.) One can show that the full set of symmetries (24.52) and (24.53) reduces the number of independent components of the Riemann tensor, in 4-dimensional spacetime, from $4^4=256$ to "just" 20.

Of these 20 independent components, 10 are contained in the *Ricci curvature tensor*—which is the contraction of the Riemann tensor on its first and third slots

$$\boxed{R_{\alpha\beta} \equiv R^{\mu}{}_{\alpha\mu\beta}\ ,} \tag{24.54}$$

and which by the symmetries (24.52) and (24.53) of Riemann is itself symmetric

$$\boxed{R_{\alpha\beta}=R_{\beta\alpha}\ .} \tag{24.55}$$

The other 10 independent components of Riemann are contained in the Weyl curvature tensor, which we will not study here; see, e.g., pp. 325 and 327 of MTW. The contraction of the Ricci tensor on its two slots,

$$\boxed{R\equiv R^{\alpha}{}_{\alpha}\ ,} \tag{24.56}$$

is called the *curvature scalar*.

One often needs to know the components of the Riemann curvature tensor in some non-local-Lorentz basis. Exercise 24.8 derives the following equation for them in an arbitrary basis:

$$\boxed{R^{\alpha}{}_{\beta\gamma\delta}=\Gamma^{\alpha}{}_{\beta\delta,\gamma}-\Gamma^{\alpha}{}_{\beta\gamma,\delta}+\Gamma^{\alpha}{}_{\mu\gamma}\Gamma^{\mu}{}_{\beta\delta}-\Gamma^{\alpha}{}_{\mu\delta}\Gamma^{\mu}{}_{\beta\gamma}-\Gamma^{\alpha}{}_{\beta\mu}c_{\gamma\delta}{}^{\mu}\ .} \tag{24.57}$$

Here $\Gamma^{\alpha}{}_{\beta\gamma}$ are the connection coefficients in the chosen basis, $\Gamma^{\alpha}{}_{\beta\gamma,\delta}$ is the result of letting the basis vector $\vec{e}_\delta$ act as a differential operator on $\Gamma^{\alpha}{}_{\beta\gamma}$, as though $\Gamma^{\alpha}{}_{\beta\gamma}$ were a scalar, and $c_{\gamma\delta}{}^{\mu}$ are the basis vectors' commutation coefficients. Calculations with this equation are usually very long and tedious, and so are carried out using symbolic-manipulation software on a computer.

****

EXERCISES

Exercise 24.7 *Example: Orders of magnitude of the radius of curvature of spacetime*
With the help of the Newtonian limit (24.43) of the Riemann curvature tensor, show that near the earth's surface the radius of curvature of spacetime has a magnitude $\mathcal{R}\sim$ (1 astronomical unit) $\equiv$ (distance from sun to earth). What is the radius of curvature of spacetime near the sun's surface? near the surface of a white-dwarf star? near the surface of a neutron star? near the surface of a one-solar-mass black hole? in intergalactic space?

Exercise 24.8 *Derivation: Components of Riemann tensor in an arbitrary basis*
By evaluating expression (24.39) in an arbitrary basis (which might not even be a coordinate basis), derive Eq. (24.57) for the components of the Riemann tensor. In your derivation

keep in mind that commas denote partial derivations *only* in a coordinate basis; in an arbitrary basis they denote the result of letting a basis vector act as a differential operator; cf. Eq. (23.32).

Exercise 24.9 *Problem: Curvature of the surface of a sphere*
On the surface of a sphere such as the earth introduce spherical polar coordinates in which the metric, written as a line element, takes the form

$$ds^2 = a^2(d\theta^2 + \sin^2\theta d\phi^2) \ , \tag{24.58}$$

where a is the sphere's radius.

(a) Show (first by hand and then by computer) that the connection coefficients for the coordinate basis $\{\partial/\partial\theta, \partial/\partial\phi\}$ are

$$\Gamma^{\theta}{}_{\phi\phi} = -\sin\theta\cos\theta \ , \quad \Gamma^{\phi}{}_{\theta\phi} = \Gamma^{\phi}{}_{\phi\theta} = \cot\theta \ , \quad \text{all others vanish.} \tag{24.59}$$

(b) Show that the symmetries (24.52) and (24.53) of the Riemann tensor guarantee that its only nonzero components in the above coordinate basis are

$$R_{\theta\phi\theta\phi} = R_{\phi\theta\phi\theta} = -R_{\theta\phi\phi\theta} = -R_{\phi\theta\theta\phi} \ . \tag{24.60}$$

(c) Show, first by hand and then by computer, that

$$R_{\theta\phi\theta\phi} = a^2\sin^2\theta \ . \tag{24.61}$$

(d) Show that in the basis

$$\{\vec{e}_{\hat\theta}, \vec{e}_{\hat\phi}\} = \left\{\frac{1}{a}\frac{\partial}{\partial\theta}\,, \frac{1}{a\sin\theta}\frac{\partial}{\partial\phi}\right\} \ , \tag{24.62}$$

the components of the metric, the Riemann tensor, the Ricci tensor, and the curvature scalar are

$$g_{\hat j\hat k} = \delta_{jk} \ , \quad R_{\hat\theta\hat\phi\hat\theta\hat\phi} = \frac{1}{a^2}, \quad R_{\hat j\hat k} = \frac{1}{a^2}g_{\hat j\hat k} \ , \quad R = \frac{2}{a^2} \ . \tag{24.63}$$

The first of these implies that the basis is orthonormal; the rest imply that the curvature is independent of location on the sphere, as it should be by spherical symmetry. [The θ-dependence in the coordinate components of Riemann, Eq. (24.61), like the θ-dependence in the metric component $g_{\phi\phi}$, is a result of the θ-dependence in the length of the coordinate basis vector $\vec{e}_\phi$: $|\vec{e}_\phi| = a\sin\theta$.)

Exercise 24.10 *Problem: Geodesic Deviation on a Sphere*
Consider two neighboring geodesics (great circles) on a sphere of radius a, one the equator and the other a geodesic slightly displaced from the equator (by $\Delta\theta = b$) and parallel to it at $\phi = 0$. Let $\vec{\xi}$ be the separation vector between the two geodesics, and note that at $\phi = 0$, $\vec{\xi} = b\partial/\partial\theta$. Let l be proper distance along the equatorial geodesic, so $d/dl = \vec{u}$ is its tangent vector.

(a) Show that $l = a\phi$ along the equatorial geodesic.

(b) Show that the equation of geodesic deviation (24.40) reduces to

$$\frac{d^2\xi^\theta}{d\phi^2} = -\xi^\theta \ , \quad \frac{d^2\xi^\phi}{d\phi^2} = 0 \ . \tag{24.64}$$

(c) Solve this, subject to the above initial conditions, to obtain

$$\xi^\theta = b\cos\phi \ , \quad \xi^\phi = 0 \ . \tag{24.65}$$

Verify, by drawing a picture, that this is precisely what one would expect for the separation vector between two great circles.

****

24.7 Curvature Coupling Delicacies in the Equivalence Principle, and Some Nongravitational Laws of Physics in Curved Spacetime[4]

If one knows a local, special relativistic, nongravitational law of physics in geometric, frame-independent form [for example, the expression for the stress-energy tensor of a perfect fluid in terms of its 4-velocity $\vec{u}$ and its rest-frame mass-energy density ρ and pressure P

$$\mathbf{T} = (\rho + P)\vec{u}\otimes\vec{u} + P\mathbf{g} \tag{24.66}$$

Eq. (23.59)], then the equivalence principle guarantees that in general relativity the law will assume the same geometric, frame-independent form. One can see that this is so by the same method as we used to derive the general relativistic equation of motion $\nabla_{\vec{p}}\vec{p} = 0$ for free particles [Eq. (24.22) and associated discussion]: (*i*) rewrite the special relativistic law in terms of components in a global Lorentz frame [$T^{\alpha\beta} = (\rho + P)u^\alpha u^\beta + Pg^{\alpha\beta}$], (*ii*) then infer from the equivalence principle that this same component form of the law will hold, unchanged, in a local Lorentz frame in general relativity, and (*iii*) then deduce that this component law is the local Lorentz frame version of the original geometric law [$\mathbf{T} = (\rho + P)\vec{u}\otimes\vec{u} + P\mathbf{g}$], now lifted into general relativity.

Thus, *when the local, nongravitational laws of physics are known in frame-independent form, one need not distinguish between whether they are special relativistic or general relativistic*.

In this conclusion the word *local* is crucial: The equivalence principle is strictly valid only at the spatial origin of a local Lorentz frame; and, correspondingly, it is in danger of failure for any law of physics that cannot be formulated solely in terms of quantities which reside at the

[4]See MTW Chap. 16.

spatial origin—i.e., along a timelike geodesic. For the above example, $\mathbf{T} = (\rho+P)\vec{u}\otimes\vec{u}+P\mathbf{g}$, there is no problem; and for the local law of conservation of 4-momentum $\vec{\nabla}\cdot\mathbf{T} = 0$ there is no problem. However, for the global law of conservation of 4-momentum

$$\int_{\partial\mathcal{V}} T^{\alpha\beta} d\Sigma_\beta = 0 \tag{24.67}$$

[Eq. (1.96) and Fig. 1.16], there is serious trouble: This law is severely nonlocal, since it involves integration over a finite, closed 3-surface $\partial\mathcal{V}$ in spacetime. Thus, the equivalence principle fails for it. The failure shows up especially clearly when one notices (as we discussed in Sec. 23.3.4) that the quantity $T^{\alpha\beta}d\Sigma_\beta$ which the integral is trying to add up over $\partial\mathcal{V}$ has one empty slot, named α; i.e., it is a vector. This means that to compute the integral (24.67) we must transport the contributions $T^{\alpha\beta}d\Sigma_\beta$ from the various tangent spaces in which they normally live, to the tangent space of some single, agreed upon location, where they are to be added. By what rule should the transport be done? In special relativity one uses parallel transport, so the components of the vector are held fixed in any global Lorentz frame. However, it turns out that spacetime curvature makes parallel transport dependent on the path of the transport (and correspondingly, a vector is changed by parallel transport around a closed curve). As a result, the integral $\int_{\partial\mathcal{V}} T^{\alpha\beta}d\Sigma_\beta$ depends not only on the common location to which one transports each surface element's contribution in order to add them, it also depends on the path of the transport, which in general is quite arbitrary. This dependence makes the integral ill defined and correspondingly causes a breakdown, in general relativity, in the global law of 4-momentum conservation.

Another instructive example is the law by which a freely moving particle transports its spin angular momentum. The spin angular momentum is readily defined in the momentary local Lorentz rest frame of the particle's center of mass; there it is a 4-vector with vanishing time component, and with space components given by the familiar integral

$$S_i = \int_{\text{interior of body}} \epsilon_{ijk} x^j T^{k0} dx\, dy\, dz \ , \tag{24.68}$$

where T^{k0} are the components of the momentum density. In special relativity the law of angular momentum conservation (e.g., MTW Sec. 5.11) guarantees that the Lorentz-frame components S^α of this spin angular momentum remain constant, so long as no external torques act on the particle. This conservation law can be written in special relativistic, frame-independent notation, as Eq. (23.91), specialized to a non-accelerated particle:

$$\nabla_{\vec{u}}\vec{S} = 0 \ ; \tag{24.69}$$

i.e., the spin vector $\vec{S}$ is parallel transported along the world line of the particle (which has 4-velocity $\vec{u}$). If this were a *local* law of physics, it would take this same form, unchanged, in general relativity, i.e., in curved spacetime. Whether the law is local or not depends, clearly, on the size of the particle. If the particle is vanishingly small in its own rest frame, then the law is local and (24.69) will be valid in general relativity. However, if the particle has finite size, the law (24.69) is in danger of failing—and, indeed it does fail if the particle's finite size is accompanied by a finite quadrupole moment. In that case, the coupling of the quadrupole

moment $\mathcal{I}_{\alpha\beta}$ to the curvature of spacetime $R^{\alpha}{}_{\beta\gamma\delta}$ produces a torque on the "particle", causing a breakdown in (24.69):

$$\boxed{S^{\alpha}{}_{;\mu}u^{\mu} = \epsilon^{\alpha\beta\gamma\delta}\mathcal{I}_{\beta\mu}R^{\mu}{}_{\nu\gamma\zeta}u_{\delta}u^{\nu}u^{\zeta}\ .} \tag{24.70}$$

The earth is a good example: the Riemann tensor $R^{\alpha}{}_{\beta\gamma\delta}$ produced at earth by the moon and sun couples to the earth's centrifugal-flattening-induced quadrupole moment $\mathcal{I}_{\mu\nu}$; and the resulting torque (24.70) causes the earth's spin axis to precess relative to the distant stars, with a precession period of 26,000 years—sufficiently fast to show up clearly in historical records as well as in modern astronomical measurements. For details see, e.g., Ex. 16.4 of MTW.

This example illustrates the fact that, if a small amount of nonlocality is present in a physical law, then when lifted from special relativity into general relativity, the law will acquire a small *curvature-coupling* modification.

What is the minimum amount of nonlocality that can produce curvature-coupling modifications in physical laws? As a rough rule of thumb, the minimum amount is double gradients: Because the connection coefficients vanish at the origin of a local Lorentz frame, the local Lorentz components of a single gradient are the same as the components in a global Lorentz frame, e.g., $A^{\alpha}{}_{;\beta} = \partial A^{\alpha}/\partial x^{\beta}$. However, because spacetime curvature prevents the spatial derivatives of the connection coefficients from vanishing at the origin of a local Lorentz frame, any law that involves double gradients is in danger of acquiring curvature-coupling corrections when lifted into general relativity. As an example, it turns out that the wave equation for the electromagnetic vector 4-potential, which in Lorenz gauge takes the form $A^{\alpha;\mu}{}_{\mu} = 0$ in flat spacetime, becomes in curved spacetime

$$\boxed{A^{\alpha;\mu}{}_{\mu} = R^{\alpha\mu}A_{\mu}\ ,} \tag{24.71}$$

where $R^{\alpha\mu}$ is the Ricci curvature tensor; see Ex. 24.11 below. [Note: in Eq. (24.71), and always, all indices that follow the semicolon represent differentiation slots; i.e., $A^{\alpha;\mu}{}_{\mu} \equiv A^{\alpha;\mu}{}_{;\mu}$.]

The curvature-coupling ambiguities that occur when one lifts slightly nonlocal laws from special relativity into general relativity using the equivalence principle are very similar to "factor-ordering ambiguities" that occur when one lifts a Hamiltonian into quantum mechanics from classical mechanics using the correspondence principle. In the equivalence principle the curvature coupling can be regarded as due to the fact that double gradients, which commute in special relativity, do not commute in general relativity. In the correspondence principle the factor ordering difficulties result from the fact that quantities that commute classically [e.g., position x and momentum p] do not commute quantum mechanically [$\hat{x}\hat{p} \neq \hat{p}\hat{x}$], so when the products of such quantities appear in a classical Hamiltonian one does not know their correct order in the quantum Hamiltonian [does xp become $\hat{x}\hat{p}$, or $\hat{p}\hat{x}$, or $\frac{1}{2}(\hat{x}\hat{p}+\hat{p}\hat{x})$?].

****

EXERCISES

Exercise 24.11 *Example: Curvature coupling in electromagnetic wave equation*

Since the Maxwell equations, written in terms of the classically measureable electromagnetic field tensor **F** [Eqs. (23.74) or (1.66)], involve only single gradients, it is reasonable to expect them to be lifted into curved spacetime without curvature-coupling additions. Assume that this is true.

It can be shown that: (*i*) if one writes the electromagnetic field tensor **F** in terms of a 4-vector potential $\vec{A}$ as

$$F_{\alpha\beta} = A_{\beta;\alpha} - A_{\alpha;\beta} \ , \tag{24.72}$$

then half of the curved-spacetime Maxwell equations, $F_{\alpha\beta;\gamma} + F_{\beta\gamma;\alpha} + F_{\gamma\alpha;\beta} = 0$ [Eqs. (23.74)] are automatically satisfied; (*ii*) **F** is unchanged by gauge transformations in which a gradient is added to the vector potential, $\vec{A} \to \vec{A} + \vec{\nabla}\psi$; and (*iii*) by such a gauge transformation one can impose the Lorentz-gauge condition $\vec{\nabla} \cdot \vec{A} = 0$ on the vector potential.

Show that, when the charge-current 4-vector vanishes, $\vec{J} = 0$, the other half of the Maxwell equations, $F^{\alpha\beta}{}_{;\beta} = 0$ [Eqs. (23.74)] become, in Lorenz gauge and in curved spacetime, the wave equation with curvature coupling, Eq. (24.71).

****

24.8 The Einstein Field Equation[5]

One crucial issue remains to be studied in this overview of the foundations of general relativity: What is the physical law that determines the curvature of spacetime? Einstein's search for that law, his *Einstein field equation*, occupied a large fraction of his efforts during the years 1913, 1914, and 1915. Several times he thought he had found it, but each time his proposed law turned out to be fatally flawed; for some flavor of his struggle see the excerpts from his writings in Sec. 17.7 of MTW.

In this section we shall briefly examine one segment of Einstein's route toward his field equation: the segment motivated by contact with Newtonian gravity.

The Newtonian potential Φ is a close analog of the general relativistic spacetime metric **g**: From Φ we can deduce everything about Newtonian gravity, and from **g** we can deduce everything about spacetime curvature. In particular, by differentiating Φ twice we can obtain the Newtonian tidal field E [Eq. (24.33)], and by differentiating the components of **g** twice we can obtain the components of the relativistic generalization of E: the components of the Riemann curvature tensor $R^{\alpha}{}_{\beta\gamma\delta}$ [Eq. (24.51) in a local Lorentz frame; Eq. (24.57) in an arbitrary basis].

In Newtonian gravity Φ is determined by Newton's field equation

$$\boxed{\boldsymbol{\nabla}^2\Phi = 4\pi G\rho \ ,} \tag{24.73}$$

which can be rewritten in terms of the tidal field $\mathcal{E}_{jk} = \partial^2\Phi/\partial x^j \partial x^k$ as

$$\mathcal{E}^j{}_j = 4\pi G\rho \ . \tag{24.74}$$

[5] See MTW Chap. 17.

Note that this equates a piece of the tidal field, its trace, to the density of mass. By analogy we can expect the Einstein field equation to equate a piece of the Riemann curvature tensor (the analog of the Newtonian tidal field) to some tensor analog of the Newtonian mass density. Further guidance comes from the demand that in nearly Newtonian situations, e.g., in the solar system, the Einstein field equation should reduce to Newton's field equation. To exploit that guidance, we can (*i*) write the Newtonian tidal field for nearly Newtonian situations in terms of general relativity's Riemann tensor, $\mathcal{E}_{jk} = R_{j0k0}$ [Eq. (24.43); valid in a local Lorentz frame], (*ii*) then take the trace and note that by its symmetries $R^0{}_{000} = 0$ so that $\mathcal{E}^j{}_j = R^\alpha{}_{0\alpha0} = R_{00}$, and (*iii*) thereby infer that the Newtonian limit of the Einstein equation should read, in a local Lorentz frame,

$$R_{00} = 4\pi G\rho \ . \tag{24.75}$$

Here R_{00} is the time-time component of the Ricci curvature tensor—which can be regarded as a piece of the Riemann tensor. An attractive proposal for the Einstein field equation should now be obvious: Since the equation should be geometric and frame-independent, and since it must have the Newtonian limit (24.75), it presumably should say $R_{\alpha\beta} = 4\pi G\times$ (a second-rank symmetric tensor that generalizes the Newtonian mass density ρ). The obvious required generalization of ρ is the stress-energy tensor $T_{\alpha\beta}$, so

$$R_{\alpha\beta} = 4\pi G T_{\alpha\beta} \ . \tag{24.76}$$

Einstein flirted extensively with this proposal for the field equation during 1913–1915. However, it, like several others he studied, was fatally flawed. When expressed in a coordinate system in terms of derivatives of the metric components $g_{\mu\nu}$, it becomes (because $R_{\alpha\beta}$ and $T_{\alpha\beta}$ both have ten independent components) ten independent differential equations for the ten $g_{\mu\nu}$. This is too many equations: By an arbitrary change of coordinates, $x^\alpha_{\rm new} = F^\alpha(x^0_{\rm old}, x^1_{\rm old}, x^2_{\rm old}, x^3_{\rm old})$ involving four arbitrary functions F^0, F^1, F^2, F^3, one should be able to impose on the metric components four arbitrary conditions, analogous to gauge conditions in electromagnetism (for example, one should be able to set $g_{00} = -1$ and $g_{0j} = 0$ everywhere); and correspondingly, the field equations should constrain only six, not ten of the components of the metric (the six g_{ij} in our example).

In November 1915 Einstein (1915), and independently Hilbert (1915) [who was familiar with Einstein's struggle as a result of private conversations and correspondence] discovered the resolution of this dilemma: Because the local law of 4-momentum conservation guarantees $T^{\alpha\beta}{}_{;\beta} = 0$ independent of the field equation, if we replace the Ricci tensor in (24.76) by a constant (to be determined) times some new curvature tensor $G^{\alpha\beta}$ that is also automatically divergence free independent of the field equation ($G^{\alpha\beta}{}_{;\beta} \equiv 0$), then the new field equation $G^{\alpha\beta} = \kappa T^{\alpha\beta}$ (with κ = constant) will not constrain all ten components of the metric. Rather, in a coordinate system the four equations $[G^{\alpha\beta} - \kappa T^{\alpha\beta}]_{;\beta} = 0$ with $\alpha = 0, 1, 2, 3$ will automatically be satisfied; they will not constrain the metric components in any way, and there will remain in the field equation only six independent constraints on the metric components, precisely the desired number.

It turns out, in fact, that from the Ricci tensor and the scalar curvature one can construct

a curvature tensor $G^{\alpha\beta}$ with the desired property:

$$\boxed{G^{\alpha\beta} \equiv R^{\alpha\beta} - \frac{1}{2} R g^{\alpha\beta} \ .} \tag{24.77}$$

Today we call this the *Einstein curvature tensor*. That it has vanishing divergence, independently of how one chooses the metric,

$$\boxed{\vec{\nabla} \cdot \mathbf{G} \equiv 0 \ ,} \tag{24.78}$$

is called the *contracted Bianchi identity* , since it can be obtained by contracting the following *Bianchi identity* on the tensor $\epsilon_\alpha{}^{\beta\mu\nu}\epsilon_\nu{}^{\gamma\delta\epsilon}$ (Sec. 13.5 of MTW):

$$\boxed{R^\alpha{}_{\beta\gamma\delta;\epsilon} + R^\alpha{}_{\beta\delta\epsilon;\gamma} + R^\alpha{}_{\beta\epsilon\gamma;\delta} = 0 \ .} \tag{24.79}$$

[This Bianchi identity holds true for the Riemann curvature tensor of any and every "manifold", i.e. of any and every smooth space; it is derived most easily by introducing a local Lorentz frame, by showing from (24.57) that in such a frame the components $R_{\alpha\beta\gamma\delta}$ of Riemann have the form (24.51) plus corrections that are quadratic in the distance from the origin, by then computing the left side of (24.79), with index α down, at the origin of that frame and showing it is zero, and by then arguing that because the origin of the frame was an arbitrary event in spacetime, and because the left side of (24.79) is the component of a tensor, the left side viewed as a frame-independent geometric object must vanish at all events in the manifold. For an extensive discussion of the Bianchi identities (24.79) and (24.78) see, e.g., Chap. 15 of MTW.]

The Einstein field equation, then, should equate a multiple of $T^{\alpha\beta}$ to the Einstein tensor $G^{\alpha\beta}$:

$$G^{\alpha\beta} = \kappa T^{\alpha\beta} \ . \tag{24.80}$$

The proportionality factor κ is determined from the Newtonian limit: By rewriting the field equation (24.80) in terms of the Ricci tensor

$$R^{\alpha\beta} - \frac{1}{2} g^{\alpha\beta} R = \kappa T^{\alpha\beta} \ , \tag{24.81}$$

then taking the trace to obtain $R = -\kappa g_{\mu\nu} T^{\mu\nu}$, then inserting this back into (24.81), we obtain

$$R^{\alpha\beta} = \kappa (T^{\alpha\beta} - \frac{1}{2} g^{\alpha\beta} g_{\mu\nu} T^{\mu\nu}) \ . \tag{24.82}$$

In nearly Newtonian situations and in a local Lorentz frame, the mass-energy density $T^{00} \cong \rho$ is far greater than the momentum density T^{j0} and also far greater than the stress T^{jk}; and correspondingly, the time-time component of the field equation (24.82) becomes

$$R^{00} = \kappa (T^{00} - \frac{1}{2} \eta^{00} \eta_{00} T^{00}) = \frac{1}{2} \kappa T^{00} = \frac{1}{2} \kappa \rho \ . \tag{24.83}$$

By comparing with the correct Newtonian limit (24.75) and noting that in a local Lorentz frame $R_{00} = R^{00}$, we see that

$$\kappa = 8\pi G \ . \tag{24.84}$$

Quantity	Conventional Units	Geometrized Units
speed of light, c	2.998×10^8 m sec^{-1}	one
Newton's gravitation constant, G	6.673×10^{-11}m^3 kg^{-1} sec^{-2}	one
G/c^2	7.425×10^{-28} m kg^{-1}	one
c^5/G	3.629×10^{52} W	one
$c^2/\sqrt{G}$	3.479×10^{24} gauss cm $= 1.160 \times 10^{24}$ volts	one
Planck's reduced constant $\hbar$	1.055×10^{-34} kg m^{2}s^{-1}	$(1.616 \times 10^{-35}$ m$)^2$
sun's mass, $M_\odot$	1.989×10^{30} kg	1.477 km
sun's radius, $R_\odot$	6.960×10^8 m	6.960×10^8 m
earth's mass, $M_\oplus$	5.977×10^{24} kg	4.438 mm
earth's radius, $R_\oplus$	6.371×10^6 m	6.371×10^6 m
Hubble constant H_o	65 ± 25 km sec^{-1}Mpc^{-1}	$[(12 \pm 5) \times 10^9$ lt yr$]^{-1}$
density to close universe, $\rho_{\rm crit}$	$9^{+11}_{-5} \times 10^{-27}$ kg m^{-3}	$7^{+8}_{-3} \times 10^{-54}$ m^{-2}

Table 24.1: Some useful quantities in conventional and geometrized units. *Note*: 1 Mpc = 10^6 parsecs (pc), 1 pc = 3.026 light year ("lt yr"), 1 lt yr = 0.946×10^{16} m, 1 AU = 1.49×10^{11} m. For other useful astronomical constants see C. W. Allen, *Astrophysical Quantities*.

By now the reader must be accustomed to our use of geometrized units in which the speed of light is unity. Just as that has simplified greatly the mathematical notation in Chapters 1, 23 and 24, so also future notation will be greatly simplified if we set Newton's gravitation constant to unity. This further geometrization of our units corresponds to equating mass units to length units via the relation

$$\boxed{1 = \frac{G}{c^2} = 7.42 \times 10^{-28} \frac{\rm m}{\rm kg} \ ; \quad \text{i.e., } 1 \text{ kg} = 7.42 \times 10^{-28} \text{ m} \ .} \tag{24.85}$$

Any equation can readily be converted from conventional units to geometrized units by removing all factors of c and G; and it can readily be converted back by inserting whatever factors of c and G one needs in order to make both sides of the equation dimensionally correct. The caption of Table 24.1 lists a few important numerical quantities in both conventional units and geometrized units. (SI units are badly suited to dealing with relativistic electrodynamics; for this reason J. D. Jackson has insisted on switching from SI to Gaussian units in the last 1/3 of the 1999 edition of his classic textbook, and we do the same in the relativity portions of this book and in Table 24.1.)

In geometrized units the Einstein field equation (24.80), with $\kappa = 8\pi G = 8\pi$ [Eq. (24.84)], assumes the following standard form, to which we shall appeal extensively in coming chapters:

$$\boxed{G^{\mu\nu} = 8\pi T^{\mu\nu} \ ; \quad \text{i.e., } \mathbf{G} = 8\pi \mathbf{T} \ .} \tag{24.86}$$

24.9 Weak Gravitational Fields

The foundations of general relativity are all now in our hands. In this concluding section of the chapter, we shall explore their predictions for the properties of weak gravitational fields,

beginning with the Newtonian limit of general relativity and then moving on to more general situations.

24.9.1 Newtonian Limit of General Relativity

A general relativistic gravitational field (spacetime curvature) is said to be *weak* if there exist "nearly globally Lorentz" coordinate systems in which the metric coefficients differ only slightly from unity:

$$\boxed{g_{\alpha\beta} = \eta_{\alpha\beta} + h_{\alpha\beta} \ , \quad \text{with} \ \ |h_{\alpha\beta}| \ll 1 \ .} \tag{24.87}$$

The Newtonian limit requires that gravity be weak in this sense throughout the system being studied. It further requires a slow-motion constraint, which has three aspects: (*i*) The sources of the gravity must have slow enough motions that, with some specific choice of the nearly globally Lorentz coordinates,

$$|h_{\alpha\beta,t}| \ll |h_{\alpha\beta,j}| \ ; \tag{24.88}$$

(*ii*) the sources' motions must be slow enough that in this frame the momentum density is very small compared to the energy density

$$|T^{j0}| \ll T^{00} \equiv \rho \ ; \tag{24.89}$$

and (*iii*) any particles on which the action of gravity is to be studied must move with low velocities; i.e., must have 4-velocities satisfying

$$|u^j| \ll u^0 \ . \tag{24.90}$$

Finally, the Newtonian limit requires that the stresses in the gravitating bodies be very small compared to their mass densities

$$|T^{jk}| \ll T^{00} \equiv \rho \ . \tag{24.91}$$

When conditions (24.87)–(24.91) are all satisfied, then at leading nontrivial order in the small dimensionless quantities $|h_{\alpha\beta}|$, $|h_{\alpha\beta,t}|/|h_{\alpha\beta,j}|$, $|T^{j0}|/T^{00}$, $|u^j|/u^0$, and $|T^{jk}|/T^{00}$ the laws of general relativity reduce to those of Newtonian theory.

The details of this reduction are an exercise for the reader [Ex. 24.12]; here we give an outline:

The low-velocity constraint $|u^j|/u^0 \ll 1$ on the 4-velocity of a particle, together with its normalization $u^\alpha u^\beta g_{\alpha\beta}$ and the near flatness of the metric (24.87), implies that

$$u^0 \cong 1 \ , \quad u^j \cong v^j \equiv \frac{dx^j}{dt} \ . \tag{24.92}$$

Since $u^0 = dt/d\tau$, the first of these relations implies that in our nearly globally Lorentz coordinate system the coordinate time is very nearly equal to the proper time of our slow-speed particle. In this way, we recover the "universal time" of Newtonian theory. The universal, Euclidean space is that of our nearly Lorentz frame, with $h_{\mu\nu}$ completely ignored because of its smallness. These universal time and universal Euclidean space become the arena in which Newtonian physics is formulated.

Equation (24.92) for the components of a particle's 4-velocity, together with $|v^j| \ll 1$ and $|h_{\mu\nu}| \ll 1$, imply that the geodesic equation for a freely moving particle at leading nontrivial order is

$$\frac{dv^j}{dt} \cong \frac{1}{2} h_{00,j} \quad \text{where } \frac{d}{dt} \equiv \frac{\partial}{\partial t} + \mathbf{v} \cdot \boldsymbol{\nabla} \ . \tag{24.93}$$

(Because our spatial coordinates are Cartesian, we can put the spatial index j up on one side of the equation and down on the other without creating any danger of error.)

By comparing Eq. (24.93) with Newton's equation of motion for the particle, we deduce that h_{00} must be related to the Newtonian gravitational potential by

$$h_{00} = -2\Phi \ , \tag{24.94}$$

so the spacetime metric in our nearly globally Lorentz coordinate system must be

$$\boxed{ds^2 = -(1+2\Phi)dt^2 + (\delta_{jk} + h_{jk}) dx^j dx^k + 2h_{0j} dt\, dx^j \ .} \tag{24.95}$$

Because gravity is weak, only those parts of the Einstein tensor that are linear in $h_{\alpha\beta}$ are significant; quadratic and higher-order contributions can be ignored. Now, by the same mathematical steps as led us to Eq. (24.51) for the components of the Riemann tensor in a local Lorentz frame, one can show that linearized Riemann tensor in our nearly global Lorentz frame have that same form, i.e. (setting $g_{\alpha\beta} = \eta_{\alpha\beta} + h_{\alpha\beta}$)

$$R_{\alpha\beta\gamma\delta} = \frac{1}{2}(h_{\alpha\delta,\beta\gamma} + h_{\beta\gamma,\alpha\delta} - h_{\alpha\gamma,\beta\delta} - h_{\beta\delta,\alpha\gamma}) \ . \tag{24.96}$$

From this equation and the slow-motion constraint $|h_{\alpha\beta,t}| \ll |h_{\alpha\beta,j}|$, we infer that the space-time-space-time components of Riemann are

$$\boxed{R_{j0k0} = -\frac{1}{2} h_{00,jk} = \Phi_{,jk} = \mathcal{E}_{jk} \ .} \tag{24.97}$$

In the last step we have used Eq. (24.94). We have thereby recovered the relation between the Newtonian tidal field $\mathcal{E}_{jk} \equiv \Phi_{,jk}$ and the Relativistic tidal field R_{j0k0}. That relation can now be used, via the train of arguments in the preceding section, to show that the Einstein field equation $G^{\mu\nu} = 8\pi T^{\mu\nu}$ reduces to the Newtonian field equation $\nabla^2\Phi = 4\pi T^{00} \equiv 4\pi\rho$.

This analysis leaves the details of h_{0j} and h_{jk} unknown, because the Newtonian limit is insensitive to them.

24.9.2 Linearized Theory

There are many systems in the universe that have weak gravity, but for which the slow-motion approximations (24.88)–(24.90) and/or weak-stress approximation (24.91) fail. Examples are electromagnetic fields and high-speed particles. For such systems we need a generalization of Newtonian theory that drops the slow-motion and weak-stress constraints, but keeps the weak-gravity constraint

$$\boxed{g_{\alpha\beta} = \eta_{\alpha\beta} + h_{\alpha\beta} \ , \quad \text{with } \ |h_{\alpha\beta}| \ll 1 \ .} \tag{24.98}$$

The obvious generalization is a linearization of general relativity in $h_{\alpha\beta}$, with no other approximations being made—the so-called *linearized theory of gravity*. In this subsection we shall develop it.

In formulating linearized theory we can regard the metric pertubation $h_{\mu\nu}$ as a gravitational field that lives in flat spacetime, and correspondingly we can carry out our mathematics as though we were in special relativity. In other words, linearized theory can be regarded as a field theory of gravity in flat spacetime—the type of theory that Einstein toyed with then rejected (Sec. 24.1 above).

In linearized theory, the Riemann tensor takes the form (24.96), but we have no right to simplify it further into the form (24.97), so we must follow a different route to the Einstein field equation:

Contracting the first and third indices in (24.96), we obtain the linearized Ricci tensor $R_{\mu\nu}$, contracting once again we obtain the scalar curvature R, and then from Eq. (24.77) we obtain for the Einstein tensor and the Einstein field equation

$$\begin{aligned} 2G_{\mu\nu} &= h_{\mu\alpha,\nu}{}^{\alpha} + h_{\nu\alpha,\mu}{}^{\alpha} - h_{\mu\nu,\alpha}{}^{\alpha} - h_{,\mu\nu} - \eta_{\mu\nu}(h_{\alpha\beta}{}^{,\alpha\beta} - h_{,\beta}{}^{\beta}) \\ &= 16\pi T_{\mu\nu} \ . \end{aligned} \tag{24.99}$$

Here all indices that follow the comma are partial-derivative indices, and

$$h \equiv \eta^{\alpha\beta} h_{\alpha\beta} \tag{24.100}$$

is the "trace" of the metric perturbation. We can simplify the field equation (24.99) by reexpressing it in terms of the quantity

$$\boxed{\bar{h}_{\mu\nu} \equiv h_{\mu\nu} - \frac{1}{2} h \eta_{\mu\nu} \ .} \tag{24.101}$$

One can easily check that this quantity has the opposite trace to that of $h_{\mu\nu}$ ($\bar{h} \equiv \bar{h}_{\alpha\beta}\eta^{\alpha\beta} = -h$), so it is called the *trace-reversed metric perturbation*. In terms of it, the field equation (24.99) becomes

$$-\bar{h}_{\mu\nu,\alpha}{}^{\alpha} - \eta_{\mu\nu}\bar{h}_{\alpha\beta,}{}^{\alpha\beta} + \bar{h}_{\mu\alpha,\nu}{}^{\alpha} + \bar{h}_{\nu\alpha,\mu}{}^{\alpha} = 16\pi T_{\mu\nu} \ . \tag{24.102}$$

We can simplify this field equation further by specializing our coordinates. We introduce a new nearly globally Lorentz coordinate system that is related to the old one by

$$\boxed{x^{\alpha}_{\rm new}(\mathcal{P}) = x^{\alpha}_{\rm old}(\mathcal{P}) + \xi_{\mu}(\mathcal{P}) \ ,} \tag{24.103}$$

where ξ_μ is a very small vectorial displacement of the coordinate grid. This change of coordinates via four arbitrary functions ($\alpha = 0, 1, 2, 3$) produces a change of the functional form of the metric perturbation $h_{\alpha\beta}$ to

$$\boxed{h^{\rm new}_{\mu\nu} = h^{\rm old}_{\mu\nu} - \xi_{\mu,\nu} - \xi_{\nu,\mu} \ ,} \tag{24.104}$$

[Ex. 24.13] and a corresponding change of the trace-reversed metric pertubation. This is linearized theory's analog of a *gauge transformation* in electromagnetic theory. Just as an

electromagnetic gauge alters the vector potential $A_\mu^{\rm new} = A_\mu^{\rm old} - \psi_{,\mu}$, so the linearized-theory gauge change alters $h_{\mu\nu}$ and $\bar{h}_{\mu\nu}$; and just as the force-producing electromagnetic field tensor $F_{\mu\nu}$ is unaffected by an electromagnetic gauge change, so the tidal-force-producing linearized Riemann tensor is left unaffected by the gravitational gauge change.

By a special choice of the four functions ξ^α, we can impose the following four gauge conditions on $\bar{h}_{\mu\nu}$:

$$\boxed{\bar{h}_{\mu\nu,}{}^{\nu} = 0 \ .} \tag{24.105}$$

These, obviously, are linearized theory's analog of the electromagnetic Lorentz gauge condition $A_\mu{}^{,\mu} = 0$, so they are called the *gravitational Lorenz gauge*. Just as the flat-spacetime Maxwell equations take the remarkably simple wave-equation form $A_{\mu,\alpha}{}^{\alpha} = 4\pi J_\mu$ in Lorenz gauge, so also the linearized Einstein equation (24.102) takes the corresponding simple wave-equation form in gravitational Lorenz gauge:

$$\boxed{-\bar{h}_{\mu\nu,\alpha}{}^{\alpha} = 16\pi T_{\mu\nu} \ .} \tag{24.106}$$

By the same method as one uses in electromagnetic theory, one can solve this gravitational field equation for the field $\bar{h}_{\mu\nu}$ produced by an arbitrary stress-energy-tensor source:

$$\boxed{\bar{h}_{\mu\nu}(t,\mathbf{x}) = \int \frac{4T_{\mu\nu}(t-|\mathbf{x}-\mathbf{x}'|,\mathbf{x}'])}{|\mathbf{x}-\mathbf{x}'|} dV_{x'}} \tag{24.107}$$

The quantity in the numerator is the stress-energy source evaluated at the "retarded time" $t' = t - |\mathbf{x}-\mathbf{x}'|$. This equation for the field, and the wave equation (24.106) that underlies it, show explicitly that dynamically changing distributions of stress-energy must generate *gravitational waves*, which propagate outward from their source at the speed of light (Einstein, 1918). We shall study these gravitational waves in Chap. 26.

24.9.3 Gravitational Field Outside a Stationary, Linearized Source

Let us specialize to a time-independent source (so $T_{\mu\nu,t} = 0$ in our chosen nearly globally Lorentz frame), and compute its external gravitational field as a power series in 1/(distance to source). We place our origin of coordinates at the source's center of mass, so

$$\int x^j T^{00} dV_x = 0 \ , \tag{24.108}$$

and in the same manner as in electromagnetic theory, we expand

$$\frac{1}{|\mathbf{x}-\mathbf{x}'|} = \frac{1}{r} + \frac{x^j x^{j'}}{r^3} + \dots \ , \tag{24.109}$$

where $r \equiv |\mathbf{x}|$ is the distance of the field point from the source's center of mass. Inserting Eq. (24.109) into the general solution (24.107) of the Einstein equation and taking note of the conservation laws $T^{\alpha j}{}_{,j} = 0$, we obtain for the source's external field

$$\boxed{\bar{h}_{00} = \frac{4M}{r} + \mathrm{O}\left(\frac{1}{r^3}\right) \ , \quad \bar{h}_{0j} = -\frac{2\epsilon_{jkm}S^k x^m}{r^3} + \mathrm{O}\left(\frac{1}{r^3}\right) \ , \quad \bar{h}_{ij} = \mathrm{O}\left(\frac{1}{r^3}\right) \ ;} \tag{24.110}$$

Here M and S^k are the source's mass and angular momentum:

$$\boxed{M \equiv \int T^{00} dV_x \ , \quad S_k \equiv \int \epsilon_{kab} x^a T^{0b} dV_x \ .} \tag{24.111}$$

see Ex. 24.14. This expansion in $1/r$, as in the electromagnetic case, is a multipolar expansion. At order $1/r$ the field is spherically symmetric and the monopole moment is the source's mass M. At order $1/r^2$ there is a "magnetic-type dipole moment", the source's spin angular momentum S_k. These are the leading-order moments in two infinite sets: the "mass multipole" moments (analog of electric moments), and the "mass-current multipole" moments (analog of magnetic moments). For details on all the higher order moments, see, e.g., Thorne (1980).

The metric perturbation can be computed by reversing the trace reversal, $h_{\alpha\beta} = \bar{h}_{\alpha\beta} - \eta_{\alpha\beta}\bar{h}$. Thereby we obtain for the spacetime metric $g_{\alpha\beta} = \eta_{\alpha\beta} + h_{\alpha\beta}$ at linear order, outside the source,

$$\boxed{ds^2 = -\left(1 - \frac{2M}{r}\right) dt^2 - \frac{4\epsilon_{jkm} S^k x^m}{r^3} dt dx^j + \left(1 + \frac{2M}{r}\right) \delta_{jk} dx^j dx^k + \mathcal{O}\left(\frac{1}{r^3}\right) dx^\alpha dx^\beta \ .} \tag{24.112}$$

In spherical polar coordinates, with the polar axis along the direction of the source's angular momentum, the leading order terms take the form

$$\boxed{ds^2 = -\left(1 - \frac{2M}{r}\right) dt^2 - \frac{4S}{r} \sin^2\theta dt d\phi + \left(1 + \frac{2M}{r}\right) (dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2) \ ,} \tag{24.113}$$

where $S \equiv |\mathbf{S}|$ is the magnitude of the source's angular momentum.

This is a very important result. It tells us that we can "read off" the mass M and angular momentum S^k from the asymptotic form of the source's metric. In the next chapter we shall devise, from the metric (24.113), physical measurements that one can make outside the source to determine its mass and angular momentum. As one would expect from Newtonian theory, the mass M will show up as the source of a "gravitational acceleration" that can be measured via Kepler's laws for an orbiting particle. It will turn out that the angular-momentum term in the metric shows up physically via a *dragging of inertial frames* that causes inertial-guidance gyroscopes near the body to precess relative to the "distant stars".

For a time-independent body with *strong* internal gravity (e.g. a black hole), the distant gravitational field will have the same general form (24.112), (24.113) as for a weakly gravitating body, but the constants M and S^k that appear in the metric will not be expressible as the integrals (24.111) over the body's interior. Nevertheless, they will be measureable by the same techniques as for a weakly gravitating body (Kepler's laws and frame dragging), and they can be interpreted as the body's total mass and angular momentum.

24.9.4 Conservation Laws for Mass, Momentum and Angular Momentum

Consider a static (unmoving) sphere $\mathcal{S}$ surrounding our time-independent source of gravity, with such a large radius r that the $O(1/r^3)$ corrections in $\bar{h}_{\mu\nu}$ and in the metric [Eqs. (24.111)–(24.113)] can be ignored. Suppose that a small amount of mass-energy E (as measured in the sphere's and source's rest frame) is injected through the sphere, into the source. Then the special relativistic law of mass-energy conservation tells us that the source's mass $M = \int T^{00} dV_x$ will increase by $\Delta M = E$. Similarly, if an energy flux T^{0j} flows through the sphere, the source's mass will change by

$$\boxed{\frac{dM}{dt} = -\int_{\mathcal{S}} T^{0j} d\Sigma_j \ ,} \tag{24.114}$$

where $d\Sigma_j$ is the sphere's outward-pointing surface-area element, and the minus sign is due to the fact that $d\Sigma_j$ points outward, not inward. Since M is the mass that appears in the source's asymptotic gravitational field $\bar{h}_{\mu\nu}$ and metric $g_{\alpha\beta}$, this conservation law can be regarded as describing how the source's gravitating mass changes when energy is injected into it.

From the special relativistic law for angular momentum conservation, we deduce a similar result: A flux $\epsilon_{ijk} x^j T^{km}$ of angular momentum through the sphere produces the following change in the angular momentum S_k that appears in the source's asymptotic field $\bar{h}_{\mu\nu}$ and metric:

$$\boxed{\frac{dS_i}{dt} = -\int_{\mathcal{S}} \epsilon_{ijk} x^j T^{km} d\Sigma_m \ .} \tag{24.115}$$

There is also a conservation law for a gravitationally measured linear momentum. That linear momentum does not show up in the asymptotic field and metric that we wrote down above [Eqs. (24.111)–(24.113)] because our coordinates were chosen to be attached to the source's center of mass—i.e., they are the Lorentz coordinates of the source's rest frame. However, if linear momentum P_j is injected through our sphere $\mathcal{S}$ and becomes part of the source, then the source's center of mass will start moving, and the asymptotic metric will acquire a new term

$$\delta g_{0j} = -4P_j/r \ , \tag{24.116}$$

where (after the injection)

$$P_j = P^j = \int T^{0j} dV_x \tag{24.117}$$

[see Eq. (24.107) with $\bar{h}^{0j} = -\bar{h}_{0j} = -h_{0j} = -\delta g_{0j}$; also see Ex 24.14b]. More generally, the rate of change of the source's total linear momentum (the P_j term in the asymptotic g_{0j}) is the integral of the inward flux of momentum (inward component of the stress tensor) across the sphere:

$$\boxed{\frac{dP_j}{dt} = -\int_{\mathcal{S}} T^{jk} d\Sigma_j \ .} \tag{24.118}$$

For a time-independent source with *strong* internal gravity, not only does the asymptotic metric, far from the source, have the same form (24.112), (24.113), (24.116) as for a weakly gravitating source; the conservation laws (24.114), (24.115), (24.118) for its gravitationally measured mass, angular momentum and linear momentum continue to hold true. The sphere $\mathcal{S}$, of course, must be placed far from the source, in a region where gravity is very weak, so linearized theory will be valid in the vicinity of $\mathcal{S}$. When this is done, then the special relativistic description of inflowing mass, angular momentum and energy is valid at $\mathcal{S}$, and the linearized Einstein equations, applied in the vicinity of $\mathcal{S}$ (and not extended into the strong-gravity region), turn out to guarantee that the M, S_j and P_j appearing in the asympototic metric evolve in accord with the conservation laws (24.114), (24.115), (24.118).

For strongly gravitating sources, these conservation laws owe their existence to the spacetime's asymptotic time-translation, rotation, and space-translation symmetries. In generic, strong-gravity regions of spacetime there are no such symmetries, and correspondingly no integral conservation laws for energy, angular momentum, or linear momentum.

If a strongly gravitating source is dynamical rather than static, it will emit gravitational waves (Chap. 26). The amplitudes of those waves, like the influence of the source's mass, die out as $1/r$ far from the source, so spacetime retains its asymptotic time-translation, rotation and space-translation symmetries. These symmetries continue to enforce integral conservation laws on the gravitationally measured mass, angular momentum and linear momentum [Eqs. (24.114), (24.115), (24.118)], but with the new requirement that one include, in the fluxes through $\mathcal{S}$, contributions from the gravitational waves' energy, angular momentum and linear momentum; see Chap. 26.

For a more detailed and rigorous derivation and discussion of these asymptotic conservation laws, see Chaps. 18 and 19 of MTW.

****

EXERCISES

Exercise 24.12 *Derivation: Newtonian limit of general relativity*
Consider a system that can be covered by a nearly globally Lorentz coordinate system in which the Newtonian-limit constraints (24.87)–(24.91) are satisfied. For such a system, flesh out the details of the text's derivation of the Newtonian limit. More specifically:

(a) Derive Eq. (24.92) for the components of the 4-velocity of a particle.

(b) Show that the geodesic equation reduces to Eq. (24.93).

(c) Show that to linear order in the metric perturbation $h_{\alpha\beta}$ the components of the Riemann tensor take the form (24.96).

(d) Show that in the slow-motion limit the space-time-space-time components of Riemann take the form (24.97).

Exercise 24.13 *Derivation: Gauge Transformations in Linearized Theory*

(a) Show that the "infinitesimal" coordinate transformation (24.103) produces the change (24.104) of the linearized metric perturbation.

(b) Exhibit a differential equation for the ξ^α that brings the metric perturbation into gravitational Lorenz gauge, i.e. that makes $h^{\rm new}_{\mu\nu}$ obey the Lorenz gauge condtion (24.105)

(c) Show that in gravitational Lorenz gauge, the Einstein field equation (24.102) reduces to (24.106).

Exercise 24.14 *Derivation: External Field of Stationary, Linearized Source*
Derive Eqs. (24.110) for the trace reversed metric perturbation outside a stationary (time-independent), linearized source of gravity. More specifically:

(a) First derive $\bar{h}_{00}$. In your derivation identify a dipolar term of the form $4D_j x^j/r^3$, and show that by placing the origin of coordinates on the center of mass, Eq. (24.108), one causes the dipole moment D_j to vanish.

(b) Next derive $\bar{h}_{0j}$. The two terms in (24.109) should give rise to two terms. The first of these is $4P_j/r$ where P_j is the source's linear momentum. Show, using the gauge condition $\bar{h}^{0\mu}_{\ \ ,\mu} = 0$ [Eq. (24.105)] that if the momentum is nonzero, then the mass dipole term of part (a) must have a nonzero time derivative, which violates our assumption of stationarity. Therefore, for this source the linear momentum must vanish. Show that the second term gives rise to the $\bar{h}_{0j}$ of Eq. (24.110). [Hint: you will have to add a perfect divergence, $(T^{0a'}x^{j'}x^{m'})_{,a'}$ to the integrand.]

(c) Finally derive $\bar{h}_{ij}$. [Hint: Show that $T^{ij} = (T^{ia}x^i)_{,a}$ and thence that the volume integral of T^{ij} vanishes; and similarly for $T^{ij}x^k$.]

****

Bibliographic Note

For a superb, detailed historical account of Einstein's intellectual struggle to formulate the laws of general relativity, see Pais (1982). For Einstein's papers of that era, in the original German and in English translation, with detailed annotations and explanations by editors with strong backgrounds in both physics and history of science, see Einstein (1989–2002). For some key papers of that era by other major contributors besides Einstein, in English translation, see Einstein, Lorentz, Minkowski and Weyl (1923).

This chapter's pedagogical approach to presenting the fundamental concepts of general relativity is strongly influenced by MTW (Misner, Thorne and Wheeler 1973), where readers will find much greater detail. See, especially, Chap. 8 for the the mathematics (differential geometry) of curved spacetime, or Chaps. 9–14 for far greater detail; Chap. 16 for the Einstein equivalence principle and how to lift laws of physics into curved spacetime; Chap. 17 for the Einstein field equations and many different ways to derive them; Chap. 18 for

Box 24.2
Important Concepts in Chapter 24

- Local Lorentz frame, Sec. 24.2
  - Nonmeshing of local Lorentz frames due to spacetime curvature, Sec. 24.3
  - Metric and connection coefficients in Local Lorentz frame, Eqs. (24.15) and (24.16)
- Principle of relativity, Sec. 24.2
- Motion of a freely falling particle: geodesic with $\nabla_{\vec{p}}\vec{p} = 0$, Sec. 24.4
  - Geodesic equation in any coordinate system, Eq. (24.26)
  - Conserved quantity associated with symmetry of the spacetime, Ex. 24.4
  - Action principle for geodesic, Ex. 24.5
- Tidal Gravity and Spacetime Curvature
  - Newtonian Tidal field $\mathcal{E}_{ij} = \partial^2\Phi/\partial x^i \partial x^j$, Sec. 24.5.1
  - Riemann curvature as tidal field; equation of geodesic deviation, Sec. 24.5.1
  - Connection of relavistic and Newtonian tidal fields, $R_{j0k0} = \mathcal{E}_{jk}$, Sec. 24.5.2
  - Tidal field outside the Earth or other spherical, gravitating body, Eq. (24.44).
- Properties of Riemann tensor (symmetries, Ricci Tensor, Curvature scalar, how to compute its components, radius of curvature of spacetime), Sec. 24.6
- Einstein's equivalence principle: Lifting the laws of physics into curved spacetime, Secs. 24.2 and 24.7
  - Curvature coupling effects, Sec. 24.7
  - Breakdown of global conservation laws for energy and momentum, Sec. 24.7
- The Einstein field equation, Sec. 24.8
  - Its connection to Newton's field equation, Sec. 24.8
  - Einstein tensor and its vanishing divergence, Sec. 24.8
- Geometrized units, Eq. (24.85) and Table 24.1
- Newtonian limit of general relativity, Sec. 24.9.1
  - Conditions for validity: weak gravity, slow motion, small stresses, Sec. 24.9.1
- Linearized theory, Sec. 24.9.2
  - Gravitational Lorenz gauge, Eq. (24.105)
  - Wave equation for metric pertubation, with stress-energy tensor as source, Eqs. (24.106), (24.107)
- Metric outside a stationary, linearized source, Sec. 24.9.3
  - Roles of mass and angular momentum in metric, Sec. 24.9.3
  - Integral conservation laws for source's mass, linear momentum and angular momentum, Sec. 24.9.4

weak gravitational fields (the Newtonian limit and Linearized Theory); and Chaps. 19 and 20 for the metric outside a stationary, linearized source and for the source's conservation laws for mass, momentum, and angular momentum.

For a superb, elementary introduction to the fundamental concepts of general relativity from a viewpoint that is somewhat less mathematical than this chapter or MTW, see Hartle (2003). We also recommend, at a somewhat elementary level, Schutz (1985), and at a more advanced level, Carroll (2004), and at a very advanced and mathematical level, Wald (1984).

Bibliography

Carroll, S. M., 2004. *Spacetime and Geometry: An Introduction to General Relativity*, San Francisco: Addison Wesley.

Einstein, Albert, 1907. "Über das Relativitätsprinzip und die aus demselben gesogenen Folgerungen," *Jahrbuch der Radioaktivität und Elektronik*, **4**, 411–462; English translation: paper 47 in *The Collected Papers of Albert Einstein*, Volume 2, Princeton University Press, Princeton, NJ.

Einstein, Albert, 1915. "Die Feldgleichungen der Gravitation," *Preuss. Akad. Wiss. Berlin, Sitzungsber*, **1915 volume**, 844–847.

Einstein, Albert, 1916. "Die Grundlage der allgemeinen Relativitätstheorie," *Annalen der Physik*, **49**, 769–822. English translation in Einstein *et al*. (1923).

Einstein, Albert, 1918. "Über Gravitationswellen," *Sitzungsberichte der Königlish Preussischen Akademie der Wissenschaften*, **1918 volume**, 154–167.

Einstein, Albert, Lorentz, Hendrik A., Minkowski, Hermann, and Weyl, Hermann, 1923. *The Principle of Relativity*, Dover, New York.

Einstein, Albert, 1989–2002. *The Collected Papers of Albert Einstein*, Volumes 2–7, Princeton University Press, Princeton, NJ; and http://www.einstein.caltech.edu/

Hartle, J. B., 2003. *Gravity: An Introduction to Einstein's General Relativity*, San Francisco: Addison-Wesley.

Hilbert, David, 1915. "Die Grundlagen der Physik," *Königl. Gesell. d. Wiss. Göttingen, Nachr., Math.-Phys. Kl.*, **1917 volume**, 53–76.

MTW: Misner, Charles W., Thorne, Kip S., and Wheeler, John A., 1973. *Gravitation*, W. H. Freeman & Co., San Francisco.

Minkowski, Hermann, 1908. "Space and Time," Address at the 80th Assembly of German Natural Scientists and Physicians, at Cologne, 21 September 1908; text published posthumously in *Annalen der Physik*, **47**, 927 (1915); English translation in Einstein *et al*. (1923).

Pais, Abraham, 1982. *Subtle is the Lord . . . : The Science and Life of Albert Einstein*, Oxford University Press: Oxford.

Schutz, B. 1980. *Geometrical Methods of Mathematical Physics*, Cambridge: Cambridge University Press.

Thorne, Kip S., 1980. "Multipole expansions of gravitational radiation," *Reviews of Modern Physics*, **52**, 299; especially Secs. VIII and X.

Wald, R. M. 1984. *General Relativity*, Chicago: University of Chicago Press.

Will, Clifford M., 1981. *Theory and Experiment in Gravitational Physics,* Cambridge University Press: Cambridge.

Will, Clifford M., 1986. *Was Einstein Right?* Basic Books, New York.

Will, Clifford M., 2006. "The Confrontation between General Relativity and Experiment," *Living Reviews in Relativity*, **9**, 3, (2006). URL (cited in April 2007): http://www.livingreviews.org/lrr-2006-3

Contents

25 Relativistic Stars and Black Holes **1**
25.1 Introduction 1
25.2 Schwarzschild's Spacetime Geometry 2
25.3 Static Stars 8
25.3.1 Birkhoff's Theorem 8
25.3.2 Stellar Interior 9
25.3.3 Local Energy and Momentum Conservation 12
25.3.4 Einstein Field Equation 14
25.3.5 Stellar Models and Their Properties 15
25.4 Gravitational Implosion of a Star to Form a Black Hole 20
25.5 Spinning Black Holes: The Kerr Spacetime 31
25.5.1 The Kerr Metric for a Spinning Black Hole 31
25.5.2 Dragging of Inertial Frames 32
25.5.3 The Light-Cone Structure, and the Horizon 32
25.5.4 Evolution of Black Holes: Rotational Energy and Its Extraction . . . 35
25.6 The Many-Fingered Nature of Time 40

Chapter 25

Relativistic Stars and Black Holes

Version 0625.1.K.pdf, 9 May 2007
Please send comments, suggestions, and errata via email to kip@tapir.caltech.edu or on paper to Kip Thorne, 130-33 Caltech, Pasadena CA 91125

Box 25.1
Reader's Guide

- This chapter relies significantly on
  - The special relativity portions of Chap. 1.
  - Chapter 23, on the transition from special relativity to general relativity.
  - Chapter 24, on the fundamental concepts of general relativity.
- Portions of this chapter are a foundation for the applications of general relativity theory to gravitational waves (Chap. 26) and to cosmology (Chap. 27).

25.1 Introduction

Having sketched the fundamentals of Einstein's theory of gravity, general relativity, we shall now illustrate his theory by means of several concrete applications: stars and black holes in this chapter, gravitational waves in Chap. 26, and the large-scale structure and evolution of the universe in Chap. 27.

While stars and black holes are the central thread of this chapter, we study them less for their own intrinsic interest than for their roles as vehicles by which to understand general relativity: Using them we shall elucidate a number of issues that we have already met: the physical and geometric interpretations of spacetime metrics and of coordinate systems, the Newtonian limit of general relativity, the geodesic motion of freely falling particles and photons, local Lorentz frames and the tidal forces measured therein, proper reference frames, the Einstein field equations, the local law of conservation of 4-momentum, and the asymptotic

structure of spacetime far from gravitating sources. Stars and black holes will also serve to introduce several new physical phenomena that did not show up in our study of the foundations of general relativity: the gravitational redshift, the "many-fingered" nature of time, event horizons, and spacetime singularities.

We begin this chapter, in Sec. 25.2, by studying the geometry of the curved spacetime outside any static star, as predicted by the Einstein field equation. In Sec. 25.3 we study general relativity's description of the interiors of static stars. In Sec. 25.4 we turn attention to the spherically symmetric gravitational implosion by which a nonrotating star is transformed into a black hole, and to the "Schwarzschild" spacetime geometry outside and inside the resulting static, spherical hole. In Sec. 25.5 we study the "Kerr" spacetime geometry of a spinnning black hole. Finally, in Sec. 25.6 we elucidate the nature of "time" in the curved spacetimes of general relativity.

25.2 Schwarzschild's Spacetime Geometry

On January 13, 1916, just seven weeks after formulating the final version of his field equation, $\mathbf{G} = 8\pi\mathbf{T}$, Albert Einstein read to a meeting of the Prussian Academy of Sciences in Berlin a letter from the eminent German astrophysicist Karl Schwarzschild. Schwarzschild, as a member of the German army, had written from the World-War-One Russian front to tell Einstein of a mathematical discovery he had made: he had found the world's first exact solution to the Einstein field equation.

Written as a line element in a special coordinate system (coordinates named t, r, θ, ϕ) that Schwarzschild invented for the purpose, Schwarzschild's solution takes the form (Schwarzschild 1916a)

$$\boxed{ds^2 = -(1-2M/r)dt^2 + \frac{dr^2}{(1-2M/r)} + r^2(d\theta^2 + \sin^2\theta d\phi^2)\ ,} \tag{25.1}$$

where M is a constant of integration. The connection coefficients, Riemann tensor, and Ricci and Einstein tensors for this metric can be computed by the methods of Chaps. 23 and 24; see Ex. 25.1. The results are tabulated in Box 25.2. The key bottom line is that the Einstein tensor vanishes. Therefore, the Schwarzschild metric (25.1) is a solution of the Einstein field equations with vanishing stress-energy tensor.

Many readers know already the lore of this subject: The Schwarzschild spacetime is reputed to represent the vacuum exterior of a nonrotating, spherical star; and also the exterior of a spherical star as it implodes to form a black hole; and also the exterior and interior of a nonrotating, spherical black hole; and also a wormhole that connects two different universes or two widely separated regions of our own universe.

How does one discover these physical interpretations of the Schwarzschild metric (25.1)? The tools for discovering them—and, more generally, the tools for interpreting physically any spacetime metric that one encounters—are a central concern of this chapter.

When presented with a line element such as (25.1), one of the first questions one is tempted to ask is "What is the nature of the coordinate system?" Since the metric coefficients

will be different in some other coordinate system, surely one must know something about the coordinates in order to interpret the line element.

Remarkably, one need not go to the inventor of the coordinates to find out their nature. Instead one can turn to the line element itself: the line element (or metric coefficients) contain full information not only about the details of the spacetime geometry, but also about the nature of the coordinates. The line element (25.1) is a good example:

Look first at the 2-dimensional surfaces in spacetime that have constant values of t and r. We can regard $\{\theta, \phi\}$ as a coordinate system on each such 2-surface; and the spacetime line element (25.1) tells us that the geometry of the 2-surface is given in terms of those coordinates by

$$^{(2)}ds^2 = r^2(d\theta^2 + \sin^2\theta d\phi^2) \tag{25.2}$$

(where the prefix $^{(2)}$ refers to the dimensionality of the surface). This is the line element (metric) of an ordinary, everyday 2-dimensional sphere expressed in standard spherical polar coordinates. Thus, we have learned that the *Schwarzschild spacetime is spherically symmetric*, and moreover that *θ and ϕ are standard spherical polar coordinates*. Here is an example of extracting from a metric information about both the coordinate-independent spacetime geometry and the coordinate system being used.

Note, further, from Eq. (25.2) that the circumferences and surface areas of the spheres $(t, r) =$ const in Schwarzschild spacetime are given by

$$\boxed{\text{circumference} = 2\pi r\ , \quad \text{area} = 4\pi r^2\ .} \tag{25.3}$$

This tells us one aspect of the geometric interpretation of the r coordinate: r is a *radial coordinate* in the sense that the circumferences and surface areas of the spheres in Schwarzschild spacetime are expressed in terms of r in the standard manner (25.3). We must not go further, however, and assert that r is radius in the sense of being the distance from the center of one of the spheres to its surface. The center, and the line from center to surface, do not lie on the sphere itself and they thus are not described by the spherical line element (25.2). Moreover, since we know that spacetime is curved, we have no right to expect that the distance from the center of a sphere to its surface will be given by distance = circumference$/2\pi = r$ as in flat spacetime.

Returning to the Schwarzschild line element (25.1), let us examine several specific regions of spacetime: At "radii" r large compared to the integration constant M, the line element (25.1) takes the form

$$ds^2 = -dt^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2)\ . \tag{25.4}$$

This is the line element of flat spacetime, $ds^2 = -dt^2 + dx^2 + dy^2 + dz^2$ written in spherical polar coordinates $[x = r\sin\theta\cos\phi,\ y = r\sin\theta\sin\phi,\ z = r\cos\theta]$. Thus, *Schwarzschild spacetime is asymptotically flat in the region of large radii $r/M \to \infty$*. This is just what one might expect physically when one gets far away from all sources of gravity. Thus, it is reasonable to presume that the Schwarzschild spacetime geometry is that of some sort of isolated, gravitating body which is located in the region $r \sim M$.

The large-r line element (25.4) not only reveals that Schwarzschild spacetime is asymptotically flat; it also shows that in the asymptotically flat region the Schwarzschild t is the time

Box 25.2
Connection Coefficients and Curvature Tensors for Schwarzschild

The coordinate basis vectors for the Schwarzschild solution are

$$\vec{e}_t = \frac{\partial}{\partial t},\ \ \vec{e}_r = \frac{\partial}{\partial r},\ \ \vec{e}_\theta = \frac{\partial}{\partial \theta},\ \ \vec{e}_\phi = \frac{\partial}{\partial \phi};\ \ \vec{e}^{\,t} = \vec{\nabla} t,\ \ \vec{e}^{\,r} = \vec{\nabla} r,\ \ \vec{e}^{\,\theta} = \vec{\nabla}\theta,\ \ \vec{e}^{\,\phi} = \vec{\nabla}\phi. \quad (1)$$

The covariant metric coefficients in this coordinate basis are [cf. Eq. (25.1)]

$$g_{tt} = -\left(1 - \frac{2M}{r}\right)\ , \quad g_{rr} = \frac{1}{(1-2M/r)}\ , \quad g_{\theta\theta} = r^2\ , \quad g_{\phi\phi} = r^2 \sin^2\theta\ ; \qquad (2a)$$

and the contravariant metric coefficients are the inverse of these

$$g^{tt} = -\frac{1}{(1-2M/r)}\ , \quad g^{rr} = \left(1 - \frac{2M}{r}\right)\ , \quad g^{\theta\theta} = \frac{1}{r^2} \quad g^{\phi\phi} = \frac{1}{r^2 \sin^2\theta}\ . \qquad (2b)$$

The nonzero connection coefficients in this coordinate basis are

$$\Gamma^t{}_{rt} = \Gamma^t{}_{tr} = \frac{M}{r^2}\frac{1}{(1-2M/r)}\ , \quad \Gamma^r{}_{tt} = \frac{M}{r^2}(1-2M/r)\ , \quad \Gamma^r{}_{rr} = -\frac{M}{r^2}\frac{1}{(1-2M/r)}\ ,$$

$$\Gamma^r{}_{\theta\theta} = -r(1-2M/r)\ , \quad \Gamma^\theta{}_{r\theta} = \Gamma^\theta{}_{\theta r} = \Gamma^\phi{}_{r\phi} = \Gamma^\phi{}_{\phi r} = \frac{1}{r}\ , \qquad (3)$$

$$\Gamma^r{}_{\phi\phi} = -r\sin^2\theta(1-2M/r)\ , \quad \Gamma^\theta{}_{\phi\phi} = -\sin\theta\cos\theta\ , \quad \Gamma^\phi{}_{\theta\phi} = \Gamma^\phi{}_{\phi\theta} = \cot\theta\ ,$$

The orthonormal basis associated with the above coordinate basis is

$$\vec{e}_{\hat{0}} = \frac{\partial/\partial t}{\sqrt{1-2M/r}}\ , \quad \vec{e}_{\hat{r}} = \sqrt{1-\frac{2M}{r}}\frac{\partial}{\partial r}\ , \quad \vec{e}_{\hat{\theta}} = \frac{1}{r}\frac{\partial}{\partial\theta}\ , \quad \vec{e}_{\hat{\phi}} = \frac{1}{r\sin\theta}\frac{\partial}{\partial\phi}\ . \qquad (4)$$

The nonzero connection coefficients in this orthonormal basis are

$$\Gamma^{\hat{r}}{}_{\hat{t}\hat{t}} = \Gamma^{\hat{t}}{}_{\hat{r}\hat{t}} = \frac{M}{r^2\sqrt{1-2M/r}}\ , \quad \Gamma^{\hat{\phi}}{}_{\hat{\theta}\hat{\phi}} = -\Gamma^{\hat{\theta}}{}_{\hat{\phi}\hat{\phi}} = \frac{\cot\theta}{r}\ ,$$

$$\Gamma^{\hat{\theta}}{}_{\hat{r}\hat{\theta}} = \Gamma^{\hat{\phi}}{}_{\hat{r}\hat{\phi}} = -\Gamma^{\hat{r}}{}_{\hat{\theta}\hat{\theta}} = -\Gamma^{\hat{r}}{}_{\hat{\phi}\hat{\phi}} = \frac{\sqrt{1-2M/r}}{r}\ . \qquad (5)$$

The nonzero components of the Riemann tensor in this orthonormal basis are

$$R_{\hat{r}\hat{t}\hat{r}\hat{t}} = -R_{\hat{\theta}\hat{\phi}\hat{\theta}\hat{\phi}} = -\frac{2M}{r^3}\ , \quad R_{\hat{\theta}\hat{t}\hat{\theta}\hat{t}} = R_{\hat{\phi}\hat{t}\hat{\phi}\hat{t}} = -R_{\hat{r}\hat{\phi}\hat{r}\hat{\phi}} = -R_{\hat{r}\hat{\theta}\hat{r}\hat{\theta}} = \frac{M}{r^3}\ , \qquad (6)$$

and those obtainable from these via the symmetries (24.52) of Riemann. The Ricci tensor, curvature scalar, and Einstein tensor all vanish—which implies that the Schwarzschild metric is a solution of the vacuum Einstein field equations.

coordinate of a Lorentz reference frame. Notice that the region of strong spacetime curvature has a boundary (say, $r \sim 100M$) that remains forever fixed relative to the asymptotically Lorentz spatial coordinates $x = r\sin\theta\cos\phi$, $y = r\sin\theta\sin\phi$, $z = r\cos\theta$. This means that the asymptotic Lorentz frame can be regarded as the body's *asymptotic rest frame*. We conclude, then, that far from the body the Schwarzschild t coordinate becomes the Lorentz time of the body's asymptotic rest frame, and the Schwarzschild r, θ, ϕ coordinates become spherical polar coordinates in the body's asymptotic rest frame.

As we move inward from $r = \infty$, we gradually begin to see spacetime curvature. That curvature shows up, at $r \gg M$, in slight deviations of the Schwarzschild metric coefficients from those of a Lorentz frame: to first order in M/r the line element (25.1) becomes

$$ds^2 = -\left(1 - \frac{2M}{r}\right) dt^2 + \left(1 + \frac{2M}{r}\right) dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2) \ . \tag{25.5}$$

or, equivalently, in Cartesian spatial coordinates,

$$ds^2 = -\left(1 - \frac{2M}{\sqrt{x^2+y^2+z^2}}\right) dt^2 + dx^2 + dy^2 + dz^2 + \frac{2M}{r}\left(\frac{x}{r}dx + \frac{y}{r}dy + \frac{z}{r}dz\right)^2 \ . \tag{25.6}$$

It is reasonable to expect that, at these large radii where the curvature is weak, Newtonian gravity will be a good approximation to Einsteinian gravity. In Sec. 24.9.1 of the last chapter we studied in detail the transition from general relativity to Newtonian gravity, and found that, in nearly Newtonian situations if one uses a nearly globally Lorentz coordinate system (as we are doing), the line element should take the form [Eq. (24.95)]

$$ds^2 = -(1 + 2\Phi)dt^2 + (\delta_{jk} + h_{jk})dx^j dx^k + 2h_{tj}dt\, dx^j \ , \tag{25.7}$$

where $h_{\mu\nu}$ are metric corrections that are very small compared to unity and Φ (which shows up in the time-time part of the metric) is the Newtonian potential. Direct comparison of (25.7) with (25.6) shows that a Newtonian description of the body's distant gravitational field will entail a Newtonian potential given by

$$\Phi = -\frac{M}{r} \tag{25.8}$$

($\Phi = -GM/r$ in cgs units). This, of course, is the external Newtonian field of a body with mass M. Thus, *the integration constant M in the Schwarzschild line element is the mass which characterizes the body's distant, nearly Newtonian gravitational field*. This is an example of reading the mass of a body off the asymptotic form of the metric (Sec. 24.9.3).

Notice that the asymptotic metric here [Eq. (25.5)] differs in its spatial part from that in Sec. 24.9.3 [Eq. (24.113)]. This difference arises from the use of different radial coordinates here and there: If we define $\bar{r}$ by $r = \bar{r} + M$ at radii $r \gg M$, then to linear order in M/r, the asymptotic Schwarzschild metric (25.5) becomes

$$ds^2 = -\left(1 - \frac{2M}{\bar{r}}\right) dt^2 + \left(1 + \frac{2M}{\bar{r}}\right) [d\bar{r}^2 + \bar{r}^2(d\theta^2 + \sin^2\theta d\phi^2)] \ , \tag{25.9}$$

which is the same as Eq. (24.113) with vanishing angular momentum $S_j = 0$. This easy change of the spatial part of the metric reinforces the fact that one reads the asymptotic Newtonian potential and the source's mass M off the *time-time* components of the metric, and not the spatial part of the metric.

We can describe the physical interpretation of M as the body's mass in operational terms as follows: Suppose that a test particle (e.g., a small planet) moves around our central body in a circular orbit with radius $r \gg M$. A Newtonian analysis of the orbit predicts that, as measured using Newtonian time, the period of the orbit will be $P = 2\pi(r^3/M)^{\frac{1}{2}}$. Moreover, since Newtonian time is very nearly equal to the time t of the nearly Lorentz coordinates used in (25.5) [cf. Sec. 24.9.1], and since that t is Lorentz time in the body's relativistic, asymptotic rest frame, the orbital period as measured by observers at rest in the asymptotic rest frame must be $P = 2\pi(r^3/M)^{\frac{1}{2}}$. Thus, *$M$ is the mass that appears in Kepler's laws for the orbits of test particles far from the central body*. This quantity is often called the body's "active gravitational mass," since it is the mass that characterizes the body's gravitational pull. It is also called the body's "total mass-energy" because it turns out to include all forms of mass and energy that the body possesses (rest mass, internal kinetic energy, and all forms of internal binding energy including gravitational).

We note, in passing, that one can use general relativity to deduce the Keplerian role of M without invoking the Newtonian limit: We place a test particle in the body's equatorial plane $\theta = \pi/2$ at a radius $r \gg M$, and we give it an initial velocity that lies in the equatorial plane. Then symmetry guarantees the body will remain in the equatorial plane: there is no way to prefer going toward north, $\theta < \pi/2$, or toward south, $\theta > \pi/2$. We, further, adjust the initial velocity so the particle remains always at a fixed radius. Then the only nonvanishing components $u^\alpha = dx^\alpha/d\tau$ of the particle's 4-velocity will be $u^t = dt/d\tau$ and $u^\phi = d\phi/d\tau$. The particle's orbit will be governed by the geodesic equation $\nabla_{\vec{u}}\vec{u} = 0$, where $\vec{u}$ is its 4-velocity. The radial component of this geodesic equation, computed in Schwarzschild coordinates, is [cf. Eq. (24.26) with a switch from affine parameter ζ to proper time $\tau = m\zeta$]

$$\frac{d^2r}{d\tau^2} = -\Gamma^r{}_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = -\Gamma^r{}_{tt}\frac{dt}{d\tau}\frac{dt}{d\tau} - \Gamma^r{}_{\phi\phi}\frac{d\phi}{d\tau}\frac{d\phi}{d\tau} \ . \tag{25.10}$$

(Here we have used the vanishing of all $dx^\alpha/d\tau$ except the t and ϕ components, and have used the vanishing of $\Gamma^r{}_{t\phi} = \Gamma^r{}_{\phi t}$ [Eq. (3) of Box 25.2].) Since the orbit is circular, with fixed r, the left side of (25.10) must vanish; and correspondingly the right side gives

$$\frac{d\phi}{dt} = \frac{d\phi/d\tau}{dt/d\tau} = \left(-\frac{\Gamma^r{}_{tt}}{\Gamma^r{}_{\phi\phi}}\right)^{\frac{1}{2}} = \left(\frac{M}{r^3}\right)^{\frac{1}{2}} \ , \tag{25.11}$$

where we have used the values of the connection coefficients from Eq. (3) of Box 25.2, specialized to the equatorial plane $\theta = \pi/2$. Equation (25.11) tells us that the amount of coordinate time t required for the particle to circle the central body once, $0 \le \phi \le 2\pi$, is $\Delta t = 2\pi(r^3/M)^{\frac{1}{2}}$. Since t is the Lorentz time of the body's asymptotic rest frame, this means that observers in the asymptotic rest frame will measure for the particle an orbital period $P = \Delta t = 2\pi(r^3/M)^{\frac{1}{2}}$. This, of course, is the same result as we obtained from the Newtonian limit—but our relativistic analysis shows it to be true for circular orbits of arbitrary radius r, not just for $r \gg M$.

Next we shall move inward, from the asymptotically flat region of Schwarzschild spacetime, toward smaller and smaller radii. As we do so, the spacetime geometry becomes more and more strongly curved, and the Schwarzschild coordinate system becomes less and less Lorentz. As an indication of extreme deviations from Lorentz, notice that the signs of the metric coefficients

$$\frac{\partial}{\partial t}\cdot\frac{\partial}{\partial t} = g_{tt} = -\left(1-\frac{2M}{r}\right)\ , \quad \frac{\partial}{\partial r}\cdot\frac{\partial}{\partial r} = g_{rr} = \frac{1}{(1-2M/r)} \tag{25.12}$$

get reversed as one moves from $r > 2M$ through $r = 2M$ and into the region $r < 2M$. Correspondingly, outside $r = 2M$ world lines of changing t but constant r, θ, ϕ are timelike, while inside $r = 2M$ those world lines are spacelike; and similarly outside $r = 2M$ world lines of changing r but constant t, θ, ϕ are spacelike, while inside they are timelike. In this sense, *outside $r = 2M$, t plays the role of a time coordinate and r the role of a space coordinate; while inside $r = 2M$, t plays the role of a space coordinate and r the role of a time coordinate*. Moreover, this role reversal occurs without any change in the role of r as $1/2\pi$ times the circumference of circles around the center [Eq. (25.3)]. Historically this role reversal presented for many decades severe conceptual problems, even to the best experts on general relativity. We will return to it in Sec. 25.4 below. Henceforth we shall refer to the location of role reversal, $r = 2M$, as the *gravitational radius* of the Schwarzschild spacetime. Throughout the rest of this section and all of Sec. 25.3, we shall confine attention to the region $r > 2M$, outside the gravitational radius. In Sec. 25.4 we shall seek a clear understanding of the "interior" region, $r < 2M$.

Notice that the metric coefficients in the Schwarzschild line element (25.1) are all independent of the coordinate t. This means that the geometry of spacetime itself is invariant under the translation $t \to t +$ constant. At radii $r > 2M$ where t plays the role of a time coordinate, $t \to t +$ constant is a time translation; and, correspondingly, the Schwarzschild spacetime geometry is time-translation-invariant, i.e., "static," outside the gravitational radius.

****

EXERCISES

Exercise 25.1 *Practice: Connection Coefficients and Riemann tensor in the Schwarzschild Metric*

(a) Explain why, for the Schwarzschild metric (25.1), the metric coefficients in the coordinate basis have the values given in Eqs. (2a,b) of Box 25.2.

(b) Using tensor-analysis software on a computer, derive the connection coefficients given in Eq. (3) of Box 25.2.

(c) Show that the basis vectors in Eqs. (4) of Box 25.2 are orthonormal.

(d) Using tensor-analysis software on a computer, derive the connection coefficients (5) and Riemann components (6) of Box 25.2 in the orthonormal basis.

Exercise 25.2 *Example: The Bertotti-Robinson solution of the Einstein field equation*
Bruno Bertotti (1959) and Ivor Robinson (1959) have independently solved the Einstein field equation to obtain the following metric for a universe endowed with a uniform magnetic field:

$$ds^2 = Q^2[-dt^2 + \sin^2 t dz^2 + d\theta^2 + \sin^2\theta d\phi^2] \ . \tag{25.13}$$

Here

$$Q = \text{const} \ , \quad 0 \le t \le \pi \ , \quad -\infty < z < +\infty \ , \quad 0 \le \theta \le \pi \ , \quad 0 \le \phi \le 2\pi \ . \tag{25.14}$$

If one computes the Einstein tensor from the metric coefficients of the line element (25.13) and equates it to 8π times a stress-energy tensor, one finds a stress-energy tensor which is precisely that of an electromagnetic field [Eqs. (23.76) and (23.77)] lifted, unchanged, into general relativity. The electromagnetic field is one which, as measured in the local Lorentz frame of an observer with fixed z, θ, ϕ (a "static" observer), has vanishing electric field and has a magnetic field with magnitude independent of where the observer is located in spacetime and with direction along $\partial/\partial z$. In this sense, the spacetime (25.13) is that of a homogeneous magnetic universe. Discuss the geometry of this universe and the nature of the coordinates t, z, θ, ϕ. More specifically:

(a) Which coordinate increases in a timelike direction and which coordinates in spacelike directions?

(b) Is this universe spherically symmetric?

(c) Is this universe cylindrically symmetric?

(d) Is this universe asymptotically flat?

(e) How does the geometry of this universe change as t ranges from 0 to π. [Hint: show that the curves $\{(z, \theta, \phi) = \text{const}, \ t = \tau/Q\}$ are timelike geodesics—the world lines of the observers referred to above. Then argue from symmetry, or use the result of Ex. 24.4.]

(f) Give as complete a characterization as you can of the coordinates t, z, θ, ϕ.

****

25.3 Static Stars

25.3.1 Birkhoff's Theorem

In 1923, George Birkhoff, a professor of mathematics at Harvard, proved a remarkable theorem:[1] *The Schwarzschild spacetime geometry is the unique spherically symmetric solution*

[1] For a textbook proof see Sec. 32.2 of MTW.

of the vacuum Einstein field equation $\mathbf{G} = 0$. This Birkhoff theorem can be restated in more operational terms as follows: Suppose that you find a solution of the vacuum Einstein field equation, written as a set of metric coefficients $g_{\bar{\alpha}\bar{\beta}}$ in some coordinate system $\{x^{\bar{\mu}}\}$. Suppose, further, that these $g_{\bar{\alpha}\bar{\beta}}(x^{\bar{\mu}})$ exhibit spherical symmetry, but do not coincide with the Schwarzschild expressions [Eqs. (2a) of Box 25.2]. Then Birkhoff guarantees the existence of a coordinate transformation from your coordinates $x^{\bar{\mu}}$ to Schwarzschild's coordinates x^{ν} such that, when that transformation is performed, the resulting new metric components $g_{\alpha\beta}(x^{\nu})$ have precisely the Schwarzschild form [Eq. (2a) of Box 25.2]. For an example see Ex. 25.3. This implies that, thought of as a coordinate-independent spacetime geometry, the Schwarzschild solution is completely unique.

Consider, now, a static, spherically symmetric star (e.g. the sun) residing alone in an otherwise empty universe (or, more realistically, residing in our own universe but so far from all other gravitating matter that we can ignore all other sources of gravity when studying it). Since the star's interior is spherical, it is reasonable to presume that the exterior will be spherical; and since the exterior is also vacuum ($\mathbf{T} = 0$), its spacetime geometry must be that of Schwarzschild. If the circumference of the star's surface is $2\pi R$ and its surface area is $4\pi R^2$, then that surface must reside at the location $r = R$ in the Schwarzschild coordinates of the exterior. In other words, the spacetime geometry will be described by the Schwarzschild line element (25.1) at radii $r > R$, but by something else inside the star, at $r < R$.

Since real atoms with finite rest masses reside on the star's surface, and since such atoms move along timelike world lines, it must be that the world lines $\{r = R,\ \theta = \text{const},\ \phi = \text{const},\ t \text{ varying}\}$ are timelike. From the Schwarzschild invariant interval (25.1) we read off the squared proper time $d\tau^2 = -ds^2 = (1 - 2M/R)dt^2$ along those world lines. This $d\tau^2$ is positive (timelike world line) if and only if $R > 2M$. Thus, *a static star with total mass-energy (active gravitational mass)* M *can never have a circumference smaller than* $2\pi R = 4\pi M$. Restated in conventional units:

$$\boxed{\frac{\text{circumference}}{2\pi} = R \equiv \begin{pmatrix}\text{Radius}\\ \text{of star}\end{pmatrix} > 2M = \frac{2GM}{c^2} = 3.0\,\text{km}\left(\frac{M}{M_\odot}\right) \equiv \begin{pmatrix}\text{gravitational}\\ \text{radius}\end{pmatrix}} \tag{25.15}$$

Here $M_\odot$ is the mass of the sun. The sun satisfies this constraint by a huge margin: $R = 7\times10^5$km. A one-solar-mass white-dwarf star satisfies it by a smaller margin: $R \simeq 6\times10^3$km. And a one-solar-mass neutron star satisfies it by only a modest margin: $R \simeq 10$km. For a pedagogical and detailed discussion see, e.g., Shapiro and Teukolsky (1983).

25.3.2 Stellar Interior

We shall now take a temporary detour away from our study of the Schwarzschild geometry in order to discuss the interior of a static, spherical star. We do so less because of an interest in stars than because the detour will illustrate the process of solving the Einstein field equation and the role of the contracted Bianchi identity in the solution process.

Since the star's spacetime geometry is to be static and spherically symmetric, we can introduce as coordinates in its interior (i) spherical polar angular coordinates θ and ϕ, (ii) a radial coordinate r such that the circumferences of the spheres are $2\pi r$, and (iii) a time

coordinate $\bar{t}$ such that the metric coefficients are independent of $\bar{t}$. By their geometrical definitions, these coordinates will produce a spacetime line element of the form

$$ds^2 = g_{\bar{t}\bar{t}} d\bar{t}^2 + 2 g_{\bar{t}r} d\bar{t} dr + g_{rr} dr^2 + r^2 (d\theta^2 + \sin^2\theta d\phi^2) \ , \tag{25.16}$$

with $g_{\alpha\beta}$ independent of $\bar{t}$, θ, and ϕ. Metric coefficients $g_{\bar{t}\theta}$, $g_{r\theta}$, $g_{\bar{t}\phi}$, $g_{r\phi}$ are absent from (25.16) because they would break the spherical symmetry: they would distinguish the $+\phi$ direction from $-\phi$ or $+\theta$ from $-\theta$ since they would give nonzero values for the scalar products of $\partial/\partial\phi$ or $\partial/\partial\theta$ with $\partial/\partial t$ or $\partial/\partial r$. [Recall: the metric coefficients in a coordinate basis are $g_{\alpha\beta} = \mathbf{g}(\partial/\partial x^\alpha, \partial/\partial x^\beta) = (\partial/\partial x^\alpha)\cdot(\partial/\partial x^\beta)$.] We can get rid of the off-diagonal $g_{\bar{t}r}$ term in the line element (25.16) by specializing the time coordinate: The coordinate transformation

$$\bar{t} = t - \int \left(\frac{g_{\bar{t}r}}{g_{\bar{t}\bar{t}}} \right) dr \ . \tag{25.17}$$

brings the line element into the form

$$\boxed{ds^2 = -e^{2\Phi} dt^2 + e^{2\Lambda} dr^2 + r^2 (d\theta^2 + \sin^2\theta d\phi^2) \ .} \tag{25.18}$$

Here we have introduced the names $e^{2\Phi}$ and $e^{2\Lambda}$ for the time-time and radial-radial metric coefficients. The signs of these coefficients (negative for g_{tt} and positive for g_{rr}) are dictated by the fact that inside the star, as on its surface, real atoms move along world lines of constant r, θ, ϕ and changing t, and thus those world lines must be timelike. The name $e^{2\Phi}$ ties in with the fact that, when gravity is nearly Newtonian the time-time metric coefficient $-e^{2\Phi}$ must reduce to $-(1+2\Phi)$, with Φ the Newtonian potential [Eq. (24.95)]. Thus, the Φ used in (25.18) is a generalization of the Newtonian potential to relativistic, spherical, static gravitational situations.

In order to solve the Einstein field equation for the star's interior, we must specify the stress-energy tensor. Stellar material is excellently approximated by a perfect fluid; and since our star is static, at any point inside the star the fluid's rest frame has constant r, θ, ϕ. Correspondingly, the 4-velocity of the fluid is

$$\boxed{\vec{u} = e^{-\Phi} \frac{\partial}{\partial t} \ .} \tag{25.19}$$

Here the factor $e^{-\Phi}$ guarantees that the 4-velocity will have unit length, as it must. This fluid, of course, is not freely falling. Rather, in order for a fluid element to remain always at fixed r, θ, ϕ, it must accelerate relative to local freely falling observers with a 4-acceleration $\vec{a} \equiv \nabla_{\vec{u}} \vec{u} \neq 0$; i.e., $a^\alpha = u^\alpha{}_{;\mu} u^\mu \neq 0$. Symmetry tells us that this 4-acceleration cannot have any θ or ϕ components; and orthogonality of the 4-acceleration to the 4-velocity tells us that it cannot have any t component. The r component, computed from $a^r = u^r{}_{;\mu} u^\mu = \Gamma^r{}_{00} u^0 u^0$, is $a^r = e^{-2\Lambda} \Phi_{,r}$; and thus,

$$\boxed{\vec{a} = e^{-2\Lambda} \ \Phi_{,r} \frac{\partial}{\partial r} \ .} \tag{25.20}$$

Each fluid element can be thought of as carrying with itself an orthonormal set of basis vectors

$$\boxed{\vec{e}_{\hat{0}} = \vec{u} = e^{-\Phi}\frac{\partial}{\partial t}\ , \quad \vec{e}_{\hat{r}} = e^{-\Lambda}\frac{\partial}{\partial r}\ , \quad \vec{e}_{\hat{\theta}} = \frac{1}{r}\frac{\partial}{\partial \theta}\ , \quad \vec{e}_{\hat{\phi}} = \frac{1}{r\sin\theta}\frac{\partial}{\partial \phi}\ ;} \tag{25.21a}$$

$$\boxed{\vec{e}^{\hat{0}} = e^{\Phi}\vec{\nabla}t\ , \quad \vec{e}^{\hat{r}} = e^{\Lambda}\vec{\nabla}r\ , \quad \vec{e}^{\hat{\theta}} = r\vec{\nabla}\theta\ , \quad \vec{e}^{\hat{\phi}} = r\sin\theta\vec{\nabla}\phi\ .} \tag{25.21b}$$

These basis vectors play two independent roles: (i) One can regard the tangent space of each event in spacetime as being spanned by the basis (25.21), specialized to that event. From this viewpoint, (25.21) constitutes an orthonormal, non-coordinate basis that covers every tangent space of the star's spacetime. This basis is called the fluid's *orthonormal, local-rest-frame basis*. (ii) One can focus attention on a specific fluid element, which moves along the world line $r = r_o$, $\theta = \theta_o$, $\phi = \phi_o$; and one can construct the proper reference frame of that fluid element in the same manner as we constructed the proper reference frame of an accelerated observer in flat spacetime in Sec. 23.5. That proper reference frame is a coordinate system $\{x^{\hat{\alpha}}\}$ whose basis vectors on the fluid element's world line are equal to the basis vectors (25.21):

$$\frac{\partial}{\partial x^{\hat{\mu}}} = \vec{e}_{\hat{\mu}}\ , \quad \vec{\nabla}x^{\hat{\mu}} = \vec{e}^{\hat{\mu}} \ \text{ at } x^{\hat{j}} = 0\ . \tag{25.22}$$

More specifically: the coordinates $x^{\hat{\mu}}$ are given, to second-order in spatial distance from the fluid element's world line, by

$$\begin{aligned} x^{\hat{0}} &= e^{\Phi_o}t\ , \quad x^{\hat{1}} = \int_{r_o}^{r} e^{\Lambda}dr - \frac{1}{2}e^{-\Lambda_o}r_o[(\theta-\theta_o)^2 + \sin^2\theta_o(\phi-\phi_o)^2]\ , \\ x^{\hat{2}} &= r(\theta-\theta_o) - \frac{1}{2}r_o\sin\theta_o\cos\theta_o(\phi-\phi_o)^2\ , \quad x^{\hat{3}} = r\sin\theta(\phi-\phi_o)\ , \end{aligned} \tag{25.23}$$

from which one can verify relation (25.22) with $\vec{e}_{\hat{\mu}}$ and $\vec{e}^{\hat{\mu}}$ given by (25.21). [In Eqs. (25.23) and throughout this discussion all quantities with subscripts $_o$ are evaluated on the fluid's world line.] In terms of the proper-reference-frame coordinates (25.23) the line element (25.18) takes the following form, accurate to first order in distance from the fluid element's world line:

$$ds^2 = -[1 + 2\Phi_{,r}(r-r_o)](dx^{\hat{o}})^2 + \delta_{ij}dx^{\hat{i}}\,dx^{\hat{j}}\ . \tag{25.24}$$

Notice that the quantity $\Phi_{,r}(r-r_o)$ is equal to the scalar product of (i) the spatial separation $\hat{\mathbf{x}} \equiv (r-r_o)\partial/\partial r + (\theta-\theta_o)\partial/\partial\theta + (\phi-\phi_o)\partial/\partial\phi$ of the "field point" (r, θ, ϕ) from the fluid element's world line, with (ii) the fluid's 4-acceleration (25.20), viewed as a spatial 3-vector $\mathbf{a} = e^{-2\Lambda_o}\Phi_{,r}\partial/\partial r$. Correspondingly, the spacetime line element (25.24) in the fluid element's proper reference frame takes the standard proper-reference-frame form (23.87)

$$ds^2 = -(1 + 2\mathbf{a}\cdot\hat{\mathbf{x}})(dx^{\hat{0}})^2 + \delta_{jk}dx^{\hat{j}}\,dx^{\hat{k}}\ , \tag{25.25}$$

accurate to first-order in distance from the fluid element's world line. At second order, as was discussed at the end of Sec. 24.3, there are corrections proportional to the spacetime curvature.

In the local rest frame of the fluid, i.e., when expanded on the fluid's orthonormal rest-frame basis vectors (25.21) or equally well (25.22), the components $T^{\hat{\alpha}\hat{\beta}} = (\rho+P)u^{\hat{\alpha}}u^{\hat{\beta}}+Pg^{\hat{\alpha}\hat{\beta}}$ of the fluid's stress-energy tensor take on the standard form [Eq. (23.58)]

$$\boxed{T^{\hat{0}\hat{0}} = \rho \ , \quad T^{\hat{r}\hat{r}} = T^{\hat{\theta}\hat{\theta}} = T^{\hat{\phi}\hat{\phi}} = P \ ,} \tag{25.26}$$

corresponding to a rest-frame mass-energy density ρ and isotropic pressure P. By contrast with the simplicity of these local-rest-frame components, the contravariant components $T^{\alpha\beta} = (\rho + P)u^\alpha u^\beta + Pg^{\alpha\beta}$ in the (t, r, θ, ϕ) coordinate basis are rather more complicated looking:

$$T^{tt} = e^{-2\Phi}\rho \ , \quad T^{rr} = e^{-2\Lambda}P \ , \quad T^{\theta\theta} = r^{-2}P \ , \quad T^{\phi\phi} = (r\sin\theta)^{-2}P \ . \tag{25.27}$$

This shows one advantage of using orthonormal bases: The components of vectors and tensors are generally simpler in an orthonormal basis than in a coordinate basis. A second advantage occurs when one seeks the physical interpretation of formulae. Because every orthornormal basis is the proper-reference-frame basis of some local observer (the observer with 4-velocity $\vec{u} = \vec{e}_{\hat{0}}$), components measured in such a basis have an immediate physical interpretation. For example, $T^{\hat{0}\hat{0}}$ is the total density of mass-energy measured by the local observer. By contrast, components in a coordinate basis typically do not have a simple physical interpretation.

25.3.3 Local Energy and Momentum Conservation

Before inserting the perfect-fluid stress-energy tensor (25.26) into the Einstein field equation, we shall impose on it the local law of conservation of 4-momentum, $\vec{\nabla}\cdot\mathbf{T} = 0$. In doing so we shall require from the outset that, since the star is to be static and spherical, its density ρ and pressure P must be independent of t, θ, and ϕ; i.e., like the metric coefficients Φ and Λ, they must be functions of radius r only.

The most straightforward way to impose 4-momentum conservation is to equate to zero the quantities

$$T^{\alpha\beta}{}_{;\beta} = \frac{\partial T^{\alpha\beta}}{\partial x^\beta} + \Gamma^\beta{}_{\mu\beta}T^{\alpha\mu} + \Gamma^\alpha{}_{\mu\beta}T^{\mu\beta} = 0 \tag{25.28}$$

in our coordinate basis, making use of expressions (25.27) for the mixed components of the stress-energy tensor, and the connection coefficients and metric components given in Box 25.2.

This straightforward calculation requires a lot of work. Much better is an analysis based on the local proper reference frame of the fluid. The temporal component of $\vec{\nabla}\cdot\mathbf{T} = 0$ in that reference frame, i.e. the projection of this conservation law onto the time basis vector $\vec{e}_{\hat{0}} = e^{-\Phi}\partial/\partial t = \vec{u}$, represents energy conservation as seen by the fluid. But the fluid sees and feels no changes; its density and pressure remain always constant along a fluid element's world line, and energy conservation is therefore guaranteed to be satisfied already; i.e., an evaluation of $\vec{u}\cdot(\vec{\nabla}\cdot\mathbf{T}) = 0$ must give the identity $0 = 0$, so why bother computing it? If one does bother, just to make sure of this argument, one does indeed get $0 = 0$.

The spatial components of $\vec{\nabla}\cdot\mathbf{T}=0$ in the fluid's local rest frame, by contrast, will be nontrivial. The easiest way to compute them is to introduce the tensor $\mathbf{P}\equiv\mathbf{g}+\vec{u}\otimes\vec{u}$ that projects all vectors into the 3-surface orthogonal to $\vec{u}$, i.e. into the fluid's local 3-surface of simultaneity. One can readily show that in the fluid's local proper reference frame, the components of this projection tensor are $P_{\hat{0}\hat{\alpha}}=0$, $P_{\hat{i}\hat{j}}=\delta_{ij}$, which means that $\mathbf{P}$ can be thought of as the spatial 3-metric of the fluid's local rest frame, viewed however as a spacetime tensor. The spatial part of $\vec{\nabla}\cdot\mathbf{T}=0$ is obtained by contraction with $\mathbf{P}$. Computed using index notation, this contraction gives:

$$0=[(\rho+P)u^\alpha u^\beta+Pg^{\alpha\beta}]_{;\beta}P_{\alpha\mu}=[(\rho+P)u^\beta]_{;\beta}u^\alpha P_{\alpha\mu}+(\rho+P)[u^\beta u^\alpha{}_{;\beta}]P_{\alpha\mu}+P_{;\beta}g^{\alpha\beta}P_{\alpha\mu}\ . \tag{25.29}$$

Here we have used the fact that the gradient of the metric vanishes. The first term in Eq. (25.29) vanishes, since $u^\alpha P_{\alpha\mu}=0$ (the projection of $\vec{u}$ orthogonal to $\vec{u}$ is zero). The quantity in square brackets in the second term is the fluid's 4-acceleration $\vec{a}$ [Eq. (25.20)]. The third term is the projection of the pressure gradient orthogonal to $\vec{u}$; but because the star is static, the pressure gradient didn't have any time component to begin with, so the projection accomplishes nothing; it is not needed. Therefore, Eq. (25.29) reduces to

$$(\rho+P)\vec{a}=-\vec{\nabla}P\ . \tag{25.30}$$

Recall from Exs. 1.28 and 23.11(b) that for a perfect fluid $\rho+P$ is the inertial mass per unit volume. Therefore, Eq. (25.30) says that *the fluid's inertial mass per unit volume times its 4-acceleration is equal to the negative of its pressure gradient*. Since both sides of Eq. (25.30) are purely spatially directed as seen in the fluid's local proper reference frame, we can rewrite this equation in 3-dimensional language as

$$\boxed{(\rho+P)\mathbf{a}=-\boldsymbol{\nabla}P\ .} \tag{25.31}$$

A Newtonian physicist, in the proper reference frame, would identify $-\mathbf{a}$ as the local gravitational acceleration, $\boldsymbol{g}$ [cf. Eq. (23.97)], and correspondingly would rewrite (25.30) as

$$\boldsymbol{\nabla}P=(\rho+P)\boldsymbol{g}\ . \tag{25.32}$$

Notice that this is the standard equation of hydrostatic equilibrium for a fluid in an earth-bound laboratory (or swimming pool or lake or ocean), except for the presence of the pressure P in the inertial mass per unit volume. On earth the typical pressures of fluids, even deep in the ocean, are only $P\lesssim 10^9$ dyne/cm$^2\simeq 10^{-12}$ g/cm$^3\lesssim 10^{-12}\rho$; and thus, to extremely good accuracy one can ignore the contribution of pressure to the inertial mass density. However, deep inside a neutron star P may be within a factor 2 of ρ, so the contribution of P cannot be ignored.

We can convert the law of force balance (25.30) into an ordinary differential equation for the pressure P by evaluating its components in the fluid's proper reference frame. The 4-acceleration (25.20) is purely radial; its radial component is $a^{\hat{r}}=e^{-\Lambda}\Phi_{,r}=\Phi_{,\hat{r}}$. The gradient of the pressure is also purely radial and its radial component is $P_{;\hat{r}}=P_{,\hat{r}}=e^{-\Lambda}P_{,r}$. Therefore, the law of force balance reduces to

$$\frac{dP}{dr}=-(\rho+P)\frac{d\Phi}{dr}\ . \tag{25.33}$$

25.3.4 Einstein Field Equation

Turn, now, to the Einstein field equation. In order to impose it, we must first compute in our $\{t, r, \theta, \phi\}$ coordinate system the components of the Einstein tensor $G_{\alpha\beta}$. In general, the Einstein tensor has 10 independent components. However, the symmetries of the line element (25.18) impose identical symmetries on the Einstein tensor computed from it: The only nonzero components will be $G_{\hat{t}\hat{t}}$, $G_{\hat{r}\hat{r}}$, and $G_{\hat{\theta}\hat{\theta}} = G_{\hat{\phi}\hat{\phi}}$; and these three independent components will be functions of radius r only. Correspondingly, the Einstein equation will produce three independent differential equations for our four unknowns: the metric coefficients ("gravitational potentials") Φ and Λ, and the radial distribution of density ρ and pressure P.

These three independent components of the Einstein equation will actually be redundant with the law of hydrostatic equilibrium (25.33). One can see this as follows: If we had not yet imposed the law of 4-momentum conservation, then the Einstein equation $\mathbf{G} = 8\pi\mathbf{T}$, together with the Bianchi identity $\vec{\nabla} \cdot \mathbf{G} \equiv 0$ [Eq. (24.78)], would enforce $\vec{\nabla} \cdot \mathbf{T} = 0$. More explicitly, our three independent components of the Einstein equation together would imply the law of radial force balance, i.e., of hydrostatic equilibrium (25.33). Since we have already imposed (25.33), we need evaluate only two of the three independent components of the Einstein equation; they will give us full information.

A long and rather tedious calculation (best done on a computer), based on the metric coefficients of (25.18) and on Eqs. (23.36)–(23.39), (24.57), (24.54), (24.56), and (24.77) produces for the time-time and radial-radial components of the Einstein tensor, and thence of the Einstein field equation,

$$G^{\hat{0}\hat{0}} = -\frac{1}{r^2}\frac{d}{dr}[r(1 - e^{-2\Lambda})] = 8\pi T^{\hat{0}\hat{0}} = 8\pi\rho \,, \tag{25.34}$$

$$G^{\hat{r}\hat{r}} = -\frac{1}{r^2}(1 - e^{-2\Lambda}) + \frac{2}{r}e^{-2\Lambda}\frac{d\Phi}{dr} = 8\pi T^{\hat{r}\hat{r}} = 8\pi P \,. \tag{25.35}$$

We can bring these components of the field equation into simpler form by defining a new metric coefficient $m(r)$ by

$$\boxed{e^{2\Lambda} \equiv \frac{1}{1 - 2m/r} \,.} \tag{25.36}$$

Note [cf. Eqs. (25.1), (25.18), and (25.36)] that outside the star m is equal to the star's total mass-energy M. This, plus the fact that in terms of m the time-time component of the field equation (25.34) takes the form

$$\frac{dm}{dr} = 4\pi r^2 \rho \,, \tag{25.37}$$

motivates the name *mass inside radius* r for the quantity $m(r)$. In terms of m the radial-radial component (25.35) of the field equation becomes

$$\boxed{\frac{d\Phi}{dr} = \frac{m + 4\pi r^3 P}{r(r - 2m)} \,;} \tag{25.38}$$

and combining this with (25.33) we obtain an alternative form of the equation of hydrostatic equilibrium

$$\boxed{\frac{dP}{dr} = -\frac{(\rho + P)(m + 4\pi r^3 P)}{r(r - 2m)} \, .} \tag{25.39}$$

[This form is called the Tolman-Oppenheimer-Volkoff or TOV equation because it was first derived by Tolman (1939) and first used in a practical calculation by Oppenheimer and Volkoff (1939).] Equations (25.37), (25.38), (25.39) plus an equation of state for the pressure of the stellar material P in terms of its density of total mass-energy ρ,

$$P = P(\rho) \, , \tag{25.40}$$

determine the four quantities Φ, m, ρ, and P as functions of radius.

Actually, for full determination, one also needs boundary conditions. Just as the surface of a sphere is everywhere locally Euclidean (i.e., is arbitrarily close to Euclidean in arbitrarily small regions), so also spacetime must be everywhere locally Lorentz; cf. Eqs. (24.15) and (24.16). In order that spacetime be locally Lorentz at the star's center (in particular, that circumferences of tiny circles around the center be equal to 2π times their radii), it is necessary that m vanish at the center

$$m = 0 \quad \text{at } r = 0, \quad \text{and thus} \quad \boxed{m(r) = \int_0^r 4\pi r^2 \rho dr \, ;} \tag{25.41}$$

cf. Eqs. (25.18) and (25.36). At the star's surface the interior spacetime geometry (25.18) must join smoothly to the exterior Schwarzschild geometry (25.1), and hence

$$\boxed{m = M \quad \text{and } e^{2\Phi} = 1 - 2M/r \quad \text{at } r = R \, .} \tag{25.42}$$

25.3.5 Stellar Models and Their Properties

A little thought now reveals a straightforward method of producing a relativistic stellar model: (*i*) Specify an equation of state for the stellar material $P = P(\rho)$ and specify a central density ρ_c or central pressure P_c for the star. (*ii*) Integrate the coupled hydrostatic-equilibrium equation (25.39) and "mass equation" (25.37) outward from the center, beginning with the initial conditions $m = 0$ and $P = P_c$ at the center. (*iii*) Terminate the integration when the pressure falls to zero; this is the surface of the star. (*iv*) At the surface read off the value of m; it is the star's total mass-energy M, which appears in the star's external, Schwarzschild line element (25.1). (*v*) From this M and the radius $r \equiv R$ of the star's surface, read off the value of the gravitational potential Φ at the surface [Eq. (25.42)]. (*vi*) Integrate the Einstein field equation (25.38) inward from the surface toward the center to determine Φ as a function of radius inside the star.

Just six weeks after reading to the Prussian Academy of Science the letter in which Karl Schwarzschild derived his vacuum solution (25.1) of the field equation, Albert Einstein again presented the Academy with the results of Schwarzschild's fertile mind: an exact solution for the structure of the interior of a star that has constant density ρ. [And just four months

after that, on June 29, 1916, Einstein had the sad task of announcing to the Academy that Schwarzschild had died of an illness contracted on the Russian front.]

In our notation, Schwarzschild's solution for the interior of a star is characterized by its uniform density ρ, its total mass M, and its radius R which is given in terms of ρ and M by

$$M = \frac{4\pi}{3}\rho R^3 \tag{25.43}$$

[Eq. (25.41)]. In terms of these the mass inside radius r, the pressure P, and the gravitational potential Φ are (Schwarzschild 1916b)

$$m = \frac{4\pi}{3}\rho r^3 \ , \quad P = \rho\left[\frac{(1-2Mr^2/R^3)^{\frac{1}{2}} - (1-2M/R)^{\frac{1}{2}}}{3(1-2M/R)^{\frac{1}{2}} - (1-2Mr^2/R^3)^{\frac{1}{2}}}\right] \ , \tag{25.44}$$

$$e^{\Phi} = \frac{3}{2}\left(1-\frac{2M}{R}\right)^{\frac{1}{2}} - \frac{1}{2}\left(1-\frac{2Mr^2}{R^3}\right)^{\frac{1}{2}} \ . \tag{25.45}$$

We present these details less for their specific physical content than to illustrate the solution of the Einstein field equation in a realistic, astrophysically interesting situation. For discussions of the application of this formalism to neutron stars, where relativistic deviations from Newtonian theory can be rather strong, see e.g., Shapiro and Teukolsky (1983). For the seminal work on the theory of neutron-star structure see Oppenheimer and Volkoff (1939).

Among the remarkable consequences of the TOV equation of hydrostatic equilibrium (25.39) for neutron-star structure are these: (*i*) If the mass m inside radius r ever gets close to $r/2$, the "gravitational pull" [right-hand side of (25.39)] becomes divergently large, forcing the pressure gradient that counterbalances it to be divergently large, and thereby driving the pressure quickly to zero as one integrates outward. This protects the static star from having M greater than $R/2$, i.e., from having its surface inside its gravitational radius. (*ii*) Although the density of matter near the center of a neutron star is above that of an atomic nucleus, where the equation of state is ill-understood, we can be confident that there is an upper limit on the masses of neutron stars, a limit in the range $1.6M_\odot \lesssim M_{\rm max} \lesssim 3M_\odot$. This mass limit cannot be avoided by postulating that a more massive neutron star develops an arbitrarily large central pressure and thereby supports itself against gravitational implosion. The reason is that an arbitrarily large central pressure is self-defeating: The "gravitational pull" which appears on the right-hand side of (25.39) is quadratic in the pressure at very high pressures (whereas it would be independent of pressure in Newtonian theory). This purely relativistic feature guarantees that if a star develops too high a central pressure, it will be unable to support itself against the resulting "quadratically too high" gravitational pull.

We conclude this section by introducing a useful technique for visualizing spacetime curvature: the embedding of the curved spacetime, or a piece of it, in a flat space of higher dimensionality.

The geometry of a curved, n-dimensional manifold is characterized by $\frac{1}{2}n(n+1)$ metric components (since those components form a symmetric $n \times n$ matrix), of which only $\frac{1}{2}n(n+1) - n = \frac{1}{2}n(n-1)$ are of coordinate-independent significance (since we are free to choose arbitrarily the n coordinates of our coordinate system and can thereby force n of the metric

components to take on any desired values, e.g., zero). If this n-dimensional manifold is embedded in a flat N-dimensional manifold, that embedding will be described by expressing $N-n$ of the embedding manifold's Euclidean (or Lorentz) coordinates in terms of the other n. Thus, the embedding will be characterized by $N-n$ functions of n variables. In order for the embedding to be possible, in general, this number of choosable functions must be at least as large as the number of significant metric coefficients $\frac{1}{2}n(n-1)$. From this argument we conclude that *the dimensionality of the embedding space must be* $N \geq \frac{1}{2}n(n+1)$. Actually, this argument analyzes only the local features of the embedding. If one wants also to preserve the global topology of the n-dimensional manifold, one must in general go to an embedding space of even higher dimensionality.

Curved spacetime has $n = 4$ dimensions and thus requires for its local embedding a flat space with $N = 10$ dimensions. This is a bit much for 3-dimensional beings like us to visualize. If, as a sop to our visual limitations, we reduce our ambitions and seek only to extract a 3-surface from curved spacetime and visualize it by embedding it in a flat space, we will require a flat space of $N = 6$ dimensions. This is still a bit much. In frustration we are driven to extract from spacetime $n = 2$ dimensional surfaces and visualize them by embedding in flat spaces with $N = 3$ dimensions. This is doable—and, indeed, instructive.

As a nice example, consider the equatorial "plane" through the spacetime of a static spherical star, at a specific "moment" of coordinate time t; i.e., consider the 2-surface $t = \text{const}$, $\theta = \pi/2$ in the spacetime of Eqs. (25.18), (25.36). The line element on this equatorial 2-surface is

$$^{(2)}ds^2 = \frac{dr^2}{1-2m/r} + r^2 d\phi^2 \ , \quad \text{where} \ \ m = m(r) = \int_0^r 4\pi r^2 \rho dr \ ; \tag{25.46}$$

cf. Eq. (25.41). We seek to construct in a 3-dimensional Euclidean space a 2-dimensional surface with precisely this same 2-geometry. As an aid, introduce in the Euclidean embedding space a cylindrical coordinate system r, z, ϕ, in terms of which the space's 3-dimensional line element is

$$^{(3)}ds^2 = dr^2 + dz^2 + r d\phi^2 \ . \tag{25.47}$$

The surface we seek to embed is axially symmetric, so we can describe its embedding by the value of z on it as a function of radius r: $z = z(r)$. Inserting this (unknown) embedding function into (25.47), we obtain for the surface's 2-geometry

$$^{(2)}ds^2 = [1 + (dz/dr)^2]dr^2 + r^2 d\phi^2 \ ; \tag{25.48}$$

and comparing with our original expression (25.46) for the 2-geometry we obtain a differential equation for the embedding function:

$$\frac{dz}{dr} = \left(\frac{1}{1-2m/r} - 1 \right)^{\frac{1}{2}} \ . \tag{25.49}$$

If we set $z = 0$ at the star's center, then the solution of this differential equation is

$$\boxed{z = \int_0^r \frac{dr}{[(r/2m) - 1]^{\frac{1}{2}}} \ .} \tag{25.50}$$

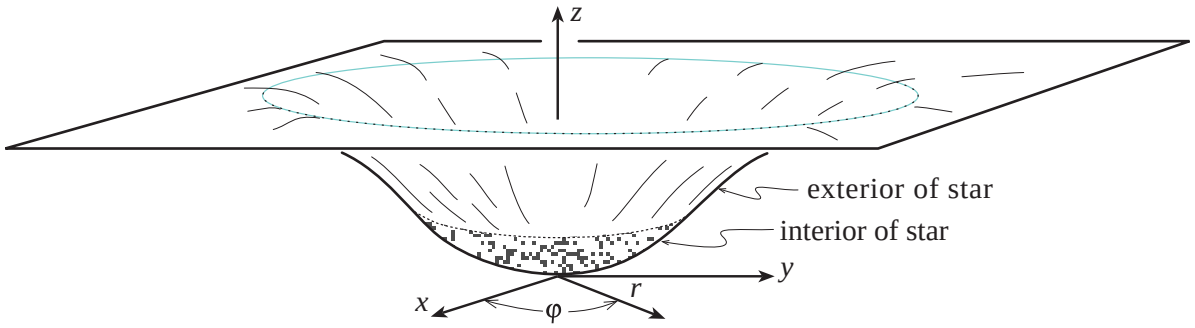


Fig. 25.1: Embedding diagram depicting an equatorial, 2-dimensional slice $t = \text{const}$, $\theta = \pi/2$ through the spacetime of a spherical star with uniform density ρ and with radius R equal to 2.5 times the gravitational radius $2M$. See Ex. 25.4 for details.

Near the star's center $m(r)$ is given by $m = (4\pi/3)\rho_c r^3$, where ρ_c is the star's central density; and outside the star $m(r)$ is equal to the star's r-independent total mass M. Correspondingly, in these two regions Eq. (25.50) reduces to

$$z = \sqrt{(2\pi/3)\rho_c}\, r^2 \quad \text{at } r \text{ very near zero .}$$
$$z = \sqrt{8M(r-2M)} + \text{constant} \quad \text{at } r > R\,, \quad \text{i.e., outside the star.} \tag{25.51}$$

Figure 25.1 shows the embedded 2-surface $z(r)$ for a star of uniform density $\rho = \text{const}$; cf. Ex. 25.4. For any other star the embedding diagram will be qualitatively similar, though quantitatively different.

The most important feature of this embedding diagram is its illustration of the fact [also clear in the original line element (25.46)] that, as one moves outward from the star's center, its circumference $2\pi r$ increases less rapidly than the proper radial distance travelled, $l = \int_0^r (1-2m/r)^{-\frac{1}{2}} dr$. As a specific example, the distance from the center of the earth to a perfect circle near the earth's surface is more than circumference/2π by about 1.5 millimeters—a number whose smallness compared to the actual radius, 6.4×10^8 cm, is a measure of the weakness of the curvature of spacetime near earth. As a more extreme example, the distance from the center of a massive neutron star to its surface is about one kilometer greater than circumference/2π—i.e., greater by an amount that is roughly 10 percent of the ~ 10 km circumference/2π. Correspondingly, in the embedding diagram for the earth (Fig. 25.1) the embedded surface would be so nearly flat that its downward dip at the center would be noticeable only with great effort; whereas the embedding diagram for a neutron star would show a downward dip about like that of Fig. 25.1.

****

EXERCISES

Exercise 25.3 *Example: Schwarzschild Geometry in Isotropic Coordinates*

(a) It turns out that the following line element is a solution of the vacuum Einstein field

equation $\mathbf{G} = 0$:

$$\boxed{ds^2 = -\left(\frac{1 - M/2\bar{r}}{1 + M/2\bar{r}}\right)^2 dt^2 + \left(1 + \frac{M}{2\bar{r}}\right)^4 [d\bar{r}^2 + \bar{r}^2(d\theta^2 + \sin^2\theta d\phi^2)] \; .} \qquad (25.52)$$

Since this solution is spherically symmetric, Birkhoff's theorem guarantees it must represent the standard Schwarzschild spacetime geometry in a coordinate system that differs from Schwarzschild's. Show that this is so by exhibiting a coordinate transformation that converts this line element into (25.1). Note: the t, $\bar{r}$, θ, ϕ coordinates are called *isotropic* because in them the spatial part of the line element is a function of $\bar{r}$ times the 3-dimensional Euclidean line element, and Euclidean geometry picks out at each point in space no preferred spatial directions, i.e., it is isotropic.

(b) Show that at large radii $r \gg M$, the line element (25.52) takes the form (24.112) discussed in Chap. 24, but with vanishing spin angular momentum $\mathbf{S} = 0$.

Exercise 25.4 *Example: Star of Uniform Density*

(a) Show that the embedding surface of Eq. (25.50) is a paraboloid of revolution everywhere outside the star.

(b) Show that in the interior of a uniform-density star, the embedding surface is a segment of a sphere.

(c) Show that the match of the interior to the exterior is done in such a way that, in the embedding space the embedded surface shows no kink (no bend) at $r = R$.

(d) Show that circumference/2π for a star is less than the distance from the center to the surface by an amount of order the star's Schwarzschild radius $2M$. Evaluate this amount analytically for a star of uniform density, and numerically (approximately) for the earth and for a neutron star.

Exercise 25.5 *Example: Gravitational Redshift*

Consider a photon emitted by an atom at rest on the surface of a static star with mass M and radius R. Analyze the photon's motion in the Schwarzschild coordinate system of the star's exterior, $r \geq R > 2M$; and, in particular, compute the "gravitational redshift" of the photon by the following steps:

(a) Since the emitting atom is very nearly an "ideal clock," it gives the emitted photon very nearly the same frequency $\nu_{\rm em}$, as measured in the atom's proper reference frame, as it would give were it in an earth laboratory or floating in free space. Thus, the proper reference frame of the emitting atom is central to a discussion of the photon's properties and behavior. Show that the basis vectors of that proper reference frame are

$$\vec{e}_{\hat{0}} = \frac{1}{\sqrt{1 - 2M/r}} \frac{\partial}{\partial t} \; , \quad \vec{e}_{\hat{r}} = \sqrt{1 - 2M/r} \frac{\partial}{\partial r} \; , \quad \vec{e}_{\hat{\theta}} = \frac{1}{r} \frac{\partial}{\partial \theta} \; , \quad \vec{e}_{\hat{\phi}} = \frac{1}{r \sin\theta} \frac{\partial}{\partial \phi} \; . \qquad (25.53)$$

As part of your proof, show that these basis vectors are orthonormal.

(b) Explain why $h\nu_{\rm em} = -p_{\hat{0}} = -\vec{p}\cdot\vec{e}_{\hat{0}}$ at the moment of photon emission. (Here and below h is Planck's constant and $\vec{p}$ is the photon's 4-momentum.)

(c) Show that the time-component of the photon 4-momentum in the Schwarzschild coordinate basis is $p_t = -\sqrt{1-2M/R}\,h\nu_{\rm em}$ at the moment of emission.

(d) Show that as the photon flies out (radially or nonradially) toward $r = \infty$, the coordinate-time component of its 4-momentum, p_t, is conserved. [*Hint*: recall the result of Ex. 24.4(a).]

(e) Show that when received by an observer at rest relative to the star and very far away from it, the photon is measured by that observer to have frequency $\nu_{\rm rec} = -p_t/h$.

(f) Show that the photon is redshifted by an amount

$$\boxed{\frac{\lambda_{\rm rec}-\lambda_{\rm em}}{\lambda_{\rm em}} = \frac{1}{\sqrt{1-2M/R}} - 1\ ,} \tag{25.54}$$

where $\lambda_{\rm rec}$ is the wavelength that the photon's spectral line exhibits at the receiver and $\lambda_{\rm em}$ is the wavelength that the emitting kind of atom would produce in an earth laboratory. Note that for a nearly Newtonian star, i.e. one with $R \gg M$, this redshift becomes $\simeq M/R = GM/Rc^2$.

(g) Evaluate this redshift for the earth, for the sun, and for a 1.4-solar-mass, 10-kilometer-radius neutron star.

Exercise 25.6 *Challenge: Mass-Radius Relation for Real Neutron Stars*
Choose a physical equation of state from the alternatives presented in Shapiro & Teukolsky (1983) and represent it numerically. Then integrate the TOV equation starting with several suitable central pressures and deduce a mass-radius relation. You should find that as the central pressure is increased, the mass passes through a maximum while the radius continues to *decrease*. (Solutions with radii smaller than that associated with the maximum mass are unstable to radial perturbations.)

****

25.4 Gravitational Implosion of a Star to Form a Black Hole

J. Robert Oppenheimer (then a professor jointly at the University of California at Berkeley and at Caltech), upon discovering with his student George Volkoff that there is a maximum mass limit for neutron stars (Oppenheimer and Volkoff 1939), was forced to consider the possibility that when it exhausts its nuclear fuel a more massive star will implode to radii $R \leq 2M$. With his graduate student Hartland Snyder, Oppenheimer just before the outbreak of

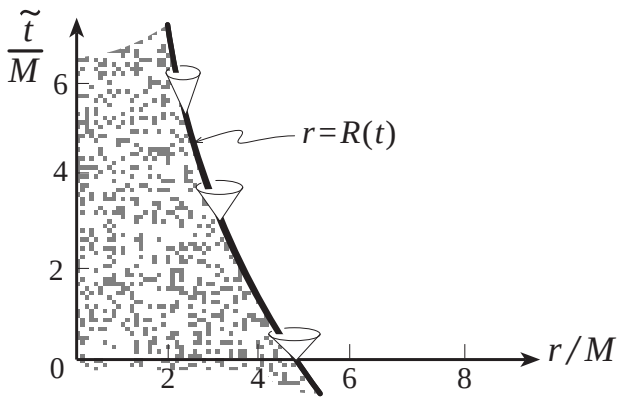


Fig. 25.2: Spacetime diagram depicting in Schwarzschild coordinates the gravitationally induced implosion of a star. The thick solid curve is the world line of the star's surface, $r = R(t)$ in the external Schwarzschild coordinates. The stippled region to the left of that world line is not correctly described by the Schwarzschild line element (25.1); it requires for its description the spacetime metric of the star's interior.

World War II investigated the details of such an implosion for the idealized case of a perfectly spherical star in which all the internal pressure is suddenly extinguished; see Oppenheimer and Snyder (1939). In this section we shall repeat their analysis, though from a more modern viewpoint and using somewhat different arguments.[2]

By Birkhoff's theorem, the spacetime geometry outside an imploding, spherical star must be that of Schwarzschild. This means, in particular, that an imploding, spherical star cannot produce any gravitational waves; such waves would break the spherical symmetry. By contrast, a star that implodes nonspherically can produce a strong burst of gravitational waves; see Chap. 26.

Since the spacetime geometry outside an imploding, spherical star is that of Schwarzschild, we can depict the motion of the star's surface by a world line in a 2-dimensional spacetime diagram with Schwarzschild coordinate time t plotted upward and Schwarzschild coordinate radius r plotted rightward (Fig. 25.2). The world line of the star's surface is an ingoing curve. The region to the left of the world line must be discarded and replaced by the spacetime of the star's interior, while the region to the right, $r > R(t)$, is correctly described by Schwarzschild.

As for a static star, so also for an imploding one, because real atoms with finite rest masses live on the star's surface, the world line of that surface, $\{r = R(t)$, θ and ϕ constant$\}$, must be timelike. Consequently, at each point along the world line it must lie within the local light cones. Let us examine those light cones:

Several of the photon world lines (25.56) are depicted in Fig. 25.3, along with some of the local light cones (25.55). The light cones are drawn with one of the two θ, ϕ angular coordinates restored. The most extreme rightward and most extreme leftward edges of each cone are short segments of the radial photon trajectories, as given by Eq. (25.55); while the in-between parts are segments of trajectories of photons with nonzero $d\theta$ and/or $d\phi$. Notice that the light cones do not have 45-degree opening angles as they do in a Lorentz frame of

[2]For further details, see MTW Chapters 31 and 32.

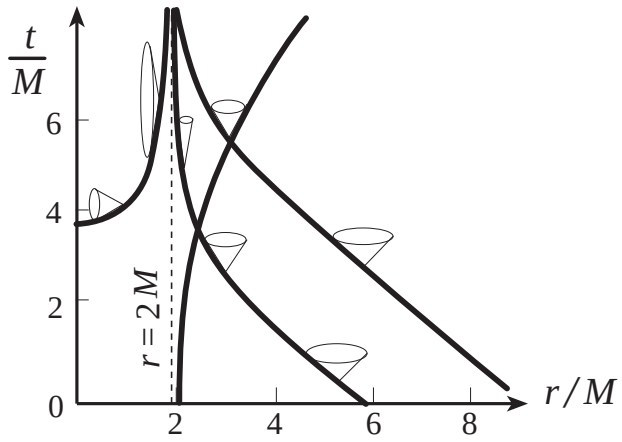


Fig. 25.3: Some radial photon world lines [Eq. (25.56)] and some light cones [Eq. (25.55)] in the Schwarzschild spacetime, depicted in Schwarzschild coordinates. We draw the light cones inside $r = 2M$ opening leftward rather than rightward for reasons explained in Sec. 25.4 below [cf. Fig. 25.5].

flat spacetime. This is a peculiarity due not to spacetime curvature, but rather to the nature of the Schwarzschild coordinates: If, at any chosen event of the Schwarzschild spacetime, we were to introduce a local Lorentz frame, then in that frame the light cones would have 45-degree opening angles. Thus, the "squeezing down" of the light cones as one approaches $r = 2M$ from $r > 2M$, in Fig. 25.3, signals not a peculiarity of the frame-independent spacetime geometry at $r = 2M$, but rather a peculiarity of the Schwarzschild coordinates there.

The radial edges of the light cones are generated by the world lines of radially traveling photons, i.e., photons with world lines of constant θ, ϕ and varying t, r. (Spherical symmetry dictates that if a photon starts out traveling radially, it will always continue to travel radially.) We could, if we wished, compute the world lines of these photons from their geodesic equation. However, knowing already that they have constant θ, ϕ we can compute them more simply from a knowledge that they must be null world lines: Setting to zero the ds^2 of Eq. (25.1), we see that along these null, radial world lines

$$0 = ds^2 = -(1 - 2M/r)dt^2 + \frac{dr^2}{1 - 2M/r} \; ; \quad \text{i.e.,} \; \frac{dt}{dr} = \pm\frac{1}{1 - 2M/r} \; . \tag{25.55}$$

Integrating this differential equation we obtain

$$r + 2M \ln|(r/2M) - 1| = \pm t + \text{const} \; . \tag{25.56}$$

Since the world line of the star's surface is confined to the interiors of the local light cones, *the squeezing down of the light cones near* $r = 2M$ [the fact that $dt/dr \to \pm\infty$ in Eq. (25.55)] *prevents the star's world line* $r = R(t)$ *from ever, in any finite coordinate time* t*, reaching the gravitational radius,* $r = 2M$*.*

This conclusion is completely general; it relies in no way whatsoever on the details of what is going on inside the star or at its surface. It is just as valid for completely realistic

stellar implosion (with finite pressure and shock waves) as for the idealized, Oppenheimer-Snyder case of zero-pressure implosion. In the special case of zero pressure, one can explore the details further:

Because no pressure forces act on the atoms at the star's surface, those atoms must move inward along radial geodesic world lines. Correspondingly, the world line of the star's surface in the external Schwarzschild spacetime must be a timelike geodesic of constant (θ, ϕ). In Ex. 25.7 the geodesic equation is solved to determine that world line $R(t)$, with a conclusion that agrees with the above argument: *Only after a lapse of infinite coordinate time t does the star's surface reach the gravitational radius $r = 2M$.* A byproduct of that calculation is equally remarkable: Although the implosion to $R = 2M$ requires infinite Schwarzschild coordinate time t, it requires only a finite proper time τ as measured by an observer who rides inward on the star's surface. In fact, the proper time is

$$\boxed{\tau \simeq \frac{\pi}{2}\left(\frac{R_o^3}{2M}\right)^{\frac{1}{2}} = 15\,\text{microseconds}\left(\frac{R_o}{2M}\right)^{3/2}\frac{M}{M_\odot} \quad \text{if } R_o \gg 2M\ ,} \tag{25.57}$$

where R_o is the star's initial radius when it first begins to implode freely, $M_\odot$ denotes the mass of the sun, and proper time τ is measured from the start of implosion. Note that this implosion time is equal to $1/4\sqrt{2}$ times the orbital period of a test particle at the radius of the star's initial surface. For a star with mass and initial radius equal to those of the sun, τ is about 30 minutes; for a neutron star that has been pushed over the maximum mass limit by accretion of matter from its surroundings, τ is about 0.1 milliseconds.

What happens to the star's surface, and an observer on it, when—after infinite coordinate time but tiny proper time—it reaches the gravitational radius? There are two possibilities: (i) the tidal gravitational forces there might be so strong that they destroy the star's surface and any observers on it; or, (ii) if the tidal forces are not that strong, then the star and observers must continue to exist, moving into a region of spacetime (presumably $r < 2M$) that is not smoothly joined onto $r > 2M$ in the Schwarzschild coordinate system. In the latter case the pathology is all due to poor properties of Schwarzschild's coordinates. In the former case it is due to an intrinsic, coordinate-independent singularity of the tide-producing Riemann curvature.

To see which is the case, we must evaluate the tidal forces felt by observers on the surface of the imploding star. Those tidal forces are produced by the Riemann curvature tensor. More specifically, if an observer's feet and head have a vector separation $\boldsymbol{\xi}$ at time τ as measured by the observer's clock, then the curvature of spacetime will exert on them a relative gravitational acceleration given by the equation of geodesic deviation, in the form appropriate to a local Lorentz frame:

$$\frac{d^2\xi^{\bar{j}}}{d\tau^2} = -R^{\bar{j}}{}_{\bar{0}\bar{k}\bar{0}}\xi^{\bar{k}} \tag{25.58}$$

[Eq. (24.42)]. Here the barred indices denote components in the observer's local Lorentz frame. The tidal forces will become infinite, and will thereby destroy the observer and all forms of matter on the star's surface, if and only if the local Lorentz Riemann components $R_{\bar{j}\bar{0}\bar{k}\bar{0}}$ diverge as the star's surface approaches the gravitational radius. Thus, to test whether

the observer and star survive, we must compute the components of the Riemann curvature tensor in the local Lorentz frame of the star's imploding surface.

The easiest way to compute those components is by a transformation from components as measured in the proper reference frames of observers who are "at rest" (fixed r, θ, ϕ) in the Schwarzschild spacetime. At each event on the world tube of the star's surface, then, we have two orthonormal frames: one (barred indices) a local Lorentz frame imploding with the star; the other (hatted indices) a proper reference frame at rest. Since the metric coefficients in these two bases have the standard flat-space form $g_{\bar{\alpha}\bar{\beta}} = \eta_{\alpha\beta}$, $g_{\hat{\alpha}\hat{\beta}} = \eta_{\alpha\beta}$, the bases must be related by a Lorentz transformation [cf. Eq. (1.47b) and associated discussion]. A little thought makes it clear that the required transformation matrix is that for a pure boost [Eq. (1.49a)]

$$L^{\hat{0}}{}_{\bar{0}} = L^{\hat{r}}{}_{\bar{r}} = \gamma \ , \quad L^{\hat{0}}{}_{\bar{r}} = L^{\hat{r}}{}_{\bar{0}} = -\beta\gamma \ , \quad L^{\hat{\theta}}{}_{\bar{\theta}} = L^{\hat{\phi}}{}_{\bar{\phi}} = 1 \ ; \quad \gamma = \frac{1}{\sqrt{1-\beta^2}} \ , \tag{25.59}$$

with β the speed of implosion of the star's surface, as measured in the proper reference frame of the static observer when the surface flies by. The transformation law for the components of the Riemann tensor has, of course, the standard form for any fourth rank tensor:

$$R_{\bar{\alpha}\bar{\beta}\bar{\gamma}\bar{\delta}} = L^{\hat{\mu}}{}_{\bar{\alpha}} L^{\hat{\nu}}{}_{\bar{\beta}} L^{\hat{\lambda}}{}_{\bar{\gamma}} L^{\hat{\sigma}}{}_{\bar{\delta}} R_{\hat{\mu}\hat{\nu}\hat{\lambda}\hat{\sigma}} \ . \tag{25.60}$$

The basis vectors of the proper reference frame are given by Eq. (25.21), specialized to the star's Schwarzschild exterior and with r set equal to the momentary radius R of the star's surface [cf. also Ex. 25.5(a)]

$$\vec{e}_{\hat{0}} = \frac{1}{\sqrt{1-2M/R}} \frac{\partial}{\partial t} \ , \quad \vec{e}_{\hat{r}} = \sqrt{1-2M/R} \frac{\partial}{\partial r} \ , \quad \vec{e}_{\hat{\theta}} = \frac{1}{R} \frac{\partial}{\partial \theta} \ , \quad \vec{e}_{\hat{\phi}} = \frac{1}{R \sin\theta} \frac{\partial}{\partial \phi} \ . \tag{25.61}$$

This is the Schwarzschild orthonormal basis used in Box 25.2; and from that Box we learn that the components of Riemann in this basis are:

$$\begin{aligned} R_{\hat{0}\hat{r}\hat{0}\hat{r}} &= -\frac{2M}{R^3} \ , \quad R_{\hat{0}\hat{\theta}\hat{0}\hat{\theta}} = R_{\hat{0}\hat{\phi}\hat{0}\hat{\phi}} = +\frac{M}{R^3} \ , \\ R_{\hat{\theta}\hat{\phi}\hat{\theta}\hat{\phi}} &= \frac{2M}{R^3} \ , \quad R_{\hat{r}\hat{\theta}\hat{r}\hat{\theta}} = R_{\hat{r}\hat{\phi}\hat{r}\hat{\phi}} = -\frac{M}{R^3} \ . \end{aligned} \tag{25.62}$$

These are the components measured by static observers.

By inserting these static-observer components and the Lorentz-transformation matrix (25.59) into the transformation law (25.60) we reach our goal: The following components of Riemann in the local Lorentz frame of the star's freely imploding surface:

$$\begin{aligned} R_{\bar{0}\bar{r}\bar{0}\bar{r}} &= -\frac{2M}{R^3} \ , \quad R_{\bar{0}\bar{\theta}\bar{0}\bar{\theta}} = R_{\bar{0}\bar{\phi}\bar{0}\bar{\phi}} = +\frac{M}{R^3} \ , \\ R_{\bar{\theta}\bar{\phi}\bar{\theta}\bar{\phi}} &= \frac{2M}{R^3} \ , \quad R_{\bar{r}\bar{\theta}\bar{r}\bar{\theta}} = R_{\bar{r}\bar{\phi}\bar{r}\bar{\phi}} = -\frac{M}{R^3} \ . \end{aligned} \tag{25.63}$$

These components are remarkable in two ways: First, they remain perfectly finite as the star's surface approaches the gravitational radius; and, correspondingly, tidal gravity cannot

destroy the star or the observers on its surface. Second, the components of Riemann are identically the same in the two orthonormal frames, hatted and barred, which move radially at finite speed β with respect to each other [expressions (25.63) are independent of β and are the same as (25.62)]. This is a result of the very special algebraic structure that Riemann's components have for the Schwarzschild spacetime; it will not be true in typical spacetimes.

From the finiteness of the components of Riemann in the surface's local Lorentz frame, we conclude that something must be wrong with Schwarzschild's t, r, θ, ϕ coordinate system in the vicinity of the gravitational radius $r = 2M$: Although nothing catastrophic happens to the star's surface as it approaches $2M$, those coordinates refuse to describe passage through $r = 2M$ in a reasonable, smooth, finite way. Thus, in order to study the implosion as it passes through the gravitational radius and beyond, we shall need a new, improved coordinate system.

Several coordinate systems have been devised for this purpose. For a study and comparison of them see, e.g., Chap. 31 of MTW. In this chapter we shall confine ourselves to one: A coordinate system devised for other purposes by Arthur Eddington (1922), then long forgotten and only rediscovered independently and used for this purpose by David Finkelstein (1958). Yevgeny Lifshitz, of Landau-Lifshitz fame, told one of the authors many years later what an enormous impact Finkelstein's coordinate system had on peoples' understanding of the implosion of stars. "You cannot appreciate how difficult it was for the human mind before Finkelstein to understand [the Oppenheimer-Snyder analysis of stellar implosion]." Lifshitz said. When, nineteen years after Oppenheimer and Snyder, the issue of the Physical Review containing Finkelstein's paper arrived in Moscow, suddenly everything was clear.

Finkelstein, a postdoctoral fellow at the Stevens Institute of Technology in Hoboken, New Jersey, found the following simple transformation which moves the region $t = \infty$, $r = 2M$ of Schwarzschild coordinates in to a finite location. His transformation involves introducing a new time coordinate

$$\boxed{\tilde{t} = t + 2M \ln |(r/2M) - 1| \, ,} \tag{25.64}$$

but leaving unchanged the radial and angular coordinates. Figure 25.4 shows the surfaces of constant Eddington-Finkelstein time $\tilde{t}$ in Schwarzschild coordinates, and the surfaces of constant Schwarzschild time t in Eddington-Finkelstein coordinates. Notice, as advertised, that $t = \infty$, $r = 2M$ is moved to a finite Eddington-Finkelstein location.

By inserting the coordinate transformation (25.64) into the Schwarzschild line element (25.1) we obtain the following line element for Schwarzschild spacetime written in Eddington-Finkelstein coordinates:

$$\boxed{ds^2 = -\left(1 - \frac{2M}{r}\right) d\tilde{t}^{\,2} + \frac{4M}{r} d\tilde{t}\, dr + \left(1 + \frac{2M}{r}\right) dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2) \, .} \tag{25.65}$$

Notice that, by contrast with the line element in Schwarzschild coordinates, none of the metric coefficients diverge as r approaches $2M$. Moreover, in an Eddington-Finkelstein spacetime diagram, by contrast with Schwarzschild, the light cones do not pinch down to slivers at $r = 2M$ [compare Figs. 25.5(a) and 25.5(b)]: The world lines of radial light rays are computable in Eddington-Finkelstein, as in Schwarzschild, by setting $ds^2 = 0$ (null world

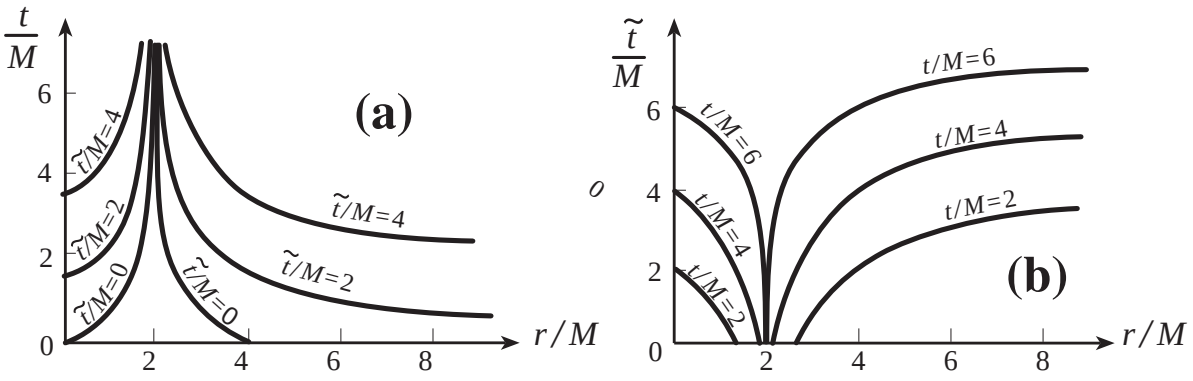


Fig. 25.4: (a) The 3-surfaces of constant Eddington-Finkelstein time coordinate $\tilde{t}$ drawn in a Schwarzschild spacetime diagram, with the angular coordinates θ, ϕ suppressed. (b) The 3-surfaces of constant Schwarzschild time coordinate t drawn in an Eddington-Finkelstein spacetime diagram, with angular coordinates suppressed.

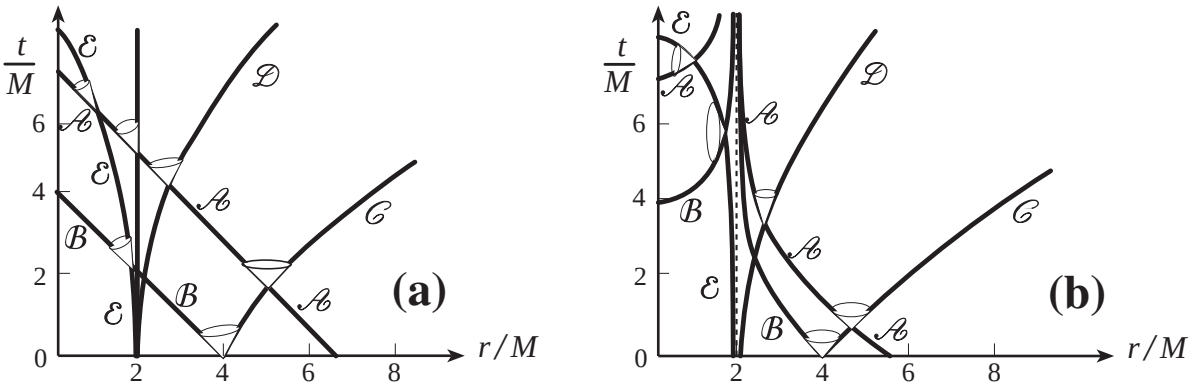


Fig. 25.5: (a) Radial light rays, and light cones, for the Schwarzschild spacetime as depicted in Eddington-Finkelstein coordinates [Eq. (25.66)]. (b) These same light rays and light cones as depicted in Schwarzschild coordinates [cf. Fig. 25.3].

lines) and $d\theta = d\phi = 0$ (radial world lines) in the line element. The result, depicted in Fig. 25.5(a), is

$$\frac{d\tilde{t}}{dr} = -1 \quad \text{for ingoing rays;} \quad \text{and } \frac{d\tilde{t}}{dr} = \left(\frac{1+2M/r}{1-2M/r}\right) \quad \text{for outgoing rays.} \tag{25.66}$$

Note that the ingoing light rays plunge unimpeded through $r = 2M$ and in to $r = 0$ along 45-degree lines in the Eddington-Finkelstein coordinate system. The outgoing light rays, by contrast, are never able to escape outward through $r = 2M$: Because of the inward tilt of the outer edge of the light cone, all light rays that begin inside $r = 2M$ are forced forever to remain inside, and in fact are drawn inexorably into $r = 0$, whereas light rays initially outside $r = 2M$ can escape to $r = \infty$.

Return, now, to the implosion of a star. The world line of the star's surface, which became asymptotically frozen at the gravitational radius when studied in Schwarzschild coordinates, plunges unimpeded through $r = 2M$ and into $r = 0$ when studied in Eddington-Finkelstein

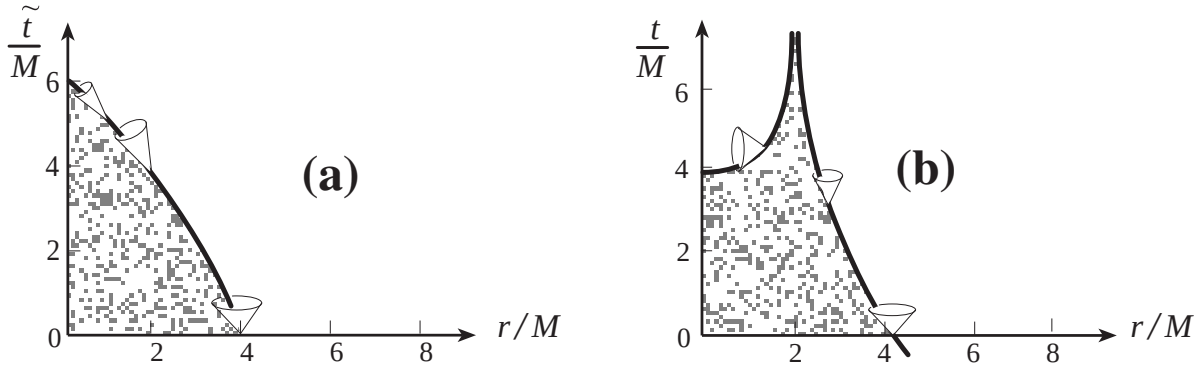


Fig. 25.6: World line of an observer on the surface of an imploding star, as depicted (a) in an Eddington-Finkelstein spacetime diagram, and (b) in a Schwarzschild spacetime diagram; see Ex. 25.7.

coordinates; see Ex. 25.7 and compare Figs. 25.6(b) and 25.6(a). Thus, in order to understand the star's ultimate fate, we must study the region $r = 0$.

As with $r = 2M$ there are two possibilities: Either the tidal forces as measured on the star's surface remain finite there, in which case something must be going wrong with the coordinate system; or else the tidal forces diverge, destroying the star. The tidal forces are computed in Ex. 25.8, with a remarkable result: They diverge. Thus, the region $r = 0$ is a *spacetime singularity*; a region where tidal gravity becomes infinitely large, destroying everything that falls into it.

This, of course, is a very unsatisfying conclusion. It is hard to believe that the correct laws of physics will predict such total destruction. In fact, they probably do not. As we shall discuss in Chap. 27, when the radius of curvature of spacetime becomes as small as $l_{\rm PW} \equiv (G\hbar/c^3)^{\frac{1}{2}} = 10^{-33}$ centimeters, space and time must cease to exist as classical entities; they, and the spacetime geometry must then become quantized; and, correspondingly, general relativity must then break down and be replaced by a quantum theory of the structure of spacetime, i.e., a quantum theory of gravity. That quantum theory will describe and govern the classically singular region at the center of a black hole. Since, however, only rough hints of the structure of that quantum theory are in hand at this time, it is not known what that theory will say about the endpoint of stellar implosion.

Unfortunately, the singularity and its quantum mechanical structure are totally invisible to observers in the external universe: The only way the singularity can possibly be seen is by means of light rays, or other signals, that emerge from its vicinity. However, because the future light cones are all directed into it (Fig. 25.6), no light-speed or sub-light-speed signals can ever emerge from it. In fact, because the outer edge of the light cone is tilted inward at every event inside the gravitational radius (Figs. 25.5 and 25.6), no signal can emerge from inside the gravitational radius to tell external observers what is going on there. In effect, the gravitational radius is an *absolute event horizon* for our universe, a horizon beyond which we cannot see—except by plunging through it, and paying the ultimate price for our momentary exploration of the hole's interior.

As most readers are aware, the region of strong, vacuum gravity left behind by the implosion of the star is called a *black hole*. The horizon, $r = 2M$, is the surface of the hole,

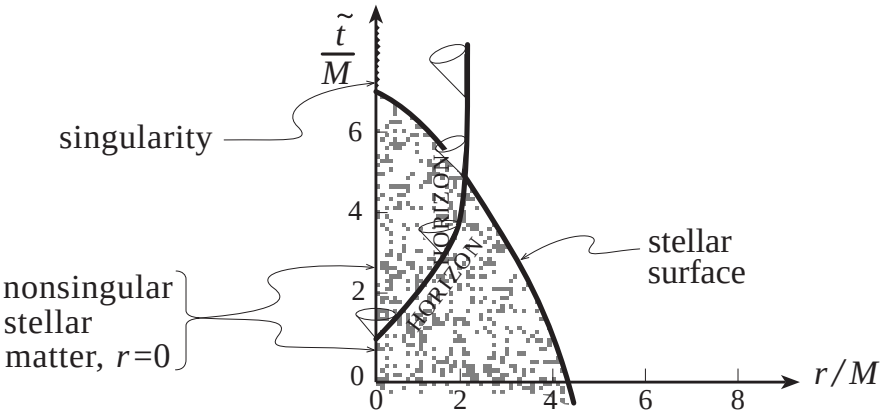


Fig. 25.7: Spacetime diagram depicting the formation and evolution of the horizon of a black hole. The coordinates outside the surface of the imploding star are those of Eddington and Finkelstein; those inside are a smooth continuation of Eddington and Finkelstein. Note that the horizon is the boundary of the region that is unable to send outgoing null geodesics to radial infinity.

and the region $r < 2M$ is its interior. The spacetime geometry of the black hole, outside and at the surface of the star which creates it by implosion, is that of Schwarzschild—though, of course, Schwarzschild had no way of knowing this in the few brief months left to him after his discovery of the Schwarzschild line element.

The horizon—defined as the boundary between spacetime regions that can and cannot communicate with the external universe—actually forms initially at the star's center, and then expands to encompass the surface at the precise moment when the surface penetrates the gravitational radius. This evolution of the horizon is depicted in an Eddington-Finkelstein-type spacetime diagram in Fig. 25.7.

Our discussion here has been confined to spherically symmetric, nonrotating black holes created by the gravitational implosion of a spherically symmetric star. Real stars, of course, are not spherical; and it was widely believed—perhaps we should say hoped—in the 1950s and 1960s that black-hole horizons and singularities would be so unstable that small nonsphericities or small rotation of the imploding star would save it from the black-hole fate. However, elegant and very general analyses carried out in the 1960s, largely by the British physicists Roger Penrose and Stephen Hawking, showed otherwise; and more recent numerical simulations on supercomputers have confirmed those analyses: Singularities are a generic outcome of stellar implosion, as are the horizons that clothe them.

****

EXERCISES

Exercise 25.7 *Example: Implosion of the Surface of a Zero-Pressure Star*
Consider the surface of a zero-pressure star, which implodes along a timelike geodesic $r = R(t)$ in the Schwarzschild spacetime of its exterior. Analyze that implosion using Schwarzschild coordinates t, r, θ, ϕ, and the exterior metric (25.1) in those coordinates.

(a) Show, using the result of Ex. 24.4(a), that the covariant time component u_t of the 4-velocity $\vec{u}$ of a particle on the star's surface is conserved along its world line. Evaluate this conserved quantity in terms of the star's mass M and the radius R_o at which it begins to implode.

(b) Use the normalization of the 4-velocity to show that the star's radius R as a function of the proper time τ since implosion began (proper time as measured on its surface) satisfies the differential equation

$$\frac{dR}{d\tau} = -[\text{const} + 2M/R]^{\frac{1}{2}} \; ; \tag{25.67}$$

and evaluate the constant. Compare this with the equation of motion for the surface as predicted by Newtonian gravity, with proper time τ replaced by Newtonian time. (It is a coincidence that the two equations are identical.)

(c) Show from the equation of motion (25.67) that the star implodes through the horizon $R = 2M$ in a finite proper time of order (25.57). Show that this proper time has the magnitudes cited in Eq. (25.57) and the sentences following it.

(d) Show, further, that when studied in Eddington-Finkelstein coordinates, the surface's implosion to $r = 2M$ requires only finite coordinate time $\tilde{t}$; in fact, a time of the same order of magnitude as the proper time (25.57). [Hint: from the Eddington-Finkelstein line element (25.65) and Eq. (25.57) derive a differential equation for $d\tilde{t}/d\tau$ along the world line of the star's surface, and use it to examine the behavior of $d\tilde{t}/d\tau$ near $R = 2M$.] Show, further, that expression (25.67) remains valid all the way through the gravitational radius and in to $r = 0$. From this conclude that the proper time and Eddington-Finkelstein time to reach $r = 0$ are also of order (25.57).

(e) Show that the world line of the star's surface as depicted in an Eddington-Finkelstein spacetime diagram has the form shown in Fig. 25.6(a), and that in a Schwarzschild spacetime diagram it has the form shown in 25.6(b).

Exercise 25.8 *Example: Gore at the Singularity*

(a) Knowing the world line of the surface of an imploding star in Eddington-Finkelstein coordinates, draw that world line in a Schwarzschild spacetime diagram. Note that as the world line approaches $r = 0$, it asymptotes to the curve $\{(t, \theta, \phi) = \text{const}, \; r \text{ variable}\}$. Explain why this is required by the light-cone structure near $r = 0$.

(b) Show that the curve to which it asymptotes, $\{(t, \theta, \phi) = \text{const}, \; r \text{ variable}\}$ is a timelike geodesic for $r < 2M$. [*Hint:* use the result of Ex. 24.4(a).]

(c) Show that the basis vectors of the infalling observer's local Lorentz frame near $r = 0$ are related to the Schwarzschild coordinate basis by

$$\vec{e}_{\hat{0}} = -\left(\frac{2M}{r} - 1\right)^{\frac{1}{2}} \frac{\partial}{\partial r} \; , \quad \vec{e}_{\hat{1}} = \left(\frac{2M}{r} - 1\right)^{-\frac{1}{2}} \frac{\partial}{\partial t} \; , \quad \vec{e}_{\hat{2}} = \frac{1}{r}\frac{\partial}{\partial \theta} \; , \quad \vec{e}_{\hat{3}} = \frac{1}{r \sin\theta}\frac{\partial}{\partial \phi} \; . \tag{25.68}$$

What are the components of the Riemann tensor in that local Lorentz frame?

(d) Show that the tidal forces produced by the Riemann tensor stretch an infalling observer in the radial, $\vec{e}_{\hat{1}}$, direction and squeeze the observer in the tangential, $\vec{e}_{\hat{2}}$ and $\vec{e}_{\hat{3}}$, directions; and show that the stretching and squeezing forces become infinitely strong as the observer approaches $r = 0$.

(e) Idealize the body of an infalling observer to consist of a head of mass $\mu \simeq 20$kg and feet of mass $\mu \simeq 20$kg separated by a distance $h \simeq 2$ meters, as measured in the observer's local Lorentz frame, and with the separation direction radial. Compute the stretching force between head and feet, as a function of proper time τ, as the observer falls into the singularity. Assume that the hole has the mass $M = 5 \times 10^9 M_\odot$ which is suggested by astronomical observations for a possible black hole at the center of the nearest giant elliptical galaxy to our own, the galaxy M87 (Sargent *et al.* 1978). How long before hitting the singularity (at what proper time τ) does the observer die, if he or she is a human being made of flesh, bone, and blood?

Exercise 25.9 *Example: Wormholes*
Our study of the Schwarzschild solution of Einstein's equations in this chapter has been confined to situations where, at small radii, the Schwarzschild geometry joins onto that of a star—either a static star, or a star that implodes to form a black hole. Suppose, by contrast, that there is no matter anywhere in the Schwarzschild spacetime. To get insight into this situation, construct an embedding diagram for the equatorial 2-surfaces $t = \text{const}$, $\theta = \pi/2$ of the vacuum Schwarzschild spacetime, using as the starting point the line element of such a 2-surface written in isotropic coordinates [Ex. 25.3]:

$$^{(2)}ds^2 = \left(1 + \frac{M}{2\bar{r}}\right)^4 (d\bar{r}^2 + \bar{r}^2 d\phi^2) \; . \tag{25.69}$$

Show that the region $0 < \bar{r} \ll M/2$ is an asymptotically flat space, that the region $\bar{r} \gg M/2$ is another asymptotically flat space, and that these two spaces are connected by a *wormhole* ("bridge," "tunnel") through the embedding space. This exercise, first carried out by Ludwig Flamm (1916) in Vienna just a few months after the discovery of the Schwarzschild solution, reveals that the pure vacuum Schwarzschild spacetime represents a wormhole that connects two different universes—or, with a change of topology, a wormhole that connects two widely separated regions of one universe. For further discussion see, e.g., Chap. 31 of MTW. For a discourse on why such wormholes almost certainly do not occur naturally in the real universe, and for analyses of whether the laws of physics as we know them allow advanced civilizations to construct wormholes, maintain them as interstellar travel devices, and convert them into "time machines," see Morris and Thorne (1988); Morris, Thorne, and Yurtsever (1988); Friedman *et al.* (1990); Kim and Thorne (1991).

****

25.5 Spinning Black Holes: The Kerr Spacetime

25.5.1 The Kerr Metric for a Spinning Black Hole

Consider a star that collapses to form a black hole, and assume for pedagogical simplicity that during the collapse no energy, momentum, or angular momentum flows through a large sphere surrounding the system. Then the asymptotic conservation laws discussed in Sec. 24.9.4 guarantee that the mass M, linear momentum P_j, and angular momentum S_j of the newborn hole, as encoded in its asymptotic metric, will be identical to those of its parent star. If (as we shall assume) our asymptotic coordinates are those of the star's rest frame so $P_j = 0$, then the hole will also be at rest in those coordinates, i.e. it will also have $P_j = 0$.

If the star was non-spinning so $S_j = 0$, then the hole will also have $S_j = 0$, and a powerful theorem due to Werner Israel guarantees that—after it has settled down into a quiescent state—the hole's spacetime geometry will be that of Schwarzschild.

If, instead, the star was spinning so $S_j \neq 0$, then the final, quiescent hole cannot be that of Schwarzschild. Instead, according to a powerful theorem due to Hawking, Carter, Robinson, and others, its spacetime geometry will be the following exact, vacuum solution to the Einstein field equation (which is called the *Kerr solution* because it was discovered by the New-Zealand mathematician Roy Kerr[3]):

$$\boxed{ds^2 = -\alpha^2 dt^2 + \frac{\rho^2}{\Delta} dr^2 + \rho^2 d\theta^2 + \varpi^2 (d\phi - \omega dt)^2 \ .} \tag{25.70a}$$

Here

$$\boxed{\Delta = r^2 + a^2 - 2Mr \ , \quad \rho^2 = r^2 + a^2 \cos^2\theta \ , \quad \Sigma^2 = (r^2 + a^2)^2 - a^2 \Delta \sin^2\theta \ ,}$$

$$\boxed{\alpha^2 = \frac{\rho^2}{\Sigma^2}\Delta \ , \quad \varpi^2 = \frac{\Sigma^2}{\rho^2}\sin^2\theta \ , \quad \omega = \frac{2aMr}{\Sigma^2} \ .} \tag{25.70b}$$

In this line element $\{t, r, \theta, \phi\}$ are the coordinates, and there are two constants, M and a. The physical meanings of M and a can be deduced from the asymptotic form of this Kerr metric at large radii:

$$ds^2 = -\left(1 - \frac{2M}{r}\right) dt^2 - \frac{4Ma}{r} \sin^2\theta d\phi dt + \left[1 + \mathcal{O}\left(\frac{M}{r}\right)\right] [dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2)] \ . \tag{25.71}$$

By comparing with the standard asymptotic metric in spherical coordinates, Eq. (24.113), we see that M is the mass of the black hole, $Ma \equiv J_H$ is the magnitude of its spin angular momentum, and its spin points along the polar axis, $\theta = 0$. Evidently, then, the constant a is the hole's angular momentum per unit mass; it has the same dimensions as M: length (in geometrized units).

It is easy to verify that, in the limit $a \to 0$, the Kerr metric (25.70) reduces to the Schwarzschild metric (25.1), and the coordinates $\{t, r, \theta, \phi\}$ in which we have written it (called "Boyer-Lindquist coordinates") reduce to Schwarzschild's coordinates.

[3]Kerr, R. P. 1963 Phys Rev. Lett. 11, 237.

Just as it is convenient to read the covariant metric components $g_{\alpha\beta}$ off the line element (25.70a) via $ds^2 = g_{\alpha\beta}dx^\alpha dx^\beta$, so also it is convenient to read the contravariant metric components $g^{\alpha\beta}$ off an expression for the wave operator $\Box \equiv \vec{\nabla}\cdot\vec{\nabla} = g^{\alpha\beta}\nabla_\alpha\nabla_\beta$. (Here $\nabla_\alpha \equiv \nabla_{\vec{e}_\alpha}$ is the directional derivative along the basis vector $\vec{e}_\alpha$.) For the Kerr metric (25.70a), a straightforward inversion of the matrix $||g_{\alpha\beta}||$ gives the $||g^{\alpha\beta}||$ embodied in the following equation:

$$\boxed{\Box = \frac{-1}{\alpha^2}(\nabla_t + \omega\nabla_\phi)^2 + \frac{\Delta}{\rho^2}{\nabla_r}^2 + \frac{1}{\rho^2}{\nabla_\theta}^2 + \frac{1}{\varpi^2}{\nabla_\phi}^2 \ .} \tag{25.72}$$

25.5.2 Dragging of Inertial Frames

As we shall see in Chap. 26, the spin of a black hole (or any other system in asymptotically flat spacetime) can be measured by its influence on the orientation of gyroscopes in the asymptotic region: The spin drags inertial frames into rotational motion around the black hole, thereby causing gyroscopes to precess.

This frame dragging also shows up in the geodesic trajectories of freely falling particles. Consider, for concreteness, a particle dropped from rest far outside the black hole. Its initial 4-velocity will be $\vec{u} = \partial/\partial t$, and correspondingly, in the distant, flat region of spacetime, the covariant components of $\vec{u}$ will be $u_t = -1$, $u_r = u_\theta = u_\phi = 0$.

Now, the Kerr metric coefficients $g_{\alpha\beta}$, like those of Schwarzschild, are independent of t and ϕ; i.e., the Kerr metric is symmetric under time translation (it is "stationary") and under rotation about the hole's spin axis (it is "axially symmetric"). These symmetries impose corresponding conservation laws on the infalling particle [Ex. 24.4(a)]: u_0 and u_ϕ are conserved;, i.e. they retain their initial values $u_0 = -1$ and $u_\phi = 0$ as the particle falls. By raising indices, $u^\alpha = g^{\alpha\beta}u_\beta$, using the metric coefficients embodied in Eq. (25.72), we learn the evolution of the contravariant 4-velocity components, $u^t = -g^{tt} = 1/\alpha^2$, $u^\phi = -g^{t\phi} = \omega/\alpha^2$. These in turn imply that as the particle falls, it acquires an *angular velocity* around the hole's spin axis given by

$$\boxed{\Omega \equiv \frac{d\phi}{dt} = \frac{d\phi/d\tau}{dt/d\tau} = \frac{u^\phi}{u^t} = \omega \ .} \tag{25.73}$$

(The coordinates ϕ and t are tied to the rotational and time-translation symmetries of the spacetime, so they are very special; that is why we can use them to define a physically meaningful angular velocity.)

At large radii, $\omega = 4aM/r \to 0$ as $r \to \infty$. Therefore, when first dropped, the particle falls radially inward. However, as the particle nears the hole and picks up speed, it acquires a significant angular velocity around the hole's spin axis. The physical cause of this is *frame dragging:* The hole's spin drags inertial frames into rotation around the spin axis, and that inertial rotation drags the inertially falling particle into a circulatory orbital motion.

25.5.3 The Light-Cone Structure, and the Horizon

Just as for a Schwarzschild hole, so also for Kerr, the light-cone structure is a powerful tool for identifying the horizon and exploring the spacetime geometry near it.

At any event in spacetime, the tangents to the light cone are those displacements $\{dt, dr, d\theta, d\phi\}$ along which $ds^2 = 0$. The outermost and innermost edges of the cone are those for which $(dr/dt)^2$ is maximal. By setting expression (25.70a) to zero we see that dr^2 has its maximum value, for a given dt^2, when $d\phi = \omega dt$ and $d\theta = 0$. In other words, *the photons that move radially outward or inward at the fastest possible rate are those whose angular motion is that of frame dragging*, Eq. (25.73). For these extremal photons, the radial motion (along the outer and inner edges of the light cone) is

$$\frac{dr}{dt} = \pm\frac{\alpha\sqrt{\Delta}}{\rho} = \pm\frac{\Delta}{\Sigma} \, . \tag{25.74}$$

Now, Σ is positive definite, but Δ is not; it decreases monotonically, with decreasing radius, reaching zero at

$$\boxed{r = r_H \equiv M + \sqrt{M^2 - a^2}} \tag{25.75}$$

[Eq. (25.70b)]. (We shall assume that $|a| < M$ so r_H is real, and shall justify this assumption below.) Correspondingly, the light cone closes up to a sliver then pinches off as $r \to r_H$; and it pinches onto a null curve (actually, a null geodesic) given by

$$r = r_H \, , \quad \theta = \text{constant} \, , \quad \phi = \Omega_H t + \text{constant} \, , \tag{25.76}$$

where

$$\boxed{\Omega_H = \omega(r = r_H) = \frac{a}{2Mr_H} \, .} \tag{25.77}$$

This light-cone structure is depicted in Fig. 25.8(a,b). The light-cone pinch off as shown there is the same as that for Schwarzschild spacetime (Fig. 25.2) except for the frame-dragging-induced angular tilt $d\phi/dt = \omega$ of the light cones. In the Schwarzschild case, as $r \to 2M$, the light cones pinch onto the geodesic world lines $\{r = 2M, \theta = \text{const}, \phi = \text{const}\}$ of photons that travel along the horizon. These null world lines are called the horizon's *generators*. In the Kerr case the light-cone pinchoff reveals that *the horizon is at* $r = r_H$, and *the horizon generators are null geodesics that travel around and around the horizon with angular velocity* Ω_H. This motivates us to regard the horizon itself as having the rotational angular velocity Ω_H.

Whenever a finite-rest-mass particle falls into a spinning black hole, its world line, as it nears the horizon, is constrained always to lie inside the light cone. The light-cone pinch off then constrains its motion to approach, asymptotically, the horizon generators. Therefore, as seen in Boyer-Lindquist coordinates, the particle is dragged into an orbital motion, just above the horizon, with asymptotic angular velocity $d\phi/dt = \Omega_H$, and it travels around and around the horizon "forever" (for infinite Boyer-Lindquist coordinate time t), and never (as $t \to \infty$) manages to cross through the horizon.

As in the Schwarzschild case, so also in Kerr, this infall to $r = r_H$ requires only finite proper time τ as measured by the particle, and the particle feels only finite tidal forces (only finite values of the components of Riemann in its proper reference frame). Therefore, as for Schwarzschild spacetime, the "barrier" to infall through $r = r_H$ must be an illusion produced by a pathology of the Boyer-Lindquist coordinates at $r = r_H$.

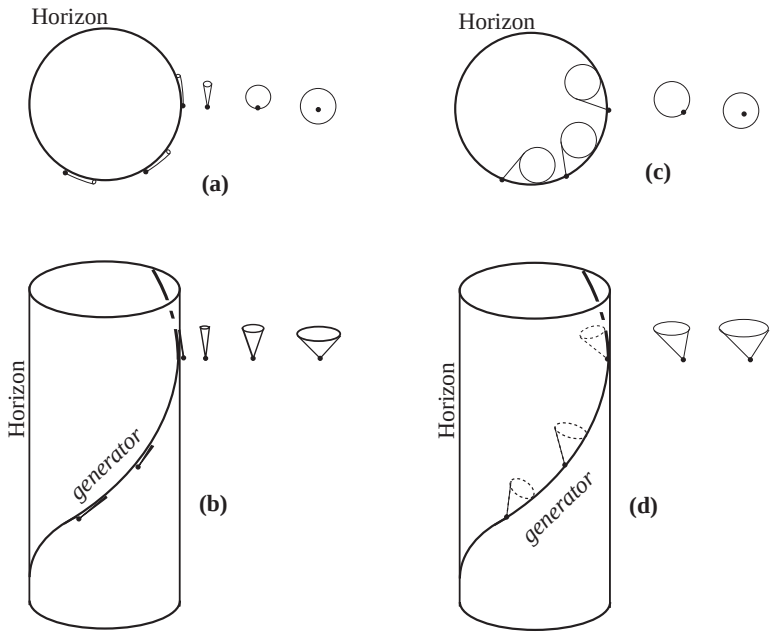


Fig. 25.8: (a) and (b): Light-cone structure of Kerr spacetime depicted in Boyer-Lindquist coordinates. Drawing (b) is a spacetime diagram; drawing (a) is the same diagram as viewed from above. (c) and (d): The same light-cone structure in Kerr coordinates.

This coordinate pathology can be removed by a variety of different coordinate transformations. One is the following change of the time and angular coordinates:

$$\boxed{\tilde{t} = t + \int \frac{2Mr}{\Delta} dr \ , \quad \tilde{\phi} = \phi + \int \frac{a}{\Delta} dr \ .} \tag{25.78}$$

The new (tilded) coordinates are a variant of a coordinate system originally introduced by Kerr, so we shall call them "Kerr coordinates". By inserting the coordinate transformation (25.78) into the line element (25.70a), we obtain the following form of the Kerr metric, in Kerr coordinates:

$$\boxed{ds^2 = -\alpha^2 d\tilde{t}^2 + \frac{4Mr\rho^2}{\Sigma^2} dr d\tilde{t} + \frac{\rho^2(\rho^2 + 2Mr)}{\Sigma^2} dr^2 + \rho^2 d\theta^2 + \varpi^2 \left[d\tilde{\phi} - \omega d\tilde{t} - \frac{a(\rho^2 + 2Mr)}{\Sigma^2} dr \right]^2} \tag{25.79}$$

It is easy to verify that when $a \to 0$ (so Kerr spacetime becomes Schwarzschild), the Kerr coordinates (25.78) become those of Eddington and Finkelstein [Eq. (25.64)], and the Kerr line element (25.79) becomes the Eddington-Finkelstein one [Eq. (25.65)]. Similarly, when one explores the light-cone structure for a spinning black hole in the Kerr coordinates [Fig. 25.8(c,d)], one finds a structure like that of Eddington-Finkelstein [Fig. 25.5(a)]: At large radii, $r \gg M$, the light cones have their usual 45-degree form, but as one moves inward toward the horizon, they begin to tilt inward. In addition to the inward tilt, there is a frame-dragging-induced tilt in the direction of the hole's rotation, $+\phi$. At the horizon the outermost edge of the light cone is tangent to the horizon generators; and in Kerr

coordinates, as in Boyer-Lindquist, these generators rotate around the horizon with angular velocity $d\tilde{\phi}/d\tilde{t} = \Omega_H$ [cf. Eq. (25.78), which says that at fixed r, $\tilde{t} = t + \text{constant}$ and $\tilde{\phi} = \phi + \text{constant}$].

This light-cone structure demonstrates graphically that the horizon is at the radius $r = r_H$. Outside there, the outer edge of the light cone tilts toward increasing r and so it is possible to escape to infinity. Inside there the outer edge of the light cone tilts inward, and all forms of matter and energy are forced to move inward, toward a singularity whose structure, presumably, is governed by the laws of quantum gravity.[4]

25.5.4 Evolution of Black Holes: Rotational Energy and Its Extraction

When a spinning star collapses to form a black hole, its centrifugal forces will flatten it, and the dynamical growth of flattening will produce gravitational radiation (Chap. 26). The newborn hole will also be flattened and will not have the Kerr shape; but rather quickly, within a time $\Delta t \sim$ 10's or 100's of $M \sim 10\mu\text{s}(M/M_\odot)$, the deformed hole will shake off its deformations as gravitational waves and settle down into the Kerr shape. This is the conclusion of extensive analyses, both analytic and numerical.

Many black holes are in binary orbits with stellar companions, and pull gas off their companions and swallow it. Other black holes accrete gas from interstellar space. Any such accretion causes the hole's mass and spin to evolve in accord with the conservation laws (24.114) and (24.115). One might have thought that by accreting a large amount of angular momentum, a hole's angular momentum per unit mass a could grow larger than its mass M. If this were to happen, then $r_H = M + \sqrt{M^2 - a^2}$ would cease to be a real radius—a fact that signals the destruction of the hole's horizon: As a grows to exceed M, the inward light-cone tilt gets reduced so that everywhere the outer edge of the cone points toward increasing r, which means that light, particles, and information are no longer trapped.

Remarkably, however, it appears that the laws of general relativity forbid a ever to grow larger than M. As accretion pushes a/M upward toward unity, the increasing drag of inertial frames causes a big increase of the hole's cross section to capture material with negative angular momentum (which will spin the hole down) and a growing resistance to capturing any further material with large positive angular momentum. Infalling particles that might try to push a/M over the limit get flung back out by huge centrifugal forces, before they can reach the horizon. A black hole, it appears, is totally resistant against having its horizon destroyed.

In 1969, Roger Penrose discovered that a large fraction of the mass of a spinning black hole is in the form of rotational energy, stored in the whirling spacetime curvature outside the hole's horizon; and this rotational energy can be extracted. Penrose discovered this by the following thought experiment:

[4]Much hooplah has been made of the fact that in the Kerr spacetime it is possible to travel inward, through a "Cauchy horizon" and then into another universe. However, the Cauchy horizon, located at $r = M - \sqrt{M^2 - a^2}$ is highly unstable against perturbations, which convert it into a singularity with infinte spacetime curvature. For details of this instability and the singularity, see, e.g., Brady, Droz and Morsink (1998) and references therein.

From far outside the hole, you throw a massive particle into the vicinity of the hole's horizon. Assuming you are at rest with respect to the hole, your 4-velocity is $\vec{U} = \partial/\partial t$. Denote by $E^{\rm in} = -\vec{p}^{\rm in} \cdot \vec{U} = -\vec{p}^{\rm in} \cdot (\partial/\partial t) = -p_t^{\rm in}$ the energy of the particle (rest mass plus kinetic), as measured by you; cf. Eq. (1.38). As the particle falls, $p_t^{\rm in}$ is conserved because of the Kerr metric's time-translation symmetry. Arrange that, as the particle nears the horizon, it splits into two particles, one (labeled "plunge") plunges through the horizon and the other (labeled "out") flies back out to large radii, where you catch it. Denote by $E^{\rm plunge} \equiv -p_t^{\rm plunge}$ the conserved energy of the plunging particle and by $E^{\rm out} \equiv -p_t^{\rm out}$ that of the out-flying particle. Four-momentum conservation at the event of the split dictates that $\vec{p}^{\rm in} = \vec{p}^{\rm plunge} + \vec{p}^{\rm out}$, which implies this same conservation law for all the components of the 4-momenta, in particular

$$E^{\rm out} = E^{\rm in} - E^{\rm plunge} \ . \tag{25.80}$$

Now, it is a remarkable fact that the Boyer-Lindquist time basis vector $\partial/\partial t$ has a squared length $\partial/\partial t \cdot \partial/\partial t = g_{tt} = -\alpha^2 + \varpi^2\omega^2$ that becomes positive at radii

$$\boxed{r < r_{\rm ergo} = M + \sqrt{M^2 - a^2\cos^2\theta}\ ,} \tag{25.81}$$

which is larger than r_H everywhere except on the hole's spin axis, $\theta = 0, \pi$. The region $r_H < r < r_{\rm ergo}$ is called the hole's *ergosphere*. If the split into two particles occurs in the ergosphere, then it is possible to arrange the split such that the scalar product of the *timelike* vector $\vec{p}^{\rm plunge}$ with the *spacelike* vector $\partial/\partial t$ is *positive*, which means that the plunging particle's conserved energy $E^{\rm plunge} = -\vec{p}^{\rm plunge} \cdot (\partial/\partial t)$ is *negative*; whence [by Eq. (25.80)]

$$E^{\rm out} > E^{\rm in} \ . \tag{25.82}$$

see Ex. 25.11(a).

When the outflying particle reaches your location, $r \gg M$, its conserved energy is equal to its physically measured total energy (rest-mass plus kinetic); and the fact that $E^{\rm out} > E^{\rm in}$ means that you get back more energy (rest-mass plus kinetic) than you put in. The hole's asymptotic energy-conservation law (24.114) implies that the hole's mass has decreased by precisely the amount of energy that you have extracted,

$$\Delta M = -(E^{\rm out} - E^{\rm in}) = E^{\rm plunge} < 0 \ . \tag{25.83}$$

A closer scrutiny of this process [Ex. 25.11(f)] reveals that the plunging particle must have had negative angular momentum, so it has spun the hole down a bit. The energy you extracted, in fact, came from the hole's enormous store of rotational energy, which makes up part of its mass M; and your extraction of energy has reduced that rotational energy.

Stephen Hawking has used sophisticated mathematical techniques to prove that, independently of how you carry out this thought experiment, and, indeed, independently of what is done to a black hole, general relativity requires that the horizon's surface area A_H never decrease. This is called the second law of black-hole mechanics, and it actually turns out to be a variant of the second law of thermodynamics, in disguise. A straightforward calculation (Ex. 25.10) reveals that the horizon surface area is given by

$$\boxed{A_H = 4\pi(r_H^2 + a^2) = 8\pi M r_H \quad \text{for a spinning hole,}} \tag{25.84a}$$

$$\boxed{A_H = 16\pi M^2 \quad \text{for a nonspinning hole, } a = 0.} \tag{25.84b}$$

Dimitrious Christodoulou has shown (cf. Ex. 25.11) that, in the Penrose process, the nondecrease of A_H is the only constraint on how much energy one can extract, so by a sequence of optimally designed particle injections and splits that keep A_H unchanged, one can reduce the mass of the hole to

$$\boxed{M_{\rm irr} = \sqrt{\frac{A_H}{16\pi}} = \sqrt{\frac{M(M + \sqrt{M^2 - a^2}\,)}{2}}\,,} \tag{25.85}$$

but no smaller. This is called the hole's irreducible mass. The hole's total mass is the sum of its irreducible mass and its rotational energy $M_{\rm rot}$; so the rotational energy is

$$\boxed{M_{\rm rot} = M - M_{\rm irr} = M\left[1 - \sqrt{\frac{1}{2}\left(1 + \sqrt{1 - a^2/M^2}\right)}\,\right]\,.} \tag{25.86}$$

For the fastest possible spin, $a = M$, this gives $M_{\rm rot} = M(1 - 1/\sqrt{2}) \simeq 0.2929M$. This is the maximum amount of energy that can be extracted, and it is enormous compared to the energy $\sim 0.005M$ that can be released by thermonuclear burning in a star with mass M.

The Penrose process of throwing in particles and splitting them in two is highly idealized, and of little or no importance in Nature. However, Nature seems to have found a very effective method for extracting rotational energy from spinning black holes: the *Blandford-Znajek process* in which magnetic fields, threading through a black hole and held on the hole by a surrounding disk of hot plasma, extract energy electromagnetically. This process is thought to power the gigantic jets that shoot out of the nuclei of some galaxies, and might also be the engine for some powerful gamma-ray bursts.

****

EXERCISES

Exercise 25.10 *Derivation: Surface Area of a Spinning Black Hole*
From the Kerr metric (25.71) derive Eq. (25.84) for the surface area of a spinning black hole's horizon—i.e., the surface area of the two-dimensional surface $\{r = r_H,\ t = \text{constant}\}$.

Exercise 25.11 *Example: Penrose Process, Hawking Radiation, and Thermodynamics of Black Holes*

(a) Consider the Penrose process, described in the text, in which a particle flying inward toward a spinning hole's horizon splits in two inside the ergosphere, and one piece

plunges into the hole while the other flies back out. Show that it is always possible to arrange this process so the plunging particle has negative energy, $E^{\rm plunge} = -\vec{p}^{\rm plunge} \cdot \partial/\partial t < 0$. [Hint: Perform a calculation in a local Lorentz frame in which $\partial/\partial t$ points along a spatial basis vector, $\vec{e}_{\hat{1}}$. Why is it possible to find such a local Lorentz frame?]

(b) Around a spinning black hole consider the vector field

$$\vec{\xi}_H \equiv \partial/\partial t + \Omega_H \partial/\partial\phi \ , \tag{25.87}$$

where Ω_H is the Horizon's angular velocity. Show that in the horizon (at radius $r = r_H$) this vector field is null and is tangent to the horizon generators. Show that all other vectors in the horizon are spacelike.

(c) In the Penrose process, the plunging particle changes the hole's mass by an amount ΔM and its spin angular momentum by an amount ΔJ_H. Show that

$$\Delta M - \Omega_H \Delta J_H = -\vec{p}^{\rm plunge} \cdot \vec{\xi}_H \ . \tag{25.88}$$

Here $\vec{p}^{\rm plunge}$ and $\vec{\xi}_H$ are to be evaluated at the event where the particle plunges through the horizon, so they both reside in the same tangent space. [Hint: the angular momentum carried into the horizon is the quantity $p_\phi^{\rm plunge} = \vec{p}^{\rm plunge} \cdot \partial/\partial\phi$. Why? This quantity is conserved along the plunging particle's world line. Why?]

(d) Show that if $\vec{A}$ is any future directed timelike vector and $\vec{K}$ is any null vector, both living in the tangent space at the same event in spacetime, then $\vec{A} \cdot \vec{K} < 0$. [Hint: Perform a calculation in a specially chosen local Lorentz frame.] Thereby conclude that $-\vec{p}^{\rm plunge} \cdot \vec{\xi}_H$ is positive, whatever may be the world line and rest mass of the plunging particle.

(e) Show that, in order for the plunging particle to decrease the hole's mass, it must also decrease the hole's angular momentum; i.e., it must spin the hole down a bit.

(f) The second law of black-hole mechanics says that, whatever may be the particle's world line and rest mass, when the particle plunges through the horizon it causes the horizon's surface area A_H to increase. This suggests that the always positive quantity $\Delta M - \Omega_H \Delta J_H = -\vec{p}^{\rm plunge} \cdot \vec{\xi}_H$ might be a multiple of the increase ΔA_H of the horizon area. Show that this is, indeed, the case:

$$\boxed{\Delta M = \Omega_H \Delta J_H + \frac{g_H}{8\pi} \Delta A_H \ ,} \tag{25.89}$$

where g_H is given in terms of the hole's mass M and the radius r_H of its horizon by

$$\boxed{g_H = \frac{r_H - M}{2 M r_H} \ .} \tag{25.90}$$

[You might want to do the algebra, based on Kerr-metric formulae, on a computer.] The quantity g_H is called the hole's "surface gravity" for a variety of reasons, including

the fact that an observer who hovers just above a horizon generator, blasting his or her rocket engines to avoid falling into the hole, has a 4-acceleration with magnitude g_H/α and thus feels a "gravitational acceleration" of this magnitude; here $\alpha = g^{tt}$ is a component of the Kerr metric, Eqs. (25.70a) and (25.72). This gravitational acceleration is arbitrarily large for an observer arbitrarily close to the horizon (where Δ and hence α is arbitrarily close to zero); when renormalized by α to make it finite, the accelation is g_H.

(g) Stephen Hawking has shown, using quantum field theory, that a black hole's horizon emits thermal (black-body) radiation. The temperature of this "Hawking radiation", as measured by the observer who hovers just above the horizon, is proportional to the gravitational acceleration g_H/α that the observer measures, with a proportionality constant $\hbar/2\pi k_B$, where $\hbar$ is Planck's constant and k_B is Boltzmann's constant. As this thermal radiation climbs out of the horizon's vicinity and flies off to large radii, its frequencies and temperature get redshifted by the factor α, so as measured by distant observers the temperature is

$$\boxed{T_H = \frac{\hbar}{2\pi k_B} g_H \ .} \tag{25.91}$$

This suggests a reinterpretation of Eq. (25.89) as the first law of thermodynamics for a black hole:

$$\boxed{\Delta M = \Omega_H \Delta J_H + T_H \Delta S_H \ ,} \tag{25.92}$$

where S_H is the hole's entropy; cf. Eq. (3.69). Show that this entropy is related to the horizon's surface area by

$$\boxed{S_H = k_B \frac{A_H}{4\ell_p^2} \ ,} \tag{25.93}$$

where $\ell_p = \sqrt{\hbar G/c^3} = 1.616 \times 10^{-33}$ cm is the Planck length (with G Newton's gravitation constant and c the speed of light). Because $S_H \propto A_H$, the second law of black-hole mechanics is actually the second law of thermodynamics in disguise. A black hole's entropy always increases. [Actually, the emission of the Hawking radiation will decrease the hole's entropy and surface area; but general relativity doesn't know about this because general relativity is a classical theory, and Hawking's prediction of the thermal radiation is based on quantum theory. Thus, the Hawking radiation violates the second law of black hole mechanics. It does not, however, violate the second law of thermodynamics, because the entropy carried into the surrounding universe by the Hawking radiation exceeds the magnitude of the decrease of the hole's entropy. The total entropy of hole plus universe increases.]

(h) For a ten solar mass, nonspinning black hole, what is the temperature of the Hawking radiation in degrees Kelvin, and what is the hole's entropy in units of the Boltzmann constant?

(i) Reread the discussions of black-hole thermodynamics and entropy in the expanding universe in Secs. 3.11.2 and 3.11.3, which rely on the results of this exercise.

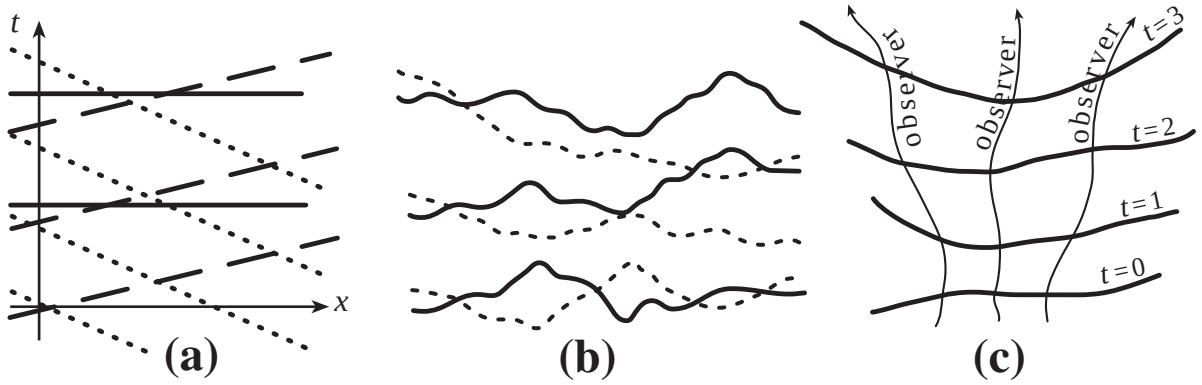


Fig. 25.9: Spacetime diagrams showing the slices of simultaneity as defined by various families of observers. Diagram (a) is in flat spacetime, and the three families (those with solid slices, those with dashed, and those with dotted) are inertial, so their slices of constant time are those of global Lorentz frames. Diagram (b) is in curved spacetime, and the two families' slices of simultaneity illustrate the "many fingered" nature of time. Diagram (c) illustrates the selection of an arbitrary foliation of spacelike hypersurfaces of simultaneity, and the subsequent construction of the world lines of observers who move orthogonal to those hypersurfaces, i.e., for whom light-ray synchronization will define those hypersurfaces as simultaneities.

****

25.6 The Many-Fingered Nature of Time

We conclude this chapter with a discussion of a concept which John Archibald Wheeler (the person who has most clarified the conceptual underpinnings of general relativity) calls the *many-fingered nature of time*.

In the flat spacetime of special relativity there are preferred families of observers: Each such family lives in a global Lorentz reference frame and uses that frame to split spacetime into space plus time. The hypersurfaces of constant time ("slices of simultaneity") which result from that split are flat hypersurfaces which slice through all of spacetime [Fig. 25.9(a)]. Of course, different preferred families live in different global Lorentz frames and thus split up spacetime into space plus time in different manners [e.g., the dotted slices of constant time in Fig. 25.9(a) as contrasted to the dashed ones]. As a result, there is no universal concept of time in special relativity; but, at least, there are some strong restrictions on time: Each family of observers will agree that another family's slices of simultaneity are flat slices.

In general relativity, i.e., in curved spacetime, even this restriction is gone: In a generic curved spacetime there are no flat hypersurfaces, and hence no candidates for flat slices of simultaneity. Hand in hand with this goes the fact that, in a generic curved spacetime there are no global Lorentz frames, and thus no preferred families of observers. A family of observers who are all initially at rest with respect to each other, and each of whom moves freely (inertially), will soon acquire relative motion because of tidal forces. As a result,

their slices of simultaneity (defined locally by Einstein light-ray synchronization, and then defined globally by patching together the little local bits of slices) may soon become rather contorted. Correspondingly, as is shown in Fig. 25.9(b), different families of observers will slice spacetime up into space plus time in manners that can be quite distorted, relative to each other—with "fingers" of one family's time slices pushing forward, ahead of the other family's here, and lagging behind there, and pushing ahead in some other place.

In curved spacetime it is best to not even restrict oneself to inertial (freely falling) observers. For example, in the spacetime of a static star, or of the exterior of a Schwarzschild black hole, the family of static observers [observers whose world lines are $\{(r, \theta, \phi) = \text{const}, t \text{ varying}\}$] are particularly simple; their world lines mold themselves to the static structure of spacetime in a simple, static manner. However, these observers are not inertial; they do not fall freely. This need not prevent us from using them to split up spacetime into space plus time, however. Their proper reference frames produce a perfectly good split; and when one uses that split, in the case of a black hole, one obtains a 3-dimensional-space version of the laws of black-hole physics which is a useful tool in astrophysical research; see Thorne, Price, and Macdonald (1986).

For any family of observers, accelerated or inertial, the slices of simultaneity as defined by Einstein light-ray synchronization (or equivalently by the space slices of the observer's proper reference frames) are the 3-surfaces orthogonal to the observers' world lines; cf. Fig. 25.9(c). To see this most easily, pick a specific event along a specific observer's world line, and study the slice of simultaneity there from the viewpoint of a local Lorentz frame in which the observer is momentarily at rest. Light-ray synchronization guarantees that, locally, the observer's slice of simultaneity will be the same as that of this local Lorentz frame; and, since the frame's slice is orthogonal to its own time direction and that time direction is the same as the direction of the observer's world line, the slice is orthogonal to the observer's world line. By the discussion in Sec. 23.5, the slice is also the same, locally (to first order in distance away from the world line), as a slice of constant time in the observer's proper reference frame.

If the observers' relative motions are sufficiently contorted (in curved spacetime or in flat), it may not be possible to mesh their local slices of simultaneity, defined in this manner, into global slices of simultaneity; i.e., there may not be any global 3-dimensional hypersurfaces orthogonal to their world lines. We can protect against this eventuality, however, by choosing the slices first: Select any foliation of spacelike slices through the curved spacetime [Fig. 25.9(c)]. Then there will be a family of timelike world lines that are everywhere orthogonal to these hypersurfaces. A family of observers who move along those world lines and who define their 3-spaces of simultaneity by local light-ray synchronization will thereby identify the orthogonal hypersurfaces as their simultaneities. Ex. 25.12 illustrates these ideas using Schwarzschild spacetime.

****

EXERCISES

Exercise 25.12 *Practice: Slices of Simultaneity in Schwarzschild Spacetime*

(a) One possible choice of slices of simultaneity for Schwarzschild spacetime is the set of 3-surfaces $t = \text{const}$, where t is the Schwarzschild time coordinate. Show that the unique family of observers for whom these are the simultaneities are the static observers, with world lines $\{(r, \theta, \phi) = \text{const},\ t \text{ varying}\}$. Explain why these slices of simultaneity and families of observers exist only outside the horizon of a black hole, and cannot be extended into the interior. Draw a picture of the world lines of these observers and their slices of simultaneity in an Eddington-Finkelstein spacetime diagram.

(b) A second possible choice of simultaneities is the set of 3-surfaces $\tilde{t} = \text{const}$, where $\tilde{t}$ is the Eddington-Finkelstein time coordinate. What are the world lines of the observers for whom these are the simultaneities? Draw a picture of those world lines in an Eddington-Finkelstein spacetime diagram. Note that they and their simultaneities cover the interior of the hole as well as its exterior.

****

Bibliographic Note

In our opinion, the best elementary textbook treatment of black holes and relativistic stars is that in Chaps. 12, 13, 15, and 24 of Hartle (2003); this treatment is also remarkably complete. For nonrotating, relativistic stars at an elementary level we also recommend Chap. 10 of Schutz (1980), and at a more advanced level (including stellar pulsations), Chaps. 23, 24, and 26 of MTW. For black holes at an intermediate level see Chaps. 5 and 6 of Carroll (2004), and at a more advanced level, Chap. 12 of Wald (1984) which is brief and highly mathematical, and Chaps. 31–34 of MTW which is long and less mathematical.

The above are all portions of general relativity textbooks. There are a number of books and monographs devoted solely to the theory of black holes and/or relativistic stars. Among these, we particularly recommend the following: Shapiro and Teukolsky (1983) is an astrophysically oriented book at much the same level as this chapter, but with much greater detail and extensive applications; it deals with black holes, neutron stars and white dwarf stars in astrophysical settings. Frolov and Novikov (1998) is a very thorough monograph on black holes, including their fundamental theory, and their interactions with the rest of the universe; it includes extensive references to the original literature and readable summaries of all the important issues that had been treated by black-hole researchers as of 1997. Chandrasekhar (1983) is an idiosynchratic but elegant and complete monograph on the theory of black holes and especially small perturbations of them.

Bibliography

Bertotti, Bruno, 1959. "Uniform Electromagnetic Field in the Theory of General Relativity," *Physical Review*, **116**, 1331–1333.

Box 25.3
Important Concepts in Chapter 25

- Schwarzschild spacetime geometry
  - Metric in Schwarzschild coordinates, Eq. (25.1); in isotropic coordinates, Ex. 25.3; in Eddington-Finkelstein coordinates, Eqs. (25.64, (25.65)
  - Connection coefficients and curvature tensors in Schwarzschild coordinates and in their orthonormal basis, Box 25.2
- Deducing the properties of a spacetime and the nature of the coordinates from a metric, Sec. sec:25Schwartzschild and Ex. 25.2
- Birkhoff's theorem, Sec. 25.3.1
- Relativistic stars, Sec. 25.3
  - Radius R always larger than gravitational radius $2M$, Eq. (25.15), Sec. 25.3.5
  - Metric (25.16) and stress-energy tensor (25.26)
  - Proper reference frame of fluid, Sec. 25.3.2
  - Deducing equations of structure from local energy-momentum conservation and the Einstein field equations, Secs. 25.3.3 and 25.3.4
  - Momentum conservation implies $(\rho + P)\mathbf{a} = -\boldsymbol{\nabla} P$, Sec. 25.3.3
  - Embedding Diagram, Sec. 25.3.5, Fig. 25.1, Ex. 25.4
  - Gravitational Redshift, Ex. 25.5
- Implosion of a star to form a Schwarzschild black hole, Sec. 25.4
  - To reach $R = 2M$ (horizon radius): infinite Schwarzschild coordinate time t but finite proper time τ and Eddington-Finkelstein time $\tilde{t}$, Sec. 25.4, Ex. 25.7
  - Finite tidal forces at horizon radius, Eq. (25.63)
  - Black-hole horizon: its formation and evolution, Fig. 25.7
  - Infinite tidal forces at singularity, Ex. 25.8
- Wormholes, Ex. 25.9
- Spinning black holes: the Kerr Spacetime, Sec. 25.5
  - Kerr metric: in Boyer-Lindquist coordinates, Eqs. (25.70); in Kerr coordinates, Eq. (25.79)
  - Dragging of inertial frames, Sec. 25.5.2
  - Horizon generators, Sec. 25.5.3, Fig. 25.8
  - Horizon radius r_H, Eq. (25.75); horizon angular velocity Ω_H, Eq. (25.77); horizon surface area A_H, Eqs. (25.84), horizon surface gravity g_H, Eq. (25.90)
  - Second law of black-hole mechanics and thermodynamics, Ex. 25.11
  - Hawking radiation, and black-hole temperature and entropy, Ex. 25.11
  - Rotational energy, energy extraction, ergosphere, irreducible mass, Sec. 25.5.4

Birkhoff, George, 1923. *Relativity and Modern Physics*, Harvard University Press, Cambridge, Massachusetts.

Brady, Patrick R., Droz, S., and Morsink, Sharon M., 1998. "The late-time singularity inside non-spherical black holes," *Phys. Rev. D*, **D58**, 084034.

Carroll, S. M., 2004. *Spacetime and Geometry: An Introduction to General Relativity*, San Francisco: Addison Wesley.

Chandrasekhar, Subramahnian, 1983. *The Mathematical Theory of Black Holes*, Oxford University Press, Oxford.

Eddington, Arthur S., 22. *The Mathematical Theory of Relativity*, Cambridge University Press, Cambridge.

Finkelstein, David, 1958. "Past-future asymmetry of the gravitational field of a point particle," *Physical Review*, **110**, 965–967.

Flamm, Ludwig, 1916. "Beiträge zur Einsteinschen Gravitationstheorie," *Physik Z.*, **17**, 448–454.

Friedman, John, Morris, Michael S., Novikov, Igor D., Echeverria, Fernando, Klinkhammer, Gunnar, Thorne, Kip S., and Yurtsever, Ulvi, 1990. "Cauchy problem in spacetimes with closed timelike curves," *Physical Review D*, **42**, 1915–1930.

Hartle, J. B., 2003. *Gravity: An Introduction to Einstein's General Relativity*, San Francisco: Addison-Wesley.

Kim, Sung-Won and Thorne, Kip S., 1991. "Do vacuum fluctuations prevent the creation of closed timelike curves?" *Physical Review D*, **44**, 1077–1099.

MTW: Misner, Charles W., Thorne, Kip S. and Wheeler, John A., 1973. *Gravitation*, W. H. Freeman & Co., San Francisco.

Morris, Mike and Thorne, Kip S., 1988. "Wormholes in spacetime and their use for interstellar travel: a tool for teaching general relativity," *American Journal of Physics*, **56**, 395–416.

Morris, Mike, Thorne, Kip S., and Yurtsever, Ulvi, 1988. "Wormholes, time machines, and the weak energy condition," *Physical Review Letters*, **61**, 1446–1449.

Frolov, Valery Pavlovich and Novikov, Igor Dmitrievich, 1998. *The Physics of Black Holes*, second edition, Kluwer Academic Publishers, Berlin.

Oppenheimer, J. Robert, and Snyder, Hartland, 1939. "On continued gravitational contraction," *Physical Review*, **56**, 455–459.

Oppenheimer, J. Robert, and Volkoff, George, 1939. "On massive neutron cores," *Physical Review*, **55**, 374–381.

Robinson, Ivor, 1959. "A Solution of the Maxwell-Einstein Equations," *Bull. Acad. Polon. Sci.*, **7**, 351–352.

Sargent, Wallace L. W., Young, Peter J., Boksenberg, A., Shortridge, Keith, Lynds, C. R., and Hartwick, F. D. A., 1978. "Dynamical evidence for a central mass condensation in the galaxy M87," *Astrophysical Journal*, **221**, 731–744.

Schutz, B. 1980. *Geometrical Methods of Mathematical Physics*, Cambridge: Cambridge University Press.

Schwarzschild, Karl, 1916a. "Über das Gravitationsfeld eines Massenpunktes nach der Einsteinschen Theorie," *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, **1916 vol. I**, 189–196.

Schwarzschild, Karl, 1916b. "Über das Gravitationsfeld einer Kugel aus inkompressibler Flüssigkeit nach der Einsteinschen Theorie," *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, **1916 vol. I**, 424–434.

Shapiro, Stuart L. and Teukolsky, Saul A., 1983. *Black Holes, White Dwarfs, and Neutron Stars*, Wiley, New York.

Thorne, Kip S., Price, Richard H. and Macdonald, Douglas A., 1986. *Black Holes: The Membrane Paradigm*, Yale University Press, New Haven, Conn..

Tolman, Richard Chace, 1939. "Static solutions of Einstein's field equations for spheres of fluid," *Physical Review*, **55**, 364–373.

Wald, R. M. 1984. *General Relativity*, Chicago: University of Chicago Press.

Contents

26 Gravitational Waves and Experimental Tests of General Relativity **1**
26.1 Introduction 1
26.2 Experimental Tests of General Relativity 2
26.2.1 Equivalence Principle, Gravitational redshift, and Global Positioning System 2
26.2.2 Perihelion advance of Mercury 4
26.2.3 Gravitational deflection of light, Fermat's principle and Gravitational Lenses 4
26.2.4 Shapiro time delay 7
26.2.5 Frame dragging and Gravity Probe B 8
26.2.6 Binary Pulsar 9
26.3 Gravitational Waves and their Propagation 11
26.3.1 The gravitational wave equation 11
26.3.2 The waves' two polarizations: + and × 14
26.3.3 Gravitons and their spin 18
26.3.4 Energy and Momentum in Gravitational Waves 19
26.3.5 Wave propagation in a source's local asymptotic rest frame 21
26.3.6 Wave propagation via geometric optics 23
26.3.7 Metric perturbation; TT gauge 25
26.4 The Generation of Gravitational Waves 28
26.4.1 Multipole-moment expansion 28
26.4.2 Quadrupole-moment formalism 29
26.4.3 Gravitational waves from a binary star system 33
26.5 The Detection of Gravitational Waves 38
26.5.1 Interferometer analyzed in TT gauge 40
26.5.2 Interferometer analyzed in proper reference frame of beam splitter . . 43
26.5.3 Realistic Interferometers 45

Chapter 26

Gravitational Waves and Experimental Tests of General Relativity

Version 0626.1.K.pdf, 16 May 2007.
Please send comments, suggestions, and errata via email to kip@tapir.caltech.edu or on paper to Kip Thorne, 130-33 Caltech, Pasadena CA 91125

Box 26.1
Reader's Guide

- This chapter relies significantly on
  - The special relativity portions of Chap. 1.
  - Chapter 23, on the transition from special relativity to general relativity.
  - Chapter 24, on the fundamental concepts of general relativity, especially Sec. 24.9 on weak, relativistic gravitational fields.
  - Chapter 25, on relativistic stars and black holes.
  - Sec. 6.3 on geometric optics.
- In addition, Sec. 26.2.3 on Fermat's principle and gravitational lenses is closely linked to Sec. 6.4 on gravitational lenses and Sec. 7.6 on diffraction at a caustic.
- Portions of this chapter are a foundation for Chap. 27, Cosmology.

26.1 Introduction

In 1915, when Einstein formulated general relativity, human technology was incapable of providing definitive experimental tests of his theory. Only a half century later did technology begin to catch up. In the remaining 35 years of the century, experiments improved from

accuracies of a few tens of per cent to a part in 1000 or even 10,000; and general relativity passed the tests with flying colors. In Sec. 26.2 we shall describe some of these tests, derive general relativity's predictions for them, and discuss the experimental results.

In the early twenty-first century, observations of gravitational waves will radically change the character of research on general relativity. They will produce, for the first time, tests of general relativity in strong-gravity situations. They will permit us to study relativistic effects in neutron-star and black-hole binaries with exquisite accuracies. They will enable us to map the spacetime geometries of black holes with high precision, and study observationally the large-amplitude, highly nonlinear vibrations of curved spacetime that occur when two black holes collide and merge. And (as we shall see in Chap. 27), they may enable us to probe the singularity in which the universe was born and the universe's evolution in its first tiny fraction of a second.

In this chapter we shall develop the theory of gravitational waves in much detail and shall describe the efforts to detect the waves and the sources that may be seen. More specifically, in Sec. 26.3 we shall develop the mathematical description of gravitational waves, both classically and quantum mechanically (in the language of gravitons), and shall study their propagation through flat spacetime and also, via the tools of geometric optics, through curved spacetime. Then in Sec. 26.4 we shall develop the simplest approximate method for computing the generation of gravitational waves, the "quadrupole-moment formalism"; and we shall describe and present a few details of other, more sophisticated and accurate methods based on multipolar expansions, post-Newtonian techniques, and numerical simulations on supercomputers ("numerical relativity"). In Sec. 26.5, we shall turn to gravitational-wave detection, focusing especially on detectors such as LIGO and LISA that rely on laser interferometry.

26.2 Experimental Tests of General Relativity

In this section we shall describe briefly some of the most important experimental tests of general relativity. For greater detail, see Will (1993, 2001, 2005)

26.2.1 Equivalence Principle, Gravitational redshift, and Global Positioning System

A key aspect of the equivalence principle is the prediction that all objects, whose size is extremely small compared to the radius of curvature of spacetime and on which no non-gravitational forces act, should move on geodesics. This means, in particular, that their trajectories through spacetime should be independent of their chemical composition. This is called the *weak equivalence principle* or the *universality of free fall*. Efforts to test the universality of free fall date back to Galileo's (perhaps apocryphal) experiment of dropping objects from the leaning tower of Pisa. In the twentieth century a sequence of ever-improving experiments led by Roland von Eötvös (1920), Robert Dicke (1964), Vladimir Braginsky (1972), and Eric Adelberger (1994) have led to an accuracy $\Delta a/a < 5 \times 10^{-13}$ for the difference of gravitational acceleration toward the Sun for earth-bound bodies with very different chem-

ical composition. A proposed space experiment called *STEP* has the prospect to increase this accuracy to the phenomenal level of $\Delta a/a \lesssim 1 \times 10^{-18}$.

General relativity predicts that bodies with significant self gravity (even black holes) should also fall, in a nearly homogeneous external gravitational field, with the same acceleration as a body with negligible self gravity. This prediction has been tested by comparing the gravitational accelerations of the Earth and Moon toward the Sun. Their fractional difference of acceleration [as determined by tracking the relative motions of the Moon and Earth using laser beams fired from Earth, reflected off mirrors that astronauts and cosmonauts have placed on the moon, and received back at earth] has been measured by the *LURE* Project to be $\Delta a/a \lesssim 3 \times 10^{-13}$. Since the Earth and Moon have (gravitational potential energy)/(rest-mass energy) $\simeq -5 \times 10^{-10}$ and $\simeq -2 \times 10^{-10}$ respectively, this verifies that gravitational energy falls with the same acceleration as other forms of energy to within about a part in 1000. For references and for discussions of a variety of other tests of the Equivalence Principle, see Will (1993, 2001, 2005).

From the equivalence principle, one can deduce that, for an emitter and absorber at rest in a Newtonian gravitational field Φ, light (or other electromagnetic waves) must be gravitationally redshifted by an amount $\Delta\lambda/\lambda = \Delta\Phi$, where $\Delta\Phi$ is the difference in Newtonian potential between the locations of the emitter and receiver. (See Ex. 25.5 for a general relativistic derivation when the field is that of a nonspinning, spherical central body with the emitter on the body's surface and the receiver far from the body.) Relativistic effects will produce a correction to this shift with magnitude $\sim (\Delta\Phi)^2$ [cf. Eq. (25.54)], but for experiments performed in the solar system, the currently available precision is too poor to see this correction; so such experiments test the equivalence principle and not the details of general relativity.

The highest precision test of this gravitational redshift thus far was NASA's 1976 Gravity-Probe-B Project (led by Robert Vessot), in which several atomic clocks were flown to a height of about 10,000 km above the earth, and were compared with atomic clocks on the earth via radio signals transmitted downward. After correcting for special relativistic effects due to the relative motions of the rocket's clocks and the earth clocks, the measured gravitational redshift agreed with the prediction to within the experimental accuracy of about 2 parts in 10,000.

The Global Positioning System (GPS), by which one can determine one's location on Earth to within an accuracy of about 10 meters, is based on signals transmitted from a set of earth-orbiting satellites. Each satellite's position is encoded on its transmitted signals, together with the time of transmission as measured by atomic clocks onboard the satellite. A person's GPS receiver contains a high-accuracy clock and a computer. It measures the signal arrival time and compares with the encoded transmission time to determine the distance from satellite to receiver; and it uses that distance, for several satellites, together with the encoded satellite positions, to determine (by triangulation) the receiver's location on earth.

The transmission times encoded on the signals are corrected for the gravitational redshift before transmission. Without this redshift correction, the satellite clocks would quickly get out of synchronization with all the clocks on the ground, thereby eroding the GPS accuracy; see Ex. 26.1. Thus, a good understanding of general relativity was crucial to the design of

the GPS![1]

26.2.2 Perihelion advance of Mercury

It was known at the end of the 19'th century that the point in Mercury's orbit closest to the Sun, known as its perihelion, advances at a rate of about 575″ per century with respect to the fixed stars, of which about 532″ can be accounted for by Newtonian perturbations of the other planets. The remaining $\sim 43''$ per century was a mystery until Einstein showed that it can be accounted for quantitatively by the general theory of relativity.

More specifically (as is demonstrated in Ex. 26.2), if we idealize the Sun as nonrotating and spherical so its external gravitational field is Schwarzschild, and we ignore the presence of the other planets, and we note that the radius of Mercury's orbit is very large compared to the Sun's mass (in geometrized units), then Mercury's orbit will be very nearly an ellipse; and the ellipse's perihelion will advance, from one orbit to the next, by an angle

$$\boxed{\Delta\phi = 6\pi M/p + \mathcal{O}(M^2/p^2) \quad \text{radians.}} \tag{26.1}$$

Here M is the Sun's mass and p is the ellipse's *semi latus rectum*, which is related to its semimajor axis a (half its major diameter) and its eccentricity e by $p = a(1-e^2)$. For the parameters of Mercury's orbit ($M = M_\odot \simeq 1.4766$ km, $a = 5.79089 \times 10^7$ km, $e = 0.205628$), this advance is $0.10352''$ per orbit. Since the orbital period is 0.24085 Earth years, this shift corresponds to 42.98 arc seconds per century.

Although the Sun is not precisely spherical, its tiny gravitational oblateness (as inferred from measurements of its spectrum of pulsations; Fig. 15.2) has been shown to contribute negligibly to this perihelion shift; and the frame dragging due to the Sun's rotational angular momentum is also (sadly!) negligible compared to the experimental accuracy; so 42.98″ per century would be the shift if the Sun and Mercury were the only objects in the solar system. The gravitational fields of the other planets, however, tug on Mercury's orbit, producing—according to Newtonian theory—the large additional shift of about 532″ per century. The weakness of gravity in the solar system guarantees that relativistic corrections to this additional shift are negligible, and that this shift can be added linearly to the Schwarzschild prediction of 42.98″, to within the accuracy of the measurements. When this is done and comparison is made with experiment, the 42.98″ prediction agrees with the observations to within the data's accuracy of about 1 part in 1000.

26.2.3 Gravitational deflection of light, Fermat's principle and Gravitational Lenses

Einstein not only explained the anomalous perihelion shift of Mercury. He also predicted [Ex. 26.3] that the null rays along which starlight propagates will be deflected, when passing through the curved spacetime near the Sun, by an angle

$$\boxed{\Delta\phi = 4M/b + \mathcal{O}(M^2/b^2)\ ,} \tag{26.2}$$

[1] For further details of the GPS see http://www.BeyondDiscovery.org

relative to their trajectories if spacetime were flat. Here M is the Sun's mass and b is the ray's impact parameter (distance of closest approach to the Sun's center). For comparison, theories that incorporated a Newtonian-like gravitational field into special relativity (Sec. 24.1) predicted half this deflection. The deflection was measured to an accuracy ~ 20 per cent during the 1919 solar eclipse and agreed with general relativity rather than the competing theories—a triumph that helped make Einstein world famous. Modern experiments, based on the deflection of radio waves from distant quasars, as measured using Very Long Baseline Interferometry (interfering the waves arriving at radio telescopes with transcontinental or transworld separations; Sec. 8.3), have achieved accuracies of about 1 part in 10,000, and they agree completely with general relativity. Similar accuracies are now achievable using optical interferometers in space, and may soon be achievable via optical interferometry on the ground.

These accuracies are so great that, when astronomers make maps of the sky using either radio interferometers or optical interferometers, they must now correct for gravitational deflection of the rays *not only when the rays pass near the sun, but for rays coming in from nearly all directions*. This correction is not quite as easy as Eq. (26.2) suggests, since that equation is valid only when the telescope is much farther from the Sun than the impact parameter. In the more general case, the correction is more complicated, and must include aberration due to the telescope motion as well as the effects of spacetime curvature.

As we discussed in Sec. 6.6, the gravitational deflection of light rays (or radio rays) passing through or near a cluster of galaxies can produce a spectacular array of distorted images of the light source. In Chap. 6 we deduced the details of this *gravitational lens effect* using a model in which we treated spacetime as flat, but endowed with a refractive index $n(\mathbf{x}) = 1-2\Phi(\mathbf{x})$, where $\Phi(\mathbf{x})$ is the Newtonian gravitational potential of the lensing system. This model can also be used to compute light deflection in the solar system. We shall now derive this model from general relativity:

The foundation for this model is the following general relativistic version of *Fermat's principle* [see Eq. (6.42) for the Newtonian version]: Consider any static spacetime geometry, i.e. one for which one can introduce a coordinate system in which $\partial g_{\alpha\beta}/\partial t = 0$ and $g_{jt} = 0$; so the only nonzero metric coefficients are $g_{00}(x^j)$ and $g_{0i}(x^j)$. In such a spacetime the time coordinate t is very special, since it is tied to the spacetime's temporal symmetry. An example is Schwarzschild spacetime and the Schwarzschild time coordinate t. Now, consider a light ray emitted from a spatial point $x^j = a^j$ in the static spacetime and received at a spatial point $x^j = b^j$. Assuming the spatial path along which the ray travels is $x^j(\eta)$ where η is any parameter with $x^j(0) = a^j$, $x^j(1) = b^j$, then the total coordinate time Δt required for the light's trip from a^j to b^j (as computed from the fact that the ray must be null so $ds^2 = g_{00}dt^2 + g_{ij}dx^i dx^j = 0$) is

$$\boxed{\Delta t = \int_0^1 \sqrt{\gamma_{jk}\frac{dx^j}{d\eta}\frac{dx^k}{d\eta}}d\eta \ , \quad \text{where} \quad \gamma_{jk} \equiv \frac{g_{jk}}{-g_{00}} \ .} \tag{26.3}$$

Fermat's principle says that *the actual spatial trajectory of the light path is the one that extremizes this coordinate time lapse.*

To prove this version of Fermat's principle, notice that the action (26.3) is the same as

that [Eq. (24.30)] for a geodesic in a 3-dimensional space with metric γ_{jk} and with t playing the role of proper distance traveled. Therefore, the Euler-Lagrange equation for Fermat's action principle $\delta\Delta t = 0$ is the geodesic equation in that space [Eq. (24.26)] with t the affine parameter, which [using Eq. (23.38) for the connection coefficients] can be written in the form

$$\gamma_{jk}\frac{d^2x^k}{dt^2} + \frac{1}{2}(\gamma_{jk,l} + \gamma_{jl,k} - \gamma_{kl,j})\frac{dx^k}{dt}\frac{dx^j}{dt} = 0 \ . \tag{26.4}$$

Next, take the geodesic equation (24.26) for the light ray in the real spacetime, with spacetime affine parameter ζ, and change parameters to t, thereby obtaining

$$\begin{aligned} &g_{jk}\frac{d^2x^k}{dt^2} + \Gamma_{jkl}\frac{dx^k}{dt}\frac{dx^l}{dt} - \Gamma_{j00}\frac{g_{kl}}{g_{00}}\frac{dx^k}{dt}\frac{dx^l}{dt} + \frac{d^2t d\zeta^2}{(dt/d\zeta)^2}g_{jk}\frac{dx^k}{dt} = 0 \ , \\ &\frac{d^2t/d\zeta^2}{(dt/d\zeta)^2} + 2\Gamma_{0k0}\frac{dx^k/dt}{g_{00}} = 0 \ . \end{aligned} \tag{26.5}$$

Insert the second of these equations into the first and write the connection coefficients in terms of derivatives of the spacetime metric. Then with a little algebra you can bring the result into the form (26.4) of the Fermat-principle Euler equation. *Therefore, the null geodesics of spacetime, when viewed as trajectories through the 3-space of constant t, are precisely the Fermat-principle paths, i.e. geodesics in a 3-space with metric γ_{jk} and proper-distance affine parameter t. QED*

The index-of-refraction formalism used to study gravitational lenses in Chap. 6 is easily deduced as a special case of this Fermat principle: In a nearly Newtonian situation, the linearized-theory, Lorentz-gauge, trace-reversed metric perturbation has the form (24.107) with only the time-time component being significantly large: $\bar{h}_{00} = -4\Phi$, $\bar{h}_{0j} \simeq 0$, $\bar{h}_{jk} \simeq 0$. Correspondingly, the metric perturbation [obtained by inverting Eq. (24.101)] is $h_{00} = -2\Phi$, $h_{jk} = -\delta_{jk}\Phi$, and the full spacetime metric $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$ is

$$ds^2 = -(1 + 2\Phi)dt^2 + (1 - 2\Phi)\delta_{jk}dx^j dx^k \ . \tag{26.6}$$

This is the standard spacetime metric (24.95) in the Newtonian limit, with a special choice of spatial coordinates, those of linearized-theory Lorentz gauge. The Newtonian limit includes the slow-motion constraint that time derivatives of the metric are small compared to spatial derivatives [Eq. (24.88)], so on the timescale for light to travel through a lensing system, the Newtonian potential can be regarded as static, $\Phi = \Phi(x^j)$. Therefore the Newtonian-limit metric (26.6) is static, and the coordinate time lapse along a trajectory between two spatial points, Eq. (26.3), reduces to

$$\Delta t = \int_0^1 (1 - 2\Phi)d\ell \ , \tag{26.7}$$

where $d\ell = \sqrt{\delta_{jk}dx^j dx^k}$ is distance traveled treating the coordinates as though they were Cartesian, in flat space. This is precisely the action for the Newtonian version of Fermat's principle, Eq. (6.42), with index of refraction

$$\boxed{n(x^j) = 1 - 2\Phi(x^j) \ .} \tag{26.8}$$

Therefore, the spatial trajectories of the light rays can be computed via the Newtonian Fermat principle, with the index of refraction (26.8). *QED*

Although this index-of-refraction model involves treating a special (Lorentz-gauge) coordinate system as though the spatial coordinates were Cartesian and space were flat (so $d\ell^2 = \delta_{jk} dx^j dx^k$)— which does not correspond to reality—, nevertheless, this model predicts the correct gravitational lens images. The reason is that it predicts the correct rays through the Lorentz-gauge coordinates, and when the light reaches Earth, the cumulative lensing has become so great that the fact that the coordinates here are slightly different from truly Cartesian has negligible influence on the images one sees.

26.2.4 Shapiro time delay

In 1964 Irwin Shapiro proposed a new experiment to test general relativity: Monitor the round-trip travel time for radio waves transmitted from earth and bounced off Venus or some other planet, or transponded by a spacecraft. As the line-of-sight between the Earth and the planet or spacecraft gradually moves nearer then farther from the Sun, the waves' rays will pass through regions of greater or smaller spacetime curvature, and this will influence the round-trip travel time by greater or smaller amounts. From the time evolution of the round-trip time, one can deduce the changing influence of the Sun's spacetime curvature.

One can compute the round-trip travel time with the aid of Fermat's principle. The round-trip proper time, as measured on Earth (neglecting, for simplicity, the Earth's orbital motion; i.e., pretending the Earth is at rest relative to the Sun while the light goes out and back) is $\Delta\tau_\oplus = \sqrt{1-2M/r_\oplus}\,\Delta t \simeq (1 - M/r_\oplus)\Delta t$, where M is the Sun's mass, $r_\oplus$ is the Earth's distance from the Sun's center, Δt is the round-trip coordinate time in the static solar-system coordinates, and we have used $g_{00} = 1 - 2M/r_\oplus$. Because Δt obeys Fermat's principle, it is stationary under small perturbations of the the light's spatial trajectory. This allows us to compute it using a straight-line trajectory through the spatial coordinate system. Letting b be the impact parameter (the ray's closest coordinate distance to the Sun) and x be coordinate distance along the straight-line trajectory and neglecting the gravitational fields of the planets, we have $\Phi = -M/\sqrt{x^2+b^2}$, so the coordinate time lapse out and back is

$$\Delta t = 2\int_{-\sqrt{r_\oplus^2-b^2}}^{\sqrt{r_{\rm refl}^2-b^2}} \left(1 + \frac{2M}{\sqrt{x^2+b^2}}\right) dx\ . \tag{26.9}$$

Here $r_{\rm refl}$ is the radius of the location at which the light gets reflected (or transponded) back to Earth. Performing the integral and multiplying by $\sqrt{g_{00}} \simeq 1 - M/r_\oplus$, we obtain for the round-trip travel time measured on Earth

$$\Delta\tau_\oplus = 2\,(a_\oplus + a_{\rm refl})\left(1 - \frac{M}{r_\oplus}\right) + 4M\ln\left[\frac{(a_\oplus + r_\oplus)(a_{\rm refl} + r_{\rm refl})}{b^2}\right]\ , \tag{26.10}$$

where $a_\oplus = \sqrt{r_\oplus^2 - b^2}$ and $a_{\rm refl} = \sqrt{r_{\rm refl}^2 - b^2}$.

As the Earth and the reflecting planet or transponding spacecraft move along their orbits, only one term in this round-trip time varies sharply: the term

$$\boxed{\Delta\tau_\oplus = 4M\ln(1/b^2) = 8M\ln b \simeq 40\mu{\rm s}(b/R_\odot)\ .} \tag{26.11}$$

When the planet or spacecraft passes nearly behind the Sun, as seen from Earth, b plunges to a minimum (on a timescale of hours or days) then rises back up, and correspondingly the time delay shows a sharp blip. By comparing the observed blip with the theory in a measurement with the Cassini spacecraft, this Shapiro time delay has been verified to the remarkable precision of about 1 part in 100,000 (Bertotti, Iess and Tortora 2003).

26.2.5 Frame dragging and Gravity Probe B

As we have discussed in Secs. 24.9.3 and 25.5, the rotational angular momentum $\mathbf{J}$ of a gravitating body places its imprint on the body's asymptotic spacetime metric:

$$ds^2 = -\left(1 - \frac{2M}{r}\right) dt^2 - \frac{4\epsilon_{jkm} J^k x^m}{r^3} dt dx^j + \left(1 + \frac{2M}{r}\right) \delta_{jk} dx^j dx^k \ . \tag{26.12}$$

Here, for definiteness, we are using Lorentz gauge, and M is the body's mass; cf. Eq. (24.112). The angular-momentum term drags inertial frames into rotation about the body (Sec. 25.2). One manifestation of this frame dragging is a precession of inertial-guidance gyroscopes near the body. Far from the body, a gyroscope's spin axis will remain fixed relative to distant galaxies and quasars, but near the body it will precess.

It is easy to deduce the precession in the simple case of a gyroscope whose center of mass is at rest in the coordinate system of Eq. (26.12), i.e. at rest relative to the body. The transport law for the gyroscope's spin is $\nabla_{\vec{u}} \vec{S} = \vec{u}(\vec{a} \cdot \vec{S})$ [Eq. (23.91) boosted from special relativity to general relativity via the equivalence principle]. Here $\vec{u}$ is the gyroscope's 4-velocity (so $u^j = 0$, $u^0 = 1/\sqrt{1-2M/r} \simeq 1 + M/r \simeq 1$) and $\vec{a}$ its 4-acceleration. The spatial components of this transport law are

$$S^j{}_{,t} u^0 \simeq S^j{}_{,t} = -\Gamma^j{}_{k0} S^k u^0 \simeq -\Gamma^j{}_{k0} S^k \simeq -\Gamma_{jk0} S^k \simeq \frac{1}{2}(g_{0k,j} - g_{0j,k}) S^k \ . \tag{26.13}$$

Here each $\simeq$ means "is equal, up to fractional corrections of order M/r". By inserting g_{j0} from the line element (26.12) and performing some manipulations with Levi-Civita tensors, we can bring Eq. (26.13) into the form

$$\boxed{\frac{\partial \mathbf{S}}{\partial t} = \mathbf{\Omega}_{\rm prec} \times \mathbf{S} \ , \quad \text{where} \quad \mathbf{\Omega}_{\rm prec} = \frac{1}{r^3}[-\mathbf{J} + 3(\mathbf{J} \cdot \mathbf{n})\mathbf{n}] \ .} \tag{26.14}$$

Here $\mathbf{n} = \mathbf{e}_{hatr}$ is the unit radial vector pointing away from the gravitating body. Equation (26.14) says that the gyroscope's spin angular momentum rotates (precesses) with angular velocity $\mathbf{\Omega}_{\rm prec}$ in the coordinate system (which is attached to distant inertial frames, i.e. to the distant galaxies and quasars). This is sometimes called a "gravitomagnetic precession" because the off-diagonal term g_{j0} in the metric, when thought of as a 3-vector, is $-4\mathbf{J} \times \mathbf{n}/r^2$, which has the same form as the vector potential of a magnetic dipole; and the gyroscopic precession is similar to that of a magnetized spinning body interacting with that magnetic dipole.

In magnitude, the precessional angular velocity (26.14) in the vicinity of the Earth is roughly one arcsec per century, so measuring it is a tough experimental challenge. A

team led by Francis Everitt has designed and constructed a set of superconducting gyroscopes that are currently (2005) flying in an Earth-orbiting satellite called Gravity Probe B, with the goal of measuring this precession to a precision of about 1 part in 100 (see http://einstein.stanford.edu/).

26.2.6 Binary Pulsar

Gravity in the solar system is very weak. Even at Mercury's orbit, the gravitational potential of the Sun is only $|\Phi| \sim 3 \times 10^{-8}$. Therefore, when one expands the spacetime metric in powers of Φ, current experiments with their fractional accuracies $\sim 10^{-4}$ or worse are able to see only the first-order terms beyond Newtonian theory; i.e. terms of *first post-Newtonian order*. To move on to second post-Newtonian order, $\mathcal{O}(\Phi^2)$ beyond Newton, will require major advances in technology, or observations of astronomical systems in which Φ is far larger than 3×10^{-8}.

Radio observations of binary pulsars (this subsection) provide one opportunity for such observations; gravitational-wave observations of neutron-star and black-hole binaries (Sec. 26.5) provide another.

The best binary pulsar for tests of general relativity is PSR1913+16, discovered by Russell Hulse and Joseph Taylor in 1974. This system consists of two neutron stars in a mutual elliptical orbit with period $P \sim 8$ hr and eccentricity $e \sim 0.6$. One of the stars emits pulses at a regular rate. These are received at earth with time delays due to crossing the binary orbit and other relativistic effects. We do not know *a priori* the orbital inclination or the neutron-star masses. However, we obtain one relation between these three quantities by analyzing the Newtonian orbit. A second relation comes from measuring the consequences of the combined second order Doppler shift and gravitational redshift as the pulsar moves in and out of its companion's gravitational field. A third relation comes from measuring the relativistic precession of the orbit's periastron (analog of the perihelion shift of Mercury). (The precession rate is far larger than for Mercury: about 4° per year!) From these three relations one can solve for the stars' masses and the orbital inclination, and as a check can verify that the Shapiro time delay comes out correctly. One can then use the system's parameters to predict the rate of orbital inspiral due to gravitational-radiation reaction—a phenomenon with magnitude $\sim |\Phi|^{2.5}$ beyond Newton, i.e. 2.5 post-Newtonian order (Sec. 26.4.2 below). The prediction agrees with the measurements to accuracy ~ 0.1 per cent (Weissberg and Taylor 2004) —a major triumph for general relativity!

****

EXERCISES

Exercise 26.1 *Practice: Gravitational Redshift for Global Positioning System*
The GPS satellites are in circular orbits at a height of 18,000 km above the Earth's surface. If the ticking rates of the clocks on the satellites were not corrected for the gravitational redshift, roughly how long would it take them to accumulate a time shift, relative to clocks on the earth, large enough to degrade the GPS position accuracy by 10 meters? by 1 kilometer?

Exercise 26.2 *Example: Perihelion Shift*
Consider a small satellite in non-circular orbit about a spherical body with much larger mass M, for which the external gravitational field is Schwarzschild. The satellite will follow a timelike geodesic. Orient the Schwarzschild coordinates so the satellite's orbit is in the equatorial plane, $\theta = \pi/2$.

(a) Because the metric coefficients are independent of t and ϕ, the quantities $\tilde{E} = -p_t$ and $\tilde{L} = p_\phi$ must be constants of the satellite's motion [cf. Ex. 24.4]. Show that

$$\tilde{E} = \left(1 - \frac{2M}{r}\right)\frac{dt}{d\tau}\ ,$$
$$\tilde{L} = r^2\frac{d\phi}{d\tau}\ . \tag{26.15}$$

Explain why $\tilde{E}$ has the physical interpretation of the satellite's orbital energy per unit mass (including rest-mass energy) and why $\tilde{L}$ is its angular momentum per unit mass.

(b) Introduce the coordinate $u = r^{-1}$ and use the normalization of the 4-velocity to derive the following differential equation for the orbit:

$$\left(\frac{du}{d\phi}\right)^2 = \frac{\tilde{E}^2}{\tilde{L}^2} - \left(u^2 + \frac{1}{\tilde{L}^2}\right)(1 - 2Mu)\ . \tag{26.16}$$

(c) Differentiate this equation with respect to ϕ to obtain a second order differential equation

$$\frac{d^2u}{d\phi^2} + u - \frac{M}{\tilde{L}^2} = 3Mu^2. \tag{26.17}$$

By reinstating the constants G, c, and comparing with the Newtonian orbital equation, argue that the right-hand side represents a relativistic perturbation to the Newtonian equation of motion.

(d) Assume, henceforth in this exercise, that $r \gg M$ (i.e. $u \ll 1/M$), and solve the orbital equation (26.17) by perturbation theory. More specifically: At zero order (i.e., setting the right side to zero), show that the Kepler ellipse

$$u_K = \left(\frac{M}{\tilde{L}^2}\right)(1 + e\cos\phi), \tag{26.18}$$

is a solution. Here e (a constant of integration) is the ellipse's eccentricity and $\tilde{L}^2/M$ is the ellipse's *semi latus rectum*. The orbit has its minimum radius at $\phi = 0$.

(e) By substituting u_K into the right hand side of the relativistic equation of motion (26.17), show (at first-order in the relativistic perturbation) that in one orbit the angle ϕ at which the satellite is closest to the mass advances by $\Delta\phi \simeq 6\pi M^2/\tilde{L}^2$. (Hint: Try to write the differential equation in the form $d^2u/d\phi^2 + (1+\epsilon)^2u \simeq \ldots$, where $\epsilon \ll 1$.)

(f) For the planet Mercury, the orbital period is $P = 0.241$ yr and the eccentricity is $e = 0.206$. Deduce that the relativistic contribution to the rate of advance of the *perihelion* (point of closest approach to the Sun) is $43''$ per century.

Exercise 26.3 *Example: Gravitational Deflection of Light.*
Repeat the previous exercise for a photon following a null geodesic.

(a) Show that the trajectory obeys the differential equation

$$\frac{d^2u}{d\phi^2} + u = 3Mu^2 \ . \tag{26.19}$$

(b) Obtain the zero'th order solution by ignoring the right hand side,

$$u = \frac{\sin\phi}{b} \ . \tag{26.20}$$

where b is an integration constant. Show that, in the asymptotically flat region far from the body, this is just a straight line and b is the impact parameter (distance of closest approach to the body).

(c) Substitute this solution into the right hand side and show that the perturbed trajectory satisfies

$$u = \frac{\sin\phi}{b} + \frac{M}{b^2}(1-\cos\phi)^2 \ . \tag{26.21}$$

(d) Hence show that a ray with impact parameter $b \gg M$ will be deflected through an angle

$$\alpha = \frac{4M}{b} \ ; \tag{26.22}$$

cf. Eq. (6.77) and associated discussion.

****

26.3 Gravitational Waves and their Propagation

26.3.1 The gravitational wave equation

Gravitational waves are *ripples in the curvature of spacetime* that are emitted by violent astrophysical events, and that propagate out from their sources with the speed of light. It was clear to Einstein and others, even before general relativity was fully formulated, that his theory would have to predict gravitational waves; and within months after completing the theory, Einstein (1916, 1918) worked out the basic properties of those waves.

It turns out that, after they have been emitted, gravitational waves propagate through matter with near impunity, i.e., they propagate as though in vacuum, even when other matter and fields are present. (For a proof and discussion see, e.g., Sec. 2.4.3 of Thorne, 1983). This

justifies simplifying our analysis to vacuum propagation. By contrast with most texts on gravitational waves, we shall *not* further simplify to propagation through a spacetime that is flat, aside from the waves, because it is almost as easy to analyze propagation through a curved background spacetime as a flat one.

The key to the analysis is the same two-lengthscale expansion as underlies geometric optics for any kind of wave propagating through any kind of medium (Sec. 6.3): We presume that the waves' reduced wavelength $\lambda\!\!\!^{-}$ (wavelength/2π) as measured in some relevant local Lorentz frame is very short compared to the radius of curvature of spacetime $\mathcal{R} \sim 1/\sqrt{R_{\hat{\alpha}\hat{\beta}\hat{\gamma}\hat{\delta}}}$ and the lengthscale $\mathcal{L}$ on which the background curvature changes (e.g., the radius of the Earth when the waves are near Earth):

$$\boxed{\lambda\!\!\!^{-} \ll \{\mathcal{R}, \mathcal{L}\}} \tag{26.23}$$

cf. Eq. (6.14). Then the Riemann curvature tensor can be split into two pieces: The background curvature $R^{\rm B}_{\alpha\beta\gamma\delta}$, which is the average of Riemann over a few wavelengths, plus the waves' curvature $R^{\rm GW}_{\alpha\beta\gamma\delta}$, which is the remaining, oscillatory piece:

$$\boxed{R_{\alpha\beta\gamma\delta} = R^{\rm B}_{\alpha\beta\gamma\delta} + R^{\rm GW}_{\alpha\beta\gamma\delta} \ , \quad R^{\rm B}_{\alpha\beta\gamma\delta} \equiv \langle R_{\alpha\beta\gamma\delta} \rangle \ .} \tag{26.24}$$

This is the same kind of split as we used in developing the quasilinear theory of plasma waves (Sec. 22.2.1). Similarly, we can split the spacetime metric into a sum of a smooth background part plus a gravitational-wave perturbation, denoted $h_{\alpha\beta}$

$$\boxed{g_{\alpha\beta} = g^{\rm B}_{\alpha\beta} + h_{\alpha\beta} \ ; \quad \text{where} \quad g^{\rm B}_{\alpha\beta} = \langle g_{\alpha\beta} \rangle \ .} \tag{26.25}$$

Obviously, the smooth background Riemann tensor $R^{\rm B}_{\alpha\beta\gamma\delta}$ can be computed in the usual manner from the smooth background metric $g^{\rm B}_{\alpha\beta}$.

Because the waves are generally very weak, we can regard their metric perturbation $h_{\alpha\beta}$ and Riemann tensor $R^{\rm GW}_{\alpha\beta\gamma\delta}$ as linearized fields that live in the smooth, curved background spacetime. When we do so, we can replace gradients (subscript ";") based on the full physical metric $g_{\alpha\beta}$ by gradients (subscript "|") based on the background metric so, e.g., $R^{\rm GW}_{\alpha\beta\gamma\delta;\mu} = R^{\rm GW}_{\alpha\beta\gamma\delta|\mu}$. This linearization implies that the waves' Riemann tensor can be computed from their metric perturbation via

$$R^{\rm GW}_{\alpha\beta\gamma\delta} = \frac{1}{2}(h_{\alpha\delta|\beta\gamma} + h_{\beta\gamma|\alpha\delta} - h_{\alpha\gamma|\beta\delta} - h_{\beta\delta|\alpha\gamma}) \ , \tag{26.26}$$

as one can see from the fact that this formula reduces to the right result, Eq. (24.96), in a local Lorentz frame of the background metric. We shall use the waves' Riemann tensor $R^{GW}_{\alpha\beta\gamma\delta}$ as our primary entity for describing the waves, and shall use the metric perturbation only as a computational tool—mostly when analyzing wave generation.

Notice that the combination of indices that appears on the right side of Eq. (26.26) is carefully designed to produce an entity with the symmetries of Riemann

$$R_{\alpha\beta\gamma\delta} = -R_{\beta\alpha\gamma\delta} \ , \quad R_{\alpha\beta\gamma\delta} = -R_{\alpha\beta\delta\gamma} \ , \quad R_{\alpha\beta\gamma\delta} = R_{\gamma\delta\alpha\beta} \ . \tag{26.27}$$

[Eq. (24.52)]. This combination of indices is encountered frequently in gravitational-wave theory, so it is useful to introduce the following short-hand notation for it:

$$\boxed{S_{\{\alpha\beta\gamma\delta\}} \equiv S_{\alpha\delta\beta\gamma} + S_{\beta\gamma\alpha\delta} - S_{\alpha\gamma\beta\delta} - S_{\beta\delta\alpha\gamma} \ .} \tag{26.28}$$

In terms of this notation, expression (26.26) reads

$$\boxed{R^{\rm GW}_{\alpha\beta\gamma\delta} = \frac{1}{2} h_{\{\alpha\beta|\gamma\delta\}} \ .} \tag{26.29}$$

One benefit of the two-lengthscale condition $\lambda\!\!\!^{-} \ll \mathcal{R}$ is the fact that the double gradient of the gravitational waves' Riemann tensor is far larger than the product of the waves' Riemann with the background Riemann

$$R^{\rm GW}_{\alpha\beta\gamma\delta|\mu\nu} \sim \frac{R^{\rm GW}_{\alpha\beta\gamma\delta}}{\lambda\!\!\!^{-2}} \gg \frac{R^{\rm GW}_{\alpha\beta\gamma\delta}}{\mathcal{R}^2} \sim R^{\rm GW}_{\alpha\beta\gamma\delta} R^{\rm B}_{\mu\nu\rho\sigma} \ . \tag{26.30}$$

Since the commutator of the double gradient is a sum of products of the wave Riemann with the background Riemann [generalization of Eq. (24.39) with p^α replaced by $R^{\rm GW}_{\alpha\beta\gamma\delta}$], *gradients of $R^{\rm GW}_{\alpha\beta\gamma\delta}$ commute to high accuracy*:

$$R^{\rm GW}_{\alpha\beta\gamma\delta|\mu\nu} = R^{\rm GW}_{\alpha\beta\gamma\delta|\nu\mu} \ . \tag{26.31}$$

We shall use this fact in deriving the wave equation for $R^{\rm GW}_{\alpha\beta\gamma\delta}$.

Our derivation of the wave equation will be based on a combination of the Riemann curvature's Bianchi identity

$$R^\alpha{}_{\beta\gamma\delta;\epsilon} + R^\alpha{}_{\beta\delta\epsilon;\gamma} + R^\alpha{}_{\beta\epsilon\gamma;\delta} = 0 \tag{26.32}$$

[Eq. (26.32)] and the vacuum Einstein field equation $G^{\alpha\beta} \equiv R^{\alpha\beta} - \frac{1}{2} R g^{\alpha\beta} = 0$. By contracting the vacuum field equation on its two slots, we find that the scalar curvature R vanishes, and by inserting this back into the vacuum field equation we find that the Ricci tensor vanishes:

$$R_{\alpha\beta} \ \equiv \ R^\mu{}_{\alpha\mu\beta} = 0 \quad \text{ in vacuum.} \tag{26.33}$$

By then contracting the Bianchi identity (26.32) on its first and fifth slots and invoking (26.33) we find that the Riemann tensor is divergence-free:

$$R^\mu{}_{\beta\gamma\delta;\mu} = 0 \quad \text{ in vacuum.} \tag{26.34}$$

The symmetries (26.27) guarantee that *Riemann is divergence-free not only on its first slot, but in fact, on each of its four slots.* By next taking the divergence of the Bianchi identity (26.32) on its last slot, we obtain

$$R_{\alpha\beta\gamma\delta;\mu}{}^\mu = -R_{\alpha\beta\delta\mu;\gamma}{}^\mu - R_{\alpha\beta\mu\gamma;\delta}{}^\mu \ . \tag{26.35}$$

We now split this equation into its rapidly oscillating (wave) piece and its background piece, and for the wave piece we approximate the full-spacetime gradients ";" by background-spacetime gradients "|", we commute the gradient indices on the right-hand side [Eq. (26.31)], and we use the vanishing of the divergence [Eq. (26.34)] to obtain

$$\boxed{R^{\rm GW}_{\alpha\beta\gamma\delta|\mu}{}^{\mu} = 0 \ .} \tag{26.36}$$

This is the wave equation for gravitational waves propagating through the curved, background spacetime. It is a perfect analog of the vacuum wave equation $A_{\alpha;\mu}{}^{\mu} = 0$ [Eq. (24.71)] for electromagnetic waves. Both wave equations dictate that their waves propagate at the speed of light ($c = 1$ in our geometrized units).

To get insight into the waves, we pick a region of spacetime far from the source, where the wavefronts are nearly flat, and in that region we introduce a local Lorentz frame of the background spacetime. This frame must be small compared to the background radius of curvature $\mathcal{R}$; but since $\lambda\!\!\!^{-} \ll \mathcal{R}$, the frame can still be big compared to $\lambda\!\!\!^{-}$. For example, for waves passing near and through Earth, in the frequency band $f \sim 100$ Hz of Earth-based detectors, $\mathcal{R}$ is about 10^9 km and $\lambda\!\!\!^{-}$ is about 500 km, so the local Lorentz frame could be given a size $\sim 10^5$ km (~ 10 times larger than the Earth), which is huge compared to $\lambda\!\!\!^{-}$ but small compared to $\mathcal{R}$.

In this local Lorentz frame, by virtue of Eqs. (24.15), (24.16) and (26.30), the wave equation (26.36) becomes

$$\left(-\frac{\partial^2}{\partial t^2} + \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) R^{\rm GW}_{\alpha\beta\gamma\delta} = 0 \ , \tag{26.37}$$

For simplicity, we orient the spatial axes of the local Lorentz frame so the waves propagate in the z-direction, and we neglect the curvature of the phase fronts (i.e., we treat the waves as planar). Then the solution to (26.37) is an arbitrary function of $t - z$:

$$\boxed{R^{\rm GW}_{\alpha\beta\gamma\delta} = R^{\rm GW}_{\alpha\beta\gamma\delta}(t - z) \ .} \tag{26.38}$$

This shows explicitly that the waves propagate with the speed of light.

26.3.2 The waves' two polarizations: + and ×

In this subsection we shall explore the properties of gravitational waves. Throughout the discussion we shall confine attention to the background's local Lorentz frame, far from the source, in which the waves are nearly planar and have the form (26.38).

Only two of the 20 independent components of $R^{\rm GW}_{\alpha\beta\gamma\delta}$ are independent functions of $t - z$; the other eighteen are determined in terms of those two by the following considerations: (i) The Bianchi identity (26.32), when applied to the specific functional form (26.38) and then integrated in time (with the integration constant dropped because we are studying waves that fluctuate in time), implies

$$R^{\rm GW}_{\alpha\beta xy} = 0 \ , \quad R^{\rm GW}_{\alpha\beta xz} = -R^{\rm GW}_{\alpha\beta x0} \ , \quad R^{\rm GW}_{\alpha\beta yz} = -R^{\rm GW}_{\alpha\beta y0} \ . \tag{26.39}$$

(*ii*) This, together with Riemann's symmetries (26.27), implies that all components can be expressed in terms of $R^{\rm GW}_{j0k0}$ (which is general relativity's analog of the Newtonian tidal tensor $\mathcal{E}_{jk}$). (*iii*) The vacuum Einstein equation (26.33) then implies that

$$\boxed{R^{\rm GW}_{z0z0} = R^{\rm GW}_{z0x0} = R^{\rm GW}_{x0z0} = R^{\rm GW}_{z0y0} = R^{\rm GW}_{y0z0} = 0\ ,} \tag{26.40}$$

and

$$\boxed{R^{\rm GW}_{x0x0} = -R^{\rm GW}_{y0y0} \equiv -\frac{1}{2}\ddot{h}_+(t-z)\ , \quad R^{\rm GW}_{x0y0} = R^{\rm GW}_{y0x0} \equiv -\frac{1}{2}\ddot{h}_\times(t-z)\ .} \tag{26.41}$$

Here the two independent components have been expressed in terms of dimensionless functions $h_+(t-z)$ and $h_\times(t-z)$. The double time derivatives, denoted by double dots ($\ddot{h}_+ \equiv \partial^2 h_+/\partial t^2$), are required by dimensionality: Riemann has dimensions of $1/{\rm length}^2$ or equivalently $1/{\rm time}^2$; so if h_+ and $h_\times$ are to be dimensionless, they must be differentiated twice in (26.41). The factors of $\frac{1}{2}$ are relics of the past history of general relativity research.

Equation (26.40) says that for a gravitational wave the space-time-space-time part of Riemann is *transverse*; i.e., it has no spatial components along the propagation direction (z-direction). This is completely analogous to the fact that the electric and magnetic fields of an electromagnetic wave are transverse to the propagation direction. The first of Eqs. (26.41) says that the nonvanishing, transverse-transverse part of Riemann is *traceless*. These two properties are often summarized by saying that gravitational waves are "transverse and traceless," or "TT."

The two independent functions h_+ and $h_\times$ are called the "gravitational-wave fields" for the "+ (plus) polarization state" and for the "× (cross) polarization state."

We can reconstruct all the components of the waves' Riemann tensor from these two gravitational-wave fields as follows: First define the polarization tensors

$$\boxed{\mathbf{e}^+ \equiv (\vec{e}_x \otimes \vec{e}_x - \vec{e}_y \otimes \vec{e}_y)\ , \quad \mathbf{e}^\times \equiv (\vec{e}_x \otimes \vec{e}_y + \vec{e}_y \otimes \vec{e}_x)\ ,} \tag{26.42}$$

and a second-rank gravitational-wave field

$$\boxed{h^{\rm TT}_{\alpha\beta} = h_+ e^+_{\alpha\beta} + h_\times e^\times_{\alpha\beta}\ ; \quad \text{or equivalently}} \tag{26.43a}$$

$$\boxed{h^{\rm TT}_{xx} = -h^{\rm TT}_{yy} = h_+\ , \quad h^{\rm TT}_{xy} = h^{\rm TT}_{yx} = h_\times\ , \quad \text{all other } h^{\rm TT}_{\alpha\beta} \text{ vanish.}} \tag{26.43b}$$

[The notation "TT" indicates that this field is transverse to the propagation direction (z-direction) and traceless. The relationship between this $h^{\rm TT}_{\alpha\beta}$ and the metric perturbation $h_{\alpha\beta}$ will be explained in Sec. 26.3.7 below.] Then the waves' Riemann tensor is

$$\boxed{R^{\rm GW}_{\alpha\beta\gamma\delta} = \frac{1}{2}h^{\rm TT}_{\{\alpha\beta|\gamma\delta\}}\ ; \quad \text{and in particular} \quad R^{\rm GW}_{0j0k} = -\frac{1}{2}\ddot{h}^{\rm TT}_{jk}\ .} \tag{26.44}$$

We shall seek physical insight into h_+ and $h_\times$ by studying the following idealized problem: Consider a cloud of test particles that floats freely in space and is static and spherical before

the waves pass. We shall study the wave-induced deformations of the cloud as viewed in the nearest thing there is to a rigid, orthonormal coordinate system: the local Lorentz frame (in the physical spacetime) of a "fiducial particle" that sits at the cloud's center. In that frame the displacement vector ξ^j between the fiducial particle and some other particle has components $\xi^j = x^j + \delta x^j$, where x^j is the other particle's spatial coordinate before the waves pass, and δx^j is its coordinate displacement, as produced by the waves. By inserting this into the local-Lorentz-frame variant of the equation of geodesic deviation, Eq. (24.42), and neglecting the tiny δx^k compared to x^k on the right side, we obtain

$$\frac{d^2\delta x^j}{dt^2} = -R^{\rm GW}_{j0k0}x^k = \frac{1}{2}\ddot{h}^{\rm TT}_{jk}x^k \ , \tag{26.45}$$

which can be integrated twice to give

$$\boxed{\delta x^j = \frac{1}{2}h^{\rm TT}_{jk}x^k \ .} \tag{26.46}$$

The middle expression in Eq. (26.45) is the *gravitational-wave tidal acceleration* that moves the particles back and forth relative to each other. It is completely analogous to the Newtonian tidal acceleration $-R_{j0k0}x^k = -(\partial^2\Phi/\partial x^j\partial x^k)x^k$ by which the moon raises tides on the earth's oceans [Sec. 24.5.1].

Specialize, now, to a wave with + polarization (for which $h_\times = 0$). By inserting expression (26.43) into (26.46), we obtain

$$\delta x = \frac{1}{2}h_+ x \ , \quad \delta y = -\frac{1}{2}h_+ y \ , \quad \delta z = 0 \ . \tag{26.47}$$

This displacement is shown in Fig. 26.1(a,b). Notice that, as the gravitational-wave field h_+ oscillates at the cloud's location, the cloud is left undisturbed in the z-direction (propagation direction), and in transverse planes it gets deformed into an ellipse elongated first along the x-axis (when $h_+ > 0$), then along the y-axis (when $h_+ < 0$). Because $R_{x0x0} = -R_{y0y0}$, i.e., because R_{j0k0} is traceless, the ellipse is squashed along one axis by the same amount as it is stretched along the other, i.e., the area of the ellipse is preserved during the oscillations.

The effects of the h_+ polarization state can also be described in terms of the *tidal acceleration field* that it produces in the central particle's local Lorentz frame:

$$\frac{d^2}{dt^2}\delta\mathbf{x} = \frac{1}{2}\ddot{h}_+(x\mathbf{e}_x - y\mathbf{e}_y) \ , \tag{26.48}$$

where $\ddot{h}_+ \equiv \partial^2 h_+/\partial t^2$. Notice that this acceleration field is divergence free. Because it is divergence-free, it can be represented by lines of force, analogous to electric field lines, which point along the field and have a density of lines proportional to the magnitude of the field; and when this is done, the field lines will never end. Figure 26.1(c,d) shows this acceleration field at the phases of oscillation when $\ddot{h}_+$ is positive and when it is negative. Notice that the field is quadrupolar in shape, with a field strength (density of lines) that increases linearly with distance from the origin of the local Lorentz frame. The elliptical deformations of the sphere of test particles shown in Fig. 26.1(a,b) are the responses of that

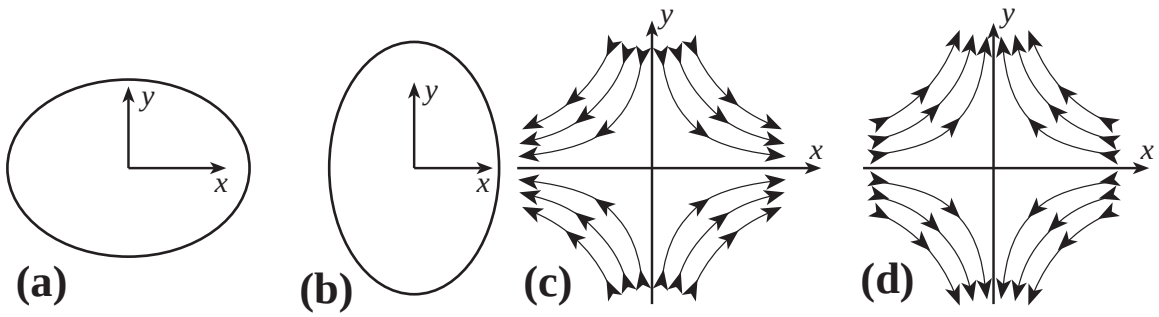


Fig. 26.1: Physical manifestations, in a particle's local Lorentz frame, of h_+ gravitational waves. (a) Transverse deformation of an initially spherical cloud of test particles at a phase of the wave when $h_+ > 0$. (b) Deformation of the cloud when $h_+ < 0$. (c) Field lines representing the acceleration field which produces the cloud's deformation, at a phase when $\ddot{h}_+ > 0$. (d) Acceleration field lines when $\ddot{h}_+ < 0$.

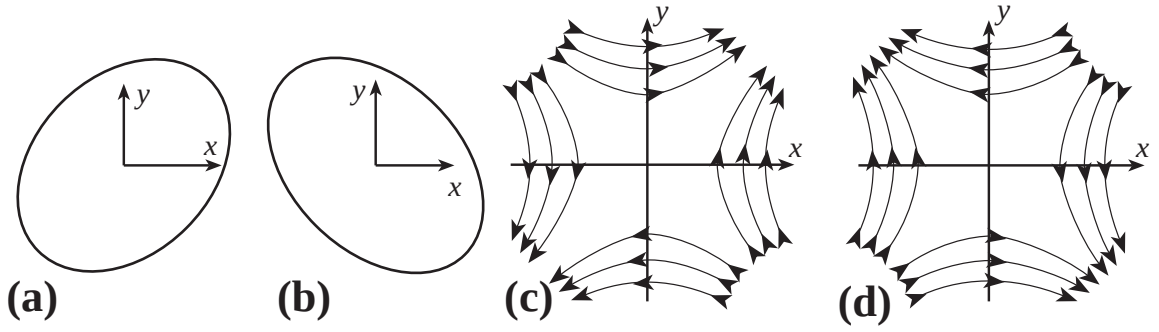


Fig. 26.2: Physical manifestations, in a particle's local Lorentz frame, of $h_\times$ gravitational waves. (a) Deformation of an initially circular sphere of test particles at a phase of the wave when $h_\times > 0$. (b) Deformation of the sphere when $h_\times < 0$. (c) Field lines representing the acceleration field which produces the sphere's deformation, at a phase of the wave when $\ddot{h}_\times > 0$. (d) Acceleration field lines when $\ddot{h}_\times < 0$.

sphere to this quadrupolar acceleration field. The polarization state which produces these accelerations and deformations is called the + state because of the orientation of the axes of the quadrupolar acceleration field [Fig. 26.1(c,d)].

Turn, next, to the × polarization state. In this state the deformations of the initially circular ring are described by

$$\delta x = \frac{1}{2} h_\times y \ , \quad \delta y = \frac{1}{2} h_\times x \ , \quad \delta z = 0 \ . \tag{26.49}$$

These deformations, like those for the + state, are purely transverse; they are depicted in Fig. 26.2(a,b). The acceleration field that produces these deformations is

$$\frac{d^2}{dt^2}\delta\mathbf{x} = \frac{1}{2}\ddot{h}_\times (y\mathbf{e}_x + x\mathbf{e}_y) \ . \tag{26.50}$$

This acceleration field, like the one for the + polarization state, is divergence free and quadrupolar; the field lines describing it are depicted in Fig. 26.2(c,d). The name "× polarization state" comes from the orientation of the axes of this quadrupolar acceleration field.

In defining the gravitational-wave fields h_+ and $h_\times$, we have relied on a choice of (local Lorentz) reference frame, i.e. a choice of local Lorentz basis vectors $\vec{e}_\alpha$. Exercise 26.4 explores how these fields change when the basis is changed. The conclusions are simple: (i) When one rotates the transverse basis vectors $\vec{e}_x$ and $\vec{e}_y$ through an angle ψ, then h_+ and $h_\times$ "rotate" through 2ψ in the sense that:

$$\boxed{(h_+ + ih_\times)_{\rm new} = (h_+ + ih_\times)_{\rm old} e^{2i\psi} \ , \quad \text{when} \quad (\vec{e}_x + i\vec{e}_y)_{\rm new} = (\vec{e}_x + i\vec{e}_y)e^{i\psi} \ .} \tag{26.51}$$

(ii) When one boosts from an old frame to a new one moving at some other speed, but chooses the old and new spatial bases such that (a) the waves propagate in the z direction in both frames and (b) the plane spanned by $\vec{e}_x$ and $\vec{\kappa} \equiv \vec{e}_0 + \vec{e}_z$ =(propagation direction in spacetime) is the same in both frames, then h_+ and $h_\times$ are the same in the two frames—i.e., they are scalars under such a boost!

26.3.3 Gravitons and their spin

Most of the above features of gravitational waves (though not expressed in this language) were clear to Einstein in 1918. Two decades later, as part of the effort to understand quantum fields, M. Fierz and Wolfgang Pauli (1939) at the Eidgenössische Technische Hochschule (ETH) in Zurich, Switzerland formulated a classical theory of linear fields of arbitrary spin so designed that the fields would be quantizable by canonical methods. Remarkably, their canonical theory for a field of spin two is identical to general relativity with nonlinear effects removed, and the plane waves of that spin-two theory are identical to the waves described above. When quantized by canonical techniques, these waves are carried by zero-rest-mass, spin-two gravitons.

One can see by the following simple argument that the gravitons which carry gravitational waves must have spin two: Consider any plane-wave field (neutrino, electromagnetic,

gravitational, ...) that propagates at the speed of light in the z-direction of a (local) Lorentz frame. At any moment of time examine any physical manifestation of that field, e.g., the acceleration field it produces on test particles. Rotate that manifestation of the field around the z axis, and ask what is the minimum angle of rotation required to bring the field back to its original configuration. Call that minimum angle, $\theta_{\rm ret}$, the waves' *return angle*. The spin S of the particles that carry the wave will necessarily be related to that return angle by

$$\boxed{S = \frac{360 \text{ degrees}}{\theta_{\rm ret}} \ .} \tag{26.52}$$

This simple formula corresponds to the elegant mathematical statement that "the waves generate an irreducible representation of order $S = 360\,\text{degrees}/\theta_{\rm ret}$ of that subgroup of the Lorentz group which leaves their propagation vector unchanged (the 'Little group' of the rotation vector)." For electromagnetic waves a physical manifestation is the electric field, which is described by a vector lying in the x–y plane; if one rotates that vector about the z-axis (propagation axis), it returns to its original orientation after a return angle $\theta_{\rm ret} = 360$ degrees. Correspondingly, the spin of the particle which carries the electromagnetic wave (the photon) is one. For neutrinos the return angle is $\theta_{\rm ret} = 720$ degrees; and correspondingly the spin of a neutrino is $\frac{1}{2}$. For gravitational waves the physical manifestations include the deformation of a sphere of test particles [Figs. 26.1(a,b) and 26.2(a,b)] and the acceleration fields [Figs. 26.1(c,d) and 26.2(c,d)]. Both the deformed, ellipsoidal spheres and the quadrupolar lines of force return to their original orientations after rotation through $\theta_{\rm ret} = 180$ degrees; and correspondingly, the graviton must have spin two.

Although Fierz and Pauli (1939) showed us how to quantize linearized general relativity, the quantization of full, nonlinear general relativity remains a difficult subject of current research, to which we shall return briefly in the next chapter.

26.3.4 Energy and Momentum in Gravitational Waves

In 1968 Richard Isaacson discovered a beautiful and powerful method to define a stress-energy tensor for a gravitational wave. This method is similar to the one by which we analyzed the back-action of a plasma wave on the plasma's background particle distribution [Eq. (22.4)]. Here, as there, we take our exact dynamical equation (the Einstein field equation here, the Vlasov equation there) and expand it to quadratic order in the wave:

$$G_{\alpha\beta} = G^{\rm B}_{\alpha\beta} + G^{(1)}_{\alpha\beta} + G^{(2)}_{\alpha\beta} = 0 \ . \tag{26.53}$$

In this equation $G_{\alpha\beta}$ is the Einstein tensor for the full spacetime metric $g_{\mu\nu} = g^{\rm B}_{\mu\nu} + h_{\mu\nu}$, $G^{\rm B}_{\alpha\beta}$ is the Einstein tensor for the background metric $g^{\rm B}_{\alpha\beta}$, $G^{(1)}_{\alpha\beta}$ is the part linear in $h_{\mu\nu}$, and $G^{(2)}_{\alpha\beta}$ is the part quadratic in $h_{\mu\nu}$. This is the analog of the quadratically expanded Vlasov equation (22.3). Here, as in the plasma case, we next split our dynamical equation into two parts, its spatial average (which is smooth on the scale $\lambda\!\!\!^{-}$) and its remaining, fluctuating piece. In the plasma case the fluctuating piece is the linear wave equation for the plasma waves; in the gravitational case it is a variant of the gravitational wave equation $R^{\rm GW}_{\alpha\beta\gamma\delta|\mu}{}^{\mu} = 0$. In the plasma case the averaged piece is Eq. (22.4) by which the waves at quadratic order in their

amplitude act back on the unperturbed particle distribution. In the gravitational case, it is the equation

$$G^{\rm B}_{\alpha\beta} = -\langle G^{(2)}_{\alpha\beta}\rangle \ , \tag{26.54}$$

by which the waves at quadratic order produce background spacetime curvature.

Equation (26.54) can be brought into the standard form for Einstein's equation in the background spacetime,

$$G^{\rm B}_{\alpha\beta} = 8\pi T^{\rm GW}_{\alpha\beta} \ , \tag{26.55}$$

by attributing to the waves a stress-energy tensor defined by

$$T^{\rm GW}_{\alpha\beta} = -\frac{1}{8\pi}\langle G^{(2)}_{\alpha\beta}\rangle \ . \tag{26.56}$$

Because this stress-energy tensor involves an average over a few wavelengths, its energy density, momentum density, energy flux, and momentum flux are *not defined* on lengthscales shorter than a wavelength. One cannot say how much energy or momentum resides in the troughs of the waves and how much in the crests. One can only say how much total energy there is in a region containing a few wavelengths. However, once one has reconciled oneself to this amount of nonlocality, one finds that $T^{\rm GW}_{\alpha\beta}$ has all the other properties that one expects of any good stress-energy tensor. Most especially, in the absence of coupling of the waves to matter (the situation we are treating), it obeys the standard conservation law

$$\boxed{T^{{\rm GW}\,\alpha\beta}{}_{|\beta} = 0 \ .} \tag{26.57}$$

This law is a direct consequence of the averaged field equation (26.56) and the contracted Bianchi identity for the background spacetime $G^{{\rm B}\,\alpha\beta}{}_{|\beta} = 0$.

By grinding out the second-order perturbation of the Einstein tensor and inserting it into Eq. (26.56), performing several integrations by parts in the average $\langle \ldots \rangle$, and invoking results to be derived in Sec. 26.3.7, one arrives at the following simple expression for $T^{\rm GW}_{\alpha\beta}$ in terms of the wave fields h_+ and $h_\times$:

$$\boxed{T^{\rm GW}_{\alpha\beta} = \frac{1}{16\pi}\langle h_{+,\alpha}h_{+,\beta} + h_{\times,\alpha}h_{\times,\beta}\rangle \ .} \tag{26.58}$$

[For details of the derivation, see Isaacson (1968) or Secs. 35.13 and 35.15 of MTW.]

Let us examine this stress-energy tensor in the background spacetime's local Lorentz frame, which we used above when exploring the properties of gravitational waves. Because $h_+ = h_+(t-z)$ and $h_\times = h_\times(t-z)$, the only nonzero components of Eq. (26.58) are

$$\boxed{T^{{\rm GW}\,00} = T^{{\rm GW}\,0z} = T^{{\rm GW}\,z0} = T^{{\rm GW}\,zz} = \frac{1}{16\pi}\langle \dot{h}_+^2 + \dot{h}_\times^2\rangle \ .} \tag{26.59}$$

This has the same form as the stress-energy tensor for a plane electromagnetic wave propagating in the z direction, and the same form as the stress-energy tensor for any collection of zero-rest-mass particles moving in the z-direction [cf. Eq. (2.30c)], as it must since the

gravitational waves are carried by zero-rest-mass gravitons just as electromagnetic waves are carried by zero-rest-mass photons.

Suppose that the waves have frequency $\sim f$ and that the amplitudes of oscillation of h_+ and $h_\times$ are $\sim h_{\rm amp}$ Then by inserting factors of G and c into Eq. (26.59) [i.e., by switching from geometrized units to conventional units] and by setting $\langle(\partial h_+/\partial t)^2\rangle \simeq 1/2(2\pi f h_{\rm amp})^2$ and similarly for $h_\times$, we obtain the following approximate expression for the energy flux in the waves:

$$T^{{\rm GW}\,0z} \simeq \frac{\pi}{4}\frac{c^3}{G^2} f^2 h_{\rm amp}^2 \simeq 300 \frac{\rm ergs}{\rm cm^2\ sec}\left(\frac{f}{1\ {\rm kHz}}\right)^2\left(\frac{h_{\rm amp}}{10^{-21}}\right)^2 . \tag{26.60}$$

The numbers in this equation correspond to a strongly emitting, highly asymmetric supernova in the Virgo cluster of galaxies. Contrast this huge gravity-wave energy flux with the peak electromagnetic flux at the height of the supernova, $\sim 10^{-9}$ erg cm^{-2} sec^{-1}; but note that the gravity waves should last for only a few milliseconds, while the strong electromagnetic output lasts for weeks.

Corresponding to the huge energy flux (26.60) in an astrophysically interesting gravitational wave is a huge *mean occupation number* for the quantum states of the gravitational-wave field, i.e., a huge value for the number of spin-2, zero-rest-mass gravitons in each quantum state. To compute that occupation number, we shall evaluate the volume in phase space occupied by the waves from a supernova and then divide by the volume occupied by each quantum state [cf. Sec. 2.3]. At a time when the waves have reached a distance r from the source, they occupy a spherical shell of area $4\pi r^2$ and thickness of order $10\lambda\!\!\!^{-}$, where $\lambda\!\!\!^{-} = 1/(2\pi f)$ is their reduced wavelength, so their volume in physical space is $\mathcal{V}_x \sim 100 r^2 \lambda\!\!\!^{-}$. As seen by observers whom the waves are passing, they come from a solid angle $\Delta\Omega \sim (2\lambda\!\!\!^{-}/r)^2$ centered on the source, and they have a spread of angular frequencies ranging from $\omega \sim \frac{1}{2}c/\lambda\!\!\!^{-}$ to $\omega \sim 2c/\lambda\!\!\!^{-}$. Since each graviton carries an energy $\hbar\omega = \hbar c/\lambda\!\!\!^{-}$ and a momentum $\hbar\omega/c = \hbar/\lambda\!\!\!^{-}$, the volume that they occupy in momentum space is $\mathcal{V}_p \sim (2\hbar/\lambda\!\!\!^{-})^3\Delta\Omega$, i.e., $\mathcal{V}_p \sim 10\hbar^3/(\lambda r^2)$. The gravitons' volume in phase space, then, is

$$\mathcal{V}_x\mathcal{V}_p \sim 1000\hbar^3 \sim 4(2\pi\hbar)^3 . \tag{26.61}$$

Since each quantum state for a zero rest-mass particle occupies a volume $(2\pi\hbar)^3$ in phase space [Eq. (2.16) with $g_s = 1$], this means that the total number of quantum states occupied by the gravitons is of order unity! Correspondingly, the mean occupation number of each occupied state is of order the total number of gravitons emitted, which (since the total energy radiated in an extremely strong supernova is of order $10^{-2}M_\odot c^2 \sim 10^{52}$ ergs, and each graviton carries an energy $\hbar c/\lambda\!\!\!^{-} \sim 10^{-23}$ erg), is

$$\bar\eta \sim 10^{75} . \tag{26.62}$$

This enormous occupation number means that the waves behave exceedingly classically; quantum-mechanical corrections to the classical theory have fractional magnitude $1/\sqrt{\bar\eta} \sim 10^{-37}$.

26.3.5 Wave propagation in a source's local asymptotic rest frame

Consider a source of gravitational waves somewhere far out in the universe. In the vicinity of the source but some wavelengths away from it (so the waves are well defined), introduce

a local Lorentz reference frame in which the source is at rest: the source's *local asymptotic rest frame*. In that frame construct spherical polar coordinates (t, r, θ, ϕ) centered on the source so the background metric is

$$ds^2 = -dt^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2) \ . \tag{26.63}$$

The wave gravitational wave equation $R_{\alpha\beta\gamma\delta|\mu}{}^{\mu} = 0$ can be solved fairly easily in this coordinate system. The solution has the form that one would expect from experience with scalar waves and electromagnetic waves in spherical coordinates, plus the description of plane gravitational waves in Sec. 26.3.2: The waves propagate radially with the speed of light, so their wave fields h_+ and $h_\times$ are rapidly varying functions of retarded time

$$\tau_r \equiv t - r \ , \tag{26.64}$$

and slowly varying functions of angle (θ, ϕ), and they die out as $1/r$:

$$h_+ = \frac{Q_+(\tau_r; \theta, \phi)}{r} \ , \quad h_\times = \frac{Q_\times(\tau_r; \theta, \phi)}{r} \ . \tag{26.65}$$

These propagation equations can be thought of as saying that Q_+ and $Q_\times$ are constant along radial null rays, i.e. curves of constant $\tau_r = t - r$, θ and ϕ; and h_+ and $h_\times$ are equal to these constantly-propagated Q's, modified by a $1/r$ falloff.

Notice that the null tangent vector to the radial rays is

$$\vec{k} = \vec{e}_t + \vec{e}_r = -\boldsymbol{\nabla}\tau_r \ . \tag{26.66}$$

By a computation in the coordinate basis, one can show that the radius factor r, which appears in the $1/r$ falloff law, evolves along the rays in accord with the equation

$$\nabla_{\vec{k}} r = r_{,\alpha} k^\alpha = \frac{1}{2}(\vec{\nabla} \cdot \vec{k}) r \ , \tag{26.67}$$

This may look like a complicated way to describe r, but its virtue is that, when the waves have left the source's vicinity and are traveling through the real, lumpy universe, the wave fields will continue to have the form (26.65), with r evolving in accord with (26.67)! [Sec. 26.3.6 below].

The wave fields are not fully meaningful until we have specified their associated polarization tensors. Those tensors can be defined along each ray using two transverse, orthonormal polarization vectors, $\vec{a}$ [the analog of $\vec{e}_x$ in Eq. (26.42)] and $\vec{b}$ [the analog of $\vec{e}_y$]:

$$\mathbf{e}^+ = (\vec{a} \otimes \vec{a} - \vec{b} \otimes \vec{b}) \ , \quad \mathbf{e}^\times = (\vec{a} \otimes \vec{b} + \vec{b} \otimes \vec{a}) \ . \tag{26.68}$$

The vectors $\vec{a}$ and $\vec{b}$ must be held constant along each ray, or equivalently must be parallel transported along the rays:

$$\nabla_{\vec{k}} \vec{a} = \nabla_{\vec{k}} \vec{b} = 0 \ . \tag{26.69}$$

It is conventional, in the source's local asymptotic rest frame, to choose

$$\vec{a} = \vec{e}_{\hat{\theta}} \ , \quad \vec{b} = \vec{e}_{\hat{\phi}} \ , \tag{26.70}$$

so the axes for the + polarization are in the θ and ϕ directions, and those for the $\times$ polarization are rotated 45 degrees to $\vec{e}_{\hat{\theta}}$ and $\vec{e}_{\hat{\phi}}$.

Once the polarization tensors have been constructed, and the wave fields (26.65) are known, then the waves' TT gravitational-wave field can be computed from the standard equation

$$h^{\rm TT}_{\alpha\beta} = h_+ e^+_{\alpha\beta} + h_\times e^\times_{\alpha\beta} \tag{26.71}$$

[Eq. (26.43)], and the waves' Riemann tensor can be computed from the obvious generalization of Eq. (26.44):

$$R^{\rm GW}_{\alpha\beta\gamma\delta} \simeq \frac{1}{2} h^{\rm TT}_{\{\alpha\beta|\gamma\delta\}} = \frac{1}{2}\frac{\partial^2 h^{\rm TT}_{\{\alpha\beta}}{\partial\tau_r^2}\frac{\partial\tau_r}{\partial x^\gamma}\frac{\partial\tau_r}{\partial x^{\delta\}}} = \frac{1}{2}\frac{\partial^2 h^{\rm TT}_{\{\alpha\beta}}{\partial\tau_r^2} k_\gamma k_{\delta\}} \ . \tag{26.72}$$

Here the derivative with respect to retarded time τ_r is taken holding (θ, ϕ, r) fixed, and the $\{\ldots\}$ on the indices has the meaning of Eq. (26.28). As an important special case, if the the basis vectors are chosen in the θ and ϕ directions [Eq. (26.70)], then the tide-producing space-time-space-time part of Riemann [Eq. (26.72)] takes form

$$R^{\rm GW}_{\hat{\theta}\hat{0}\hat{\theta}\hat{0}} = -R^{\rm GW}_{\hat{\phi}\hat{0}\hat{\phi}\hat{0}} = \frac{1}{2} h_{+,\tau_r\tau_r} \ , \quad R^{\rm GW}_{\hat{\theta}\hat{0}\hat{\phi}\hat{0}} = R^{\rm GW}_{\hat{\phi}\hat{0}\hat{\theta}\hat{0}} = \frac{1}{2} h_{\times,\tau_r\tau_r} \ , \tag{26.73}$$

which is the obvious generalization of Eq. (26.41) to radially propagating waves.

We shall demonstrate at the end of the next section that the Riemann tensor (26.72) constructed by the above procedure is, indeed, a solution of the gravitational wave equation.

26.3.6 Wave propagation via geometric optics

The two-lengthscale conditions (26.23), which underlie the definition of gravitational waves, permit us to solve the gravitational wave equation $R^{\rm GW}_{\alpha\beta\gamma\delta|\mu}{}^{\mu} = 0$ by means of geometric optics.

We developed the concepts of geometric optics for rather general types of waves in Sec. 6.3. When those techniques are applied to the gravitational wave equation, they reveal that, as the waves travel through our lumpy, bumpy universe, with its stars, galaxies, and black holes, they continue to propagate along null geodesics, just as they did in the local asymptotic rest frame where they originated. Classically the null geodesics are the waves' rays; quantum mechanically they are the world lines of the waves' gravitons. Because electromagnetic waves also propagate along null-geodesic rays (photon world lines), *gravitational waves must exhibit all the same null-ray-induced phenomena as electromagnetic waves: doppler shifts, cosmological redshifts, gravitational redshifts, gravitational deflection of rays, and gravitational lensing.*

Each ray starts out traveling radially through the local asymptotic rest frame, so it can be identified by three parameters: the direction (θ, ϕ) in which it was emitted, and the retarded time τ_r of its emission. The rays carry these three labels out through spacetime with themselves, and in particular they lay down the scalar field $\tau_r(\mathcal{P})$. As in the source's local asymptotic rest frame, so also throughout spacetime, the vector

$$\boxed{\vec{k} \equiv -\vec{\nabla}\tau_r} \tag{26.74}$$

continues to be tangent to the null rays (so $\vec{k}\cdot\vec{k}=0$), and continues to satisfy the null geodesic equation

$$\boxed{\nabla_{\vec{k}}\vec{k}=0\ ,} \tag{26.75}$$

as one can see by the following index manipulations:

$$k_{\alpha|\mu}k^\mu=-\tau_{r|\alpha\mu}k^\mu=-\tau_{r|\mu\alpha}k^\mu=k_{\mu|\alpha}k^\mu=\frac{1}{2}(\vec{k}\cdot\vec{k})_{|\alpha}=0\ . \tag{26.76}$$

[Here the second expression follows from the definition (26.74) of $\vec{k}$, the third follows from the fact that double gradients of scalars (by contrast with vectors) commute, the fourth follows from (26.74) again, the fifth from the rule for differentiating products (and the fact that the gradient of the metric vanishes), and the sixth from the fact that $\vec{k}\cdot\vec{k}=0$.] Thus, $\vec{k}\equiv-\vec{\nabla}\tau_r$ continues to be the null-geodesic tangent vector.

As in the source's local asymptotic rest frame, so throughout spacetime: (i) the Q functions Q_+ and $Q_\times$ continue to be constant along each ray, (ii) the radius function r continues to evolve via the propagation law

$$\boxed{\nabla_{\vec{k}}r=r_{,\alpha}k^\alpha=\frac{1}{2}(\vec{\nabla}\cdot\vec{k})r\ ,} \tag{26.77}$$

[Eq. (26.67)], (iii) the polarization vectors continue to be parallel transported along the rays

$$\boxed{\nabla_{\vec{k}}\vec{a}=\nabla_{\vec{k}}\vec{b}=0\ ,} \tag{26.78}$$

[Eq. (26.69)], and continue to be used to build the polarization tensors via

$$\boxed{\mathbf{e}^+=(\vec{a}\otimes\vec{a}-\vec{b}\otimes\vec{b})\ ,\quad \mathbf{e}^\times=(\vec{a}\otimes\vec{b}+\vec{b}\otimes\vec{a})\ .} \tag{26.79}$$

[Eq. (26.68)], (iv) the gravitational-wave fields continue to be constructed via the equations

$$\boxed{h_+=\frac{Q_+(\tau_r;\theta,\phi)}{r}\ ,\quad h_\times=\frac{Q_\times(\tau_r;\theta,\phi)}{r}\ ,} \tag{26.80}$$

$$\boxed{h^{\rm TT}_{\alpha\beta}=h_+e^+_{\alpha\beta}+h_\times e^\times_{\alpha\beta}} \tag{26.81}$$

[Eqs. (26.65) and (26.71)] and the Riemann tensor continues to be constructed via

$$\boxed{R^{\rm GW}_{\alpha\beta\gamma\delta}=\frac{1}{2}\frac{\partial^2 h^{\rm TT}_{\{\alpha\beta}}{\partial\tau_r^2}k_\gamma k_{\delta\}}=\frac{1}{2}\frac{\partial^2 h^{\rm TT}_{\{\alpha\beta}}{\partial\tau_r^2}\frac{\partial\tau_r}{\partial x^\gamma}\frac{\partial\tau_r}{\partial x^{\delta\}}}\simeq\frac{1}{2}h^{\rm TT}_{\{\alpha\beta|\gamma\delta\}}} \tag{26.82}$$

[Eq. (26.72)].

We shall now sketch a proof that this geometric-optics Riemann tensor does, indeed, satisfy the gravitational wave equation. The foundation for the proof is the geometric-optics condition that h_+, $h_\times$, and thence $R^{\rm GW}_{\alpha\beta\gamma\delta}$ are rapidly varying functions of retarded time τ_r

and slowly varying functions of (θ, ϕ, r). To take advantage of this, we shall use a prime to denote derivatives at fixed τ_r, so

$$R^{\rm GW}_{\alpha\beta\gamma\delta|\mu} = R^{\rm GW}_{\alpha\beta\gamma\delta,\tau_r}\frac{\partial\tau_r}{\partial x_\mu} + R^{\rm GW}_{\alpha\beta\gamma\delta|\mu'} = -R^{\rm GW}_{\alpha\beta\gamma\delta,\tau_r}k_\mu + R^{\rm GW}_{\alpha\beta\gamma\delta|\mu'} \ . \tag{26.83}$$

Taking the divergence of this on the μ index we obtain

$$R^{\rm GW}_{\alpha\beta\gamma\delta|\mu}{}^{\mu} = R^{\rm GW}_{\alpha\beta\gamma\delta,\tau_r\tau_r}k_\mu k^\mu - 2R^{\rm GW}_{\alpha\beta\gamma\delta,\tau_r|\mu'}k^\mu - R^{\rm GW}_{\alpha\beta\gamma\delta,\tau_r}k_\mu{}^{|\mu} + R^{\rm GW}_{\alpha\beta\gamma\delta|\mu'}{}^{\mu'} \ . \tag{26.84}$$

In the limit as the waves' wavelength goes to zero, the first term á priori scales as $1/\lambda^2$, the second and third as $1/\lambda$, and the fourth as $1/\lambda^0 = 1$. In the spirit of geometric optics (Sec. 6.3), we neglect the tiny fourth term. The leading-order, first term vanishes because $\vec{k}$ is null, so Eq. (26.84) reduces to

$$R^{\rm GW}_{\alpha\beta\gamma\delta|\mu}{}^{\mu} = -2R^{\rm GW}_{\alpha\beta\gamma\delta,\tau_r|\mu'}k^\mu - R^{\rm GW}_{\alpha\beta\gamma\delta,\tau_r}k_\mu{}^{|\mu} \tag{26.85}$$

The first term on the right hand side is the directional derivative of $-2R^{\rm GW}_{\alpha\beta\gamma\delta,\tau_r}$ along $\vec{k}$, i.e. along a ray. Since each ray has constant (θ, ϕ, τ_e), and since the vectors $\vec{a}$, $\vec{b}$, and $\vec{k}$ that appear in Eqs. (26.80)–(26.82) for $R^{\rm GW}_{\alpha\beta\gamma\delta}$ all are parallel transported along $\vec{k}$, the only piece of $R^{\rm GW}_{\alpha\beta\gamma\delta}$ that can vary along $\vec{k}$ is the factor $1/r$. Correspondingly,

$$R^{\rm GW}_{\alpha\beta\gamma\delta,\tau_r|\mu'}k^\mu = R^{\rm GW}_{\alpha\beta\gamma\delta,\tau_r}r\nabla_{\vec{k}}\left(\frac{1}{r}\right) \ . \tag{26.86}$$

Inserting this into Eq. (26.85) and invoking the propagation law (26.77) for r, we obtain

$$R^{\rm GW}_{\alpha\beta\gamma\delta|\mu}{}^{\mu} = 0 \ . \tag{26.87}$$

Thus, our geometric-optics formulae for $R^{\rm GW}_{\alpha\beta\gamma\delta}$ do, indeed, produce a solution to the gravitational wave equation. Moreover, since, in the source's local asymptotic rest frame, this solution reduces to the one developed in Sec. 26.3.5, the formulae for $R^{\rm GW}_{\alpha\beta\gamma\delta}$ also satisfy the gravitational wave equation.

26.3.7 Metric perturbation; TT gauge

Although the properties of gravitational waves and their propagation are best described in terms of the waves' Riemann tensor $R^{\rm GW}_{\alpha\beta\gamma\delta}$, their generation is best described in terms of the waves' metric perturbation $h_{\mu\nu}$ [cf. the linearized-theory analysis in Sec. 24.9.2]. As in linearized theory, so also here, there is gauge freedom in the waves' metric perturbation, which results from introducing a tiny, rippled displacement $\vec{\xi}$ of the coordinate lines. In a local Lorentz frame of the smooth background, the gauge change has the linearized-theory form $\delta h_{\mu\nu} = -\xi_{\mu,\nu} - \xi_{\nu,\mu}$ [Eq. (24.104)], so in an arbitrary coordinate system of the background spacetime it must be

$$\boxed{h^{\rm new}_{\mu\nu} = h^{\rm old}_{\mu\nu} - \xi_{\mu|\nu} - \xi_{\nu|\mu} \ .} \tag{26.88}$$

By choosing the background coordinates to be local Lorentz and carefully adjusting the waves' gauge, we can ensure that the waves' metric perturbation is equal to the transverse-traceless gravitational-wave field (26.43), which we originally defined in terms of the Riemann tensor; i.e., we can ensure that

$$h_{\alpha\beta} = h^{\rm TT}_{\alpha\beta} \ , \quad \text{or equivalently,}$$
$$h_{xx} = -h_{yy} = h_+(t-z) \ , \quad h_{xy} = h_{yx} = h_\times(t-z) \ , \quad \text{all other } h_{\alpha\beta} \text{ vanish.} \tag{26.89}$$

To see that this is possible, we need only verify that this metric perturbation produces the correct components of Riemann, Eq. (26.41); indeed it does, as we can see by inserting Eqs. (26.89) into expression (26.26) for the Riemann tensor. For an alternative proof see Ex. 26.6. The gauge in which the waves' metric perturbation takes the simple TT form (26.89) is called *transverse-traceless gauge*, or TT gauge, and the coordinates in which the metric perturbation takes this form are called *TT coordinates*.

TT gauge is not the only one in which the waves' metric perturbation has the plane-wave form $h_{\alpha\beta} = h_{\alpha\beta}(t-z)$ There are many other such gauges; cf. Ex. 26.6. In any local Lorentz frame of the background spacetime and any gauge for which $h_{\alpha\beta} = h_{\alpha\beta}(t-z)$, the transverse components of the waves' Riemann curvature tensor take the form [derivable from Eq. (24.51) or (24.96)]

$$R^{\rm GW}_{j0k0} = -\frac{1}{2}\frac{\partial^2 h_{jk}}{\partial t^2} \text{ for } j = x\,, y \text{ and } k = x\,, y \ . \tag{26.90}$$

By comparing with Eq. (26.41) we see that in such a gauge, the transverse part of the waves' metric perturbation must be equal to the TT gravitational-wave field:

$$h^{\rm TT}_{jk} = (h_{jk})^{\rm T} \ . \tag{26.91}$$

Here the T on the right-hand side means "throw away all components except those which are spatial and are transverse to the waves' propagation direction". Since $h^{\rm TT}_{jk}$ is trace-free as well as transverse, we are guaranteed that the transverse part of the metric perturbation h_{jk} will be trace-free; i.e. $h_{xx} + h_{yy} = 1$. To emphasize this trace-free property it is conventional to write Eq. (26.91) in the form

$$h^{\rm TT}_{jk} = (h_{jk})^{\rm TT} \ , \tag{26.92}$$

where the second T on the right-hand side means "remove the trace, if the trace is not already zero". To repeat, *Eq. (26.92) is true in any gauge where the waves' contribution to the metric has the "speed-of-light-propagation" form* $h_{\alpha\beta} = h_{\alpha\beta}(t-z)$.

If we rotate the spatial axes so the waves propagate along the unit spatial vector $\mathbf{n}$ instead of along $\vec{e}_z$, then the "speed-of-light-propagation" form of the metric becomes

$$h_{\alpha\beta} = h_{\alpha\beta}(t - \mathbf{n}\cdot\mathbf{x}) \ , \tag{26.93}$$

and the extraction of the spatial, transverse-traceless part of this metric perturbation can be achieved with the aid of the projection tensor

$$\boxed{P^{jk} \equiv \delta^{jk} - n^j\, n^k \ .} \tag{26.94}$$

Specifically,

$$\boxed{h^{\rm TT}_{jk} = (h_{jk})^{\rm TT} = P_j{}^l P_k{}^m h_{lm} - \frac{1}{2} P_{jk} P^{lm} h_{lm} \ .} \tag{26.95}$$

Here the notation is that of Cartesian coordinates with $P_j{}^k = P^{jk} = P_{jk}$.

When analyzing gravitational wave generation, the quantity most easily computed is often the trace-reversed metric perturbation, in a gauge with speed-of-light propagation, so $\bar{h}_{\alpha\beta} = \bar{h}_{\alpha\beta}(t - \mathbf{n}\cdot\mathbf{x})$. Because the projection process (26.95) removes the trace (i.e., the result is insensitive to the trace), and $\bar{h}_{jk}$ and h_{jk} differ only in their trace, we can compute the gravitational-wave field by direct TT projection of $\bar{h}_{jk}$ without bothering to evaluate h_{jk} first:

$$\boxed{h^{\rm TT}_{jk} = (\bar{h}_{jk})^{\rm TT} = P_j{}^l P_k{}^m \bar{h}_{lm} - \frac{1}{2} P_{jk} P^{lm} \bar{h}_{lm} \ .} \tag{26.96}$$

****

EXERCISES

Exercise 26.4 *Derivation: Behavior of h_+ and $h_\times$ under rotations and boosts*

(a) Derive the behavior (26.51) of h_+ and $h_\times$ under rotations in the transverse plane. [Hint: show that $\mathbf{e}^+ + i\mathbf{e}^\times$ rotates through 2ψ, and then write $h^{\rm GW}_{\alpha\beta}$ [Eq. (26.43)] in terms of $h_+ + ih_\times$ and $\mathbf{e}^+ - i\mathbf{e}^\times$.]

(b) Show that, with the orientations of spatial basis vectors described after Eq. (26.51), h_+ and $h_\times$ are unchanged by boosts.

Exercise 26.5 *Problem: Energy-Momentum Conservation in Geometric Optics Limit*
Near the end of Sec. 26.3.6, we proved that our geometric-optics formulae for $R^{\rm GW}_{\alpha\beta\gamma\delta}$ satisfy the gravitational wave equation. Use these same techniques to show that the gravitational stress-energy tensor (26.58), with h_+ and $h_\times$ given by the geometric-optics formulae (26.80), has vanishing divergence, $\vec{\nabla}\cdot\mathbf{T}^{\rm GW} = 0$.

Exercise 26.6 *Example: Transformation to TT Gauge*
Consider a plane gravitational wave propagating in the z-direction through a local Lorentz frame of the smooth background spacetime. Such a wave can be described by Linearized Theory. In Sec. 24.9.2 and Ex. 24.13 we showed that, by a careful choice of the four gauge-generating functions $\xi^\alpha(\mathcal{P})$, one can bring the trace-reversed metric perturbation into Lorentz gauge, so it satisfies the gauge condition $\bar{h}_{\alpha\beta,}{}^\beta = 0$ and the wave equation $\bar{h}_{\alpha\beta,\mu}{}^\mu = 0$, and thence has the form $\bar{h}_{\alpha\beta} = h_{\alpha\beta}(t-z)$. In general there are 10 independent components of $\bar{h}_{\alpha\beta}$, since it is a symmetric second-rank tensor, but the 4 gauge conditions reduce this from 10 to 6. *Thus, in general, the Lorentz-gauge metric perturbation for a plane gravitational wave contains six independent functions of $t-z$.* Only two of these six can represent physical degrees of freedom of the wave; the other four must be pure-gauge functions and must be removable by a further specialization of the gauge. This exercise explores that further gauge freedom.

(a) Consider any trace-reversed metric perturbation that is in Lorentz gauge. Show that a further gauge change whose generators satisfy the wave equation $\xi_{\alpha,\mu}{}^{\mu} = 0$ leaves $\bar{h}_{\alpha\beta}$ still in Lorentz gauge. Show that such a gauge change, in general, involves four free functions of *three* of the spacetime coordinates, by contrast with general gauge transformations which entail four free functions of all *four* spacetime coordinates.

(b) Consider the plane gravitational wave described in the first paragraph of this problem. Exhibit gauge-change generators ξ_α that satisfy the wave equation and that remove four of the six independent functions from $\bar{h}_{\alpha\beta}$, bringing it into TT gauge, so the components of $h_{\alpha\beta}$ are given by Eqs. (26.89).

(c) Show by an explicit calculation that the gauge change of part (b) can be achieved by throwing away all pieces of $h_{\alpha\beta}$ except the transverse ones (those that lie in the x-y plane) and by then removing the trace — i.e. by the transverse-traceless projection of Eq. (26.92).

****

26.4 The Generation of Gravitational Waves

26.4.1 Multipole-moment expansion

The electromagnetic waves emitted by a dynamical charge distribution are usually expressed as a sum over the source's multipole moments. There are two families of moments: the electric moments (moments of the electric charge distribution) and the magnetic moments (moments of the electric current distribution).

Similarly, the gravitational waves emitted by a dynamical distribution of mass-energy and momentum can be expressed as a sum over multipole moments. Again there are two families of moments: the *mass moments* (moments of the mass-energy distribution) and the *current moments* (moments of the mass-current distribution, i.e. the momentum distribution). The multipolar expansion of gravitational waves is presented in great detail in Thorne (1980). In this section we shall sketch and explain its qualitative and order-of-magnitude features.

In the source's weak-gravity near zone (if it has one), the mass moments show up in the time-time part of the metric in a form familiar from Newtonian theory

$$g_{00} = -(1 + 2\Phi) = -1 \& \frac{\mathcal{I}_0}{r} \& \frac{\mathcal{I}_1}{r^2} \& \frac{\mathcal{I}_2}{r^3} \& \dots \ . \tag{26.97}$$

[cf. Eq. (24.95)]. Here r is radius, $\mathcal{I}_\ell$ is the moment of order ℓ, and "&" means "plus terms with the form". The mass monopole moment $\mathcal{I}_0$ is the source's mass, and the mass dipole moment $\mathcal{I}_1$ can be made to vanish by placing the origin of coordinates at the center of mass.

Similarly, in the source's weak-gravity near zone, its current moments $\mathcal{S}_\ell$ show up in the space-time part of the metric

$$g_{0j} = \frac{\mathcal{S}_1}{r^2} \& \frac{\mathcal{S}_2}{r^3} \& \dots \ . \tag{26.98}$$

Just as there is no magnetic monopole moment in classical electromagnetic theory, so there is no current monopole moment in general relativity. The current dipole moment $\mathcal{S}_1$ is the source's angular momentum, so the leading-order term in the expansion (26.98) has the form (24.112), which we have used to deduce the angular momenta of gravitating bodies.

If the source has mass M, size L and internal velocities $\sim v$, then the magnitudes of its moments are

$$\mathcal{I}_\ell \sim ML^\ell \ , \quad \mathcal{S}_\ell \sim MvL^\ell \tag{26.99}$$

These formulae guarantee that the near-zone fields g_{00} and g_{0j}, as given by Eqs. (26.97) and (26.98), are dimensionless.

As the source's moments oscillate dynamically, they produce gravitational waves. Mass-energy conservation [Eq. (24.114)] prevents the mass monopole moment $\mathcal{I}_0 = M$ from oscillating; angular-momentum conservation [Eq. (24.115)] prevents the current dipole moment $\mathcal{S}_1$ = (angular momentum) from oscillating; and because the time derivative of the mass dipole moment $\mathcal{I}_1$ is the source's linear momentum, momentum conservation [Eq. (24.118)] prevents the mass dipole moment from oscillating. Therefore, the lowest-order moments that can contribute to the waves are the quadrupolar ones. The wave fields h_+ and $h_\times$ in the source's local asymptotic rest frame must (i) be dimensionless, (ii) die out as $1/r$, and (iii) be expressed as a sum over derivatives of the multipole moments. These conditions guarantee that the waves will have the following form:

$$\boxed{h_+ \sim h_\times \sim \frac{\partial^2 \mathcal{I}_2/\partial t^2}{r} \& \frac{\partial^3 \mathcal{I}_3/\partial t^3}{r} \& \ldots \& \frac{\partial^2 \mathcal{S}_2/\partial t^2}{r} \& \frac{\partial^3 \mathcal{S}_3/\partial t^3}{r} \& \ldots \ .} \tag{26.100}$$

The timescale on which the moments oscillate is $T \sim L/v$, so each time derivative produces a factor v/L. Correspondingly, the ℓ-pole contributions to the waves have magnitudes

$$\frac{\partial^\ell \mathcal{I}_\ell/\partial t^\ell}{r} \sim \frac{M}{r} v^\ell \ , \quad \frac{\partial^\ell \mathcal{S}_\ell/\partial t^\ell}{r} \sim \frac{M}{r} v^{(\ell+1)} \ . \tag{26.101}$$

This means that, for a "slow-motion source" (one with internal velocities v small compared to light so the reduced wavelength $\lambda\!\!\!^{-} \sim L/v$ is large compared to the source size L), *the mass quadrupole moment $\mathcal{I}_2$ will produce the strongest waves.* The mass octupole waves and current quadrupole waves will be weaker by $\sim v \sim L/\lambda\!\!\!^{-}$; the mass 4-pole and current octupole waves will be weaker by $\sim v^2 \sim L^2/\lambda\!\!\!^{-2}$, etc. This is analogous to the electromagnetic case, where the electric dipole waves are the strongest, the electric quadrupole and magnetic dipole are smaller by $\sim L/\lambda\!\!\!^{-}$, etc.

In the next section we shall develop the theory of mass-quadrupole gravitational waves. For the corresponding theory of higher-order multipoles, see, e.g., Sec. VIII of Thorne (1980).

26.4.2 Quadrupole-moment formalism

Consider a weakly gravitating, nearly Newtonian system, e.g. a binary star system, and write its Newtonian potential in the usual way

$$\Phi(\mathbf{x}) = -\int \frac{\rho(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|} dV_{x'} \ . \tag{26.102}$$

By using Cartesian coordinates, placing the origin of coordinates at the center of mass, and expanding

$$\frac{1}{|\mathbf{x}-\mathbf{x}'|} = \frac{1}{r} + \frac{x^j x^{j'}}{r^3} + \frac{x^j x^k (3x^{j'}x^{k'} - r'^2\delta_{jk})}{2r^5} + \dots \ , \tag{26.103}$$

we obtain the multipolar expansion of the Newtonian potential

$$\boxed{\Phi(\mathbf{x}) = -\frac{M}{r} - \frac{3\mathcal{I}_{jk}x^j x^k}{2r^5} + \dots \ .} \tag{26.104}$$

Here

$$\boxed{M = \int \rho dV_x \ , \quad \mathcal{I}_{jk} = \int \rho \left(x^j x^k - \frac{1}{3}r^2\delta_{jk}\right) dV_x} \tag{26.105}$$

are the system's mass and mass quadrupole moment. Note that *the mass quadrupole moment is equal to the second moment of the mass distribution, with its trace removed.*

As we have discussed, dynamical oscillations of the quadrupole moment produce gravitational waves. Those waves must be describable by an outgoing-wave solution to the Lorentz-gauge, linearized Einstein equations

$$\bar{h}_{\mu\nu,}{}^{\nu} = 0 \ , \quad \bar{h}_{\mu\nu,\alpha}{}^{\alpha} = 0 \tag{26.106}$$

[Eqs. (24.105) and (24.106)] that has the near-zone Newtonian limit

$$\frac{1}{2}(\bar{h}_{00} + \bar{h}_{xx} + \bar{h}_{yy} + \bar{h}_{zz}) = h_{00} = \frac{3\mathcal{I}_{jk}x^j x^k}{r} \tag{26.107}$$

[cf. Eq. (24.101)].

The desired solution can be written in the form

$$\boxed{\bar{h}_{00} = 2\left[\frac{\mathcal{I}_{jk}(t-r)}{r}\right]_{,jk} \ , \quad \bar{h}_{0j} = 2\left[\frac{\dot{\mathcal{I}}_{jk}(t-r)}{r}\right]_{,k} \ , \quad \bar{h}_{jk} = 2\frac{\ddot{\mathcal{I}}_{jk}(t-r)}{r} \ ,} \tag{26.108}$$

where the coordinates are Cartesian, $r \equiv \sqrt{\delta_{jk}x^j x^k}$, and the dots denote time derivatives. To verify that this is the desired solution: (i) Compute its divergence $\bar{h}_{\alpha\beta,}{}^{\beta}$ and obtain zero almost trivially. (ii) Notice that each Lorentz-frame component of $\bar{h}_{\alpha\beta}$ has the form $f(t-r)/r$ aside from some derivatives that commute with the wave operator, which implies that it satisfies the wave equation. (iii) Notice that in the near zone, the slow-motion assumption inherent in the Newtonian limit makes the time derivatives negligible, so $\bar{h}_{jk} \simeq 0$ and $\bar{h}_{00}$ is twice the right-hand side of Eq. (26.107), as desired.

Because the trace-reversed metric perturbation (26.108) in the wave zone has the speed-of-light-propagation form, aside from its very slow decay as $1/r$, we can compute the gravitational-wave field $h^{\rm TT}_{jk}$ from it by transverse-traceless projection, Eq. (26.96) with $\mathbf{n} = \mathbf{e}_r$:

$$\boxed{h^{\rm TT}_{jk} = 2\left[\frac{\ddot{\mathcal{I}}_{jk}(t-r)}{r}\right]^{\rm TT} \ .} \tag{26.109}$$

This is called *the quadrupole-moment formula for gravitational-wave generation*. Our derivation shows that it is valid for any nearly Newtonian source. Looking back more carefully at the derivation, one can see that, in fact, it relied only on the linearized Einstein equations and the Newtonian potential in the source's local asymptotic rest frame. Therefore, this quadrupole formula is also valid for slow-motion sources that have strong internal gravity (e.g., slowly spinning neutron stars), so long as we read the quadrupole moment $\mathcal{I}_{jk}(t-r)$ off the source's near-zone Newtonian potential (26.104) and don't try to compute it via the Newtonian volume integral (26.105).

When the source is nearly Newtonian, so the volume integral (26.105) can be used to compute the quadrupole moment, the computation of the waves is simplified by computing instead the second moment of the mass distribution

$$\boxed{I_{jk} = \int \rho x^j x^k dV_x \ ,} \tag{26.110}$$

which differs from the quadrupole moment solely in its trace. Then, because the TT projection is insensitive to the trace, the wave field (26.109) can be computed as

$$\boxed{h^{\rm TT}_{jk} = 2\left[\frac{\ddot{I}_{jk}(t-r)}{r}\right]^{\rm TT} .} \tag{26.111}$$

To get an order of magnitude feeling for the strength of the gravitational waves, notice that the second time derivative of the quadrupole moment, in order of magnitude, is the nonspherical part of the source's internal kinetic energy, $E^{\rm ns}_{\rm kin}$, so

$$\boxed{h_+ \sim h_\times \sim \frac{E^{\rm ns}_{\rm kin}}{r} = G\frac{E^{\rm ns}_{\rm kin}}{c^4 r} \ ,} \tag{26.112}$$

where the second expression is written in conventional units. Although this estimate is based on the slow-motion assumption of source size small compared to reduced wavelength, $L \ll \lambda\!\!\!^{-}$, it remains valid in order of magnitude when extrapolated into the realm of the strongest of all realistic astrophysical sources, which have $L \sim \lambda\!\!\!^{-}$. For sources in the "high-frequency" band of ground-based detectors (as we shall see below), the largest value of $E^{\rm ns}_{\rm kin}$ that is likely to occur is roughly $E^{\rm ns}_{\rm kin} \sim M_\odot \sim 1$ km, where $M_\odot$ is the mass of the Sun. The collision of two smallish black holes (masses of several solar masses) will have such an $E^{\rm ns}_{\rm kin}$. Such a source at the center of our galaxy would produce $h_+ \sim 10^{-17}$; at the center of the Virgo cluster of galaxies it would produce $h_+ \sim 10^{-20}$, and at the Hubble distance (edge of the observable universe) it would produce $h_+ \sim 10^{-23}$. This sets the sensitivity goals of ground-based detectors, Sec. 26.5.

Because the gravitational stress-energy tensor $T^{\rm GW}_{\mu\nu}$ produces background curvature via the Einstein equation $G^{\rm B}_{\mu\nu} = 8\pi T^{\rm GW}_{\mu\nu}$, just like nongravitational stress-energy tensors, it must contribute to the rate of change of the source's mass M, linear momentum P_j and angular momentum S_j [Eqs. (24.114)–(24.118)] just like other stress-energies. When one inserts the quadrupolar $T^{\rm B}_{\mu\nu}$ into Eqs. (24.114)–(24.118) and integrates over a sphere in the wave zone

of the source's local asymptotic rest frame, one finds that

$$\boxed{\frac{dM}{dt} = -\frac{1}{5}\left\langle \frac{\partial^3 \mathcal{I}_{jk}}{\partial t^3}\frac{\partial^3 \mathcal{I}_{jk}}{\partial t^3}\right\rangle ,} \tag{26.113}$$

$$\boxed{\frac{dS_i}{dt} = -\frac{2}{5}\epsilon_{ijk}\left\langle \frac{\partial^2 \mathcal{I}_{jm}}{\partial t^2}\frac{\partial^3 \mathcal{I}_{km}}{\partial t^3}\right\rangle ,} \tag{26.114}$$

and $dP_j/dt = 0$. It turns out [cf. Sec. IV of Thorne (1980)] that the dominant linear-momentum change (i.e., the dominant radiation-reaction "kick") arises from a beating of the mass quadrupole moment against the mass octupole moment, and mass quadrupole against current quadrupole.

The back reaction of the emitted waves on their source shows up not only in changes of the source's mass, momentum, and angular momentum, but also in accompanying changes of the source's internal structure. These structure changes can be deduced fully, in many cases, from dM/dt, dS_j/dt and dP_j/dt. A nearly Newtonian binary system is an example (Sec. 26.4.3 below). However, in other cases (e.g., a compact body orbiting near the horizon of a black hole), the only way to compute the structure changes is via a *gravitational-radiation-reaction force* that acts back on the system.

The simplest example of such a force is one derived by William Burke (1971) for quadrupole waves emitted by a nearly Newtonian system. Burke's quadrupolar radiation-reaction force can be incorporated into Newtonian gravitation theory by simply augmenting the system's near-zone Newtonian potential by a radiation-reaction term, computed from the fifth time derivative of the system's quadrupole moment:

$$\boxed{\Phi^{\rm react} = \frac{1}{5}\frac{\partial^5 \mathcal{I}_{jk}}{\partial t^5}x^j x^k \ .} \tag{26.115}$$

This potential satisfies the vacuum Newtonian field equation $\nabla^2\Phi \equiv \delta_{jk}\Phi_{,jk} = 0$ because $\mathcal{I}_{jk}$ is trace free.

This augmentation onto the Newtonian potential arises as a result of general relativity's outgoing-wave condition. If one were to switch to an ingoing-wave condition, $\Phi^{\rm react}$ would change sign, and if the system's oscillating quadrupole moment were joined onto standing gravitational waves, $\Phi^{\rm react}$ would go away. In Ex. 26.9, it is shown that the radiation reaction force density $-\rho\boldsymbol{\nabla}\Phi^{\rm react}$ saps energy from the system at the same rate as the gravitational waves carry it away.

Burke's gravitational radiation-reaction potential $\Phi^{\rm react}$ and force density $-\rho\boldsymbol{\nabla}\Phi^{\rm react}$ are close analogs of the radiation reaction potential [last term in Eq. (15.93)] and acceleration [right side of Eq. (15.96)] that act on an oscillating ball which emits sound waves into a surrounding fluid. Moreover, Burke's derivation of his gravitational radiation-reaction potential is conceptually the same as the derivation, in Chap. 15, of the sound-wave reaction potential.

26.4.3 Gravitational waves from a binary star system

A very important application of the quadrupole formalism is to wave emission by a nearly Newtonian binary star system. Denote the stars by indices A and B and their masses by M_A and M_B, so their total and reduced mass are (as usual)

$$\boxed{M = M_A + M_B \ , \quad \mu = \frac{M_A M_B}{M} \ ;} \tag{26.116}$$

and let the binary's orbit be circular, for simplicity, with separation a between the stars' centers of mass. Then Newtonian force balance dictates that the orbital angular velocity Ω is given by Kepler's law,

$$\boxed{\Omega = \sqrt{M/a^3} \ ,} \tag{26.117}$$

and the orbits of the two stars are

$$x_A = \frac{M_B}{M} a \cos\Omega t \ , \quad y_A = \frac{M_B}{M} a \sin\Omega t \ , \quad x_B = -\frac{M_A}{M} a \cos\Omega t \ , \quad y_B = -\frac{M_A}{M} a \sin\Omega t \ . \tag{26.118}$$

The second moment of the mass distribution, Eq. (26.110), is $I_{jk} = M_A x_A^j x_A^k + M_B x_B^j x_B^k$. Inserting the stars' time-dependent positions (26.118), we obtain as the only nonzero components

$$I_{xx} = \mu a^2 \cos^2\Omega t \ , \quad I_{yy} = \mu a^2 \sin^2\Omega t \ , \quad I_{xy} = I_{yx} = \mu a^2 \cos\Omega t \sin\Omega t \ . \tag{26.119}$$

Noting that $\cos^2\Omega t = \frac{1}{2}(1+\cos 2\Omega t)$, $\sin^2\Omega t = \frac{1}{2}(1-\cos 2\Omega t)$ and $\cos\Omega t \sin\Omega t = \frac{1}{2}\sin 2\Omega t$, and evaluating the double time derivative, we obtain

$$\begin{aligned} &\ddot{I}_{xx} = -2\mu(M\Omega)^{2/3}\cos 2\Omega t \ , \quad \ddot{I}_{yy} = 2\mu(M\Omega)^{2/3}\cos 2\Omega t \ , \\ &\ddot{I}_{xy} = \ddot{I}_{yx} = -2\mu(M\Omega)^{2/3}\sin 2\Omega t \ . \end{aligned} \tag{26.120}$$

We express this in terms of Ω rather than a because Ω is a direct gravitational-wave observable: the waves' angular frequency is 2Ω.

To compute the gravitational-wave field (26.109), we must project out the transverse part of this. The projection is best performed in an orthonormal spherical basis, since there the transverse part is just the projection into the plane spanned by $\vec{e}_{\hat{\theta}}$ and $\vec{e}_{\hat{\phi}}$, and the transverse-traceless part just has components

$$(\ddot{I}_{\hat{\theta}\hat{\theta}})^{\rm TT} = -(\ddot{I}_{\hat{\phi}\hat{\phi}})^{\rm TT} = \frac{1}{2}(\ddot{I}_{\hat{\theta}\hat{\theta}} - \ddot{I}_{\hat{\phi}\hat{\phi}}) \ , \quad (\ddot{I}_{\hat{\theta}\hat{\phi}})^{\rm TT} = \ddot{I}_{\hat{\theta}\hat{\phi}} \ . \tag{26.121}$$

Now, a little thought will save us much work: We need only compute these quantities at $\phi = 0$ (i.e., in the x-z plane), since their circular motion guarantees that their dependence on t and ϕ must be solely through the quantity $\Omega t - \phi$. At $\phi = 0$, $\vec{e}_{\hat{\theta}} = \vec{e}_x \cos\theta - \vec{e}_z \sin\theta$ and $\vec{e}_{\hat{\phi}} = \vec{e}_y$, so the only nonzero components of the transformation matrices from the Cartesian basis to the transverse part of the spherical basis are $L^x{}_{\hat{\theta}} = \cos\theta$, $L^z{}_{\hat{\theta}} = -\sin\theta$, $L^y{}_{\hat{\phi}} = 1$. Using this transformation matrix, we obtain, at $\phi = 0$, $\ddot{I}_{\hat{\theta}\hat{\theta}} = \ddot{I}_{xx}\cos^2\theta$, $\ddot{I}_{\hat{\phi}\hat{\phi}} = \ddot{I}_{yy}$, $\ddot{I}_{\hat{\theta}\hat{\phi}} = \ddot{I}_{xy}\cos\theta$.

Inserting these and expressions (26.120) into Eq. (26.121), and setting $\Omega t \to \Omega t - \phi$ to make the formulae valid away from $\phi = 0$, we obtain

$$\begin{aligned}(\ddot{I}_{\hat{\theta}\hat{\theta}})^{\rm TT} &= -(\ddot{I}_{\hat{\phi}\hat{\phi}})^{\rm TT} = -(1+\cos^2\theta)\,\mu(M\Omega)^{2/3}\cos[2(\Omega t-\phi)]\ ,\\ (\ddot{I}_{\hat{\theta}\hat{\phi}})^{\rm TT} &= +(\ddot{I}_{\hat{\phi}\hat{\theta}})^{\rm TT} = -2\cos\theta\,\mu(M\Omega)^{2/3}\sin[2(\Omega t-\phi)]\ .\end{aligned} \tag{26.122}$$

The gravitational-wave field (26.109) is $2/r$ times this quantity evaluated at the retarded time $t-r$.

We shall make the conventional choice for the polarization tensors:

$$\boxed{\mathbf{e}^{+} = (\vec{e}_{\hat{\theta}}\otimes\vec{e}_{\hat{\theta}} - \vec{e}_{\hat{\phi}}\otimes\vec{e}_{\hat{\phi}})\ ,\quad \mathbf{e}^{\times} = (\vec{e}_{\hat{\theta}}\otimes\vec{e}_{\hat{\phi}} + \vec{e}_{\hat{\phi}}\otimes\vec{e}_{\hat{\theta}})\ .} \tag{26.123}$$

Then the two scalar gravitational-wave fields are

$$\boxed{h_{+} = h^{\rm TT}_{\hat{\theta}\hat{\theta}} = \frac{2}{r}[\ddot{I}_{\hat{\theta}\hat{\theta}}(t-r)]^{\rm TT} = -2(1+\cos^2\theta)\,\frac{\mu(M\Omega)^{2/3}}{r}\cos[2(\Omega t-\Omega r-\phi)]\ ,} \tag{26.124a}$$

$$\boxed{h_{\times} = h^{\rm TT}_{\hat{\theta}\hat{\phi}} = \frac{2}{r}[\ddot{I}_{\hat{\theta}\hat{\phi}}(t-r)]^{\rm TT} = -4\cos\theta\,\frac{\mu(M\Omega)^{2/3}}{r}\sin[2(\Omega t-\Omega r-\phi)]\ .} \tag{26.124b}$$

We have expressed the ampitudes of these waves in terms of the dimensionless quantity $(M\Omega)^{2/3} = M/a = v^2$, where v is the relative velocity of the two stars.

Notice that, as viewed from the polar axis $\theta = 0$, h_+ and $h_\times$ are identical except for a $\pi/2$ phase delay, which means that the net stretch-squeeze ellipse (the combination of those in Figs. 26.1 and 26.2) rotates with angular velocity Ω. This is the gravitational-wave variant of circular polarization and arises because the binary motion as viewed from the polar axis looks circular. By contrast, as viewed by an observer in the equatorial plane $\theta = \pi/2$, $h_\times$ vanishes, so the net stretch-squeeze ellipse just oscillates along the + axes and the waves have linear polarization. This is natural, since the orbital motion as viewed by an equatorial observer is just a linear, horizontal, back-and-forth oscillation. Notice also that *the gravitational-wave frequency is twice the orbital frequency*, i.e.

$$\boxed{f = 2\frac{\Omega}{2\pi} = \frac{\Omega}{\pi}\ .} \tag{26.125}$$

To compute, via Eqs. (26.113) and (26.114), the rate at which energy and angular momentum are lost from the binary, we need to know the double and triple time derivatives of its quadrupole moment $\mathcal{I}_{jk}$. The double time derivative is just $\ddot{I}_{jk}$ with its trace removed, but Eq. (26.119) shows that $\ddot{I}_{jk}$ is already trace free so $\ddot{\mathcal{I}}_{jk} = \ddot{I}_{jk}$. Inserting Eq. (26.119) for this quantity into Eqs. (26.113) and (26.114) and performing the average over a gravitational-wave period, we find that

$$\boxed{\frac{dM}{dt} = -\frac{32}{\pi}\frac{\mu^2}{M^2}(M\Omega)^{10/3}\ ,\quad \frac{dS_z}{dt} = -\frac{1}{\Omega}\frac{dM}{dt}\ ,\quad \frac{dS_x}{dt} = \frac{dS_y}{dt} = 0\ .} \tag{26.126}$$

This loss of energy and angular momentum causes the binary to spiral inward, decreasing the stars' separation a and increasing the orbital angular velocity Ω. By comparing Eqs. (26.126) with the standard equations for the binary's orbital energy and angular momentum, $M - (\text{sum of rest masses of stars}) = E = -\frac{1}{2}\mu M/a = -\frac{1}{2}\mu(M\Omega)^{2/3}$, and $S_z = \mu a^2\Omega = \mu(M\Omega)^{2/3}/\Omega$, we obtain an equation for $d\Omega/dt$ which we can integrate to give

$$\boxed{\Omega = \pi f = \left(\frac{5}{256}\frac{1}{\mu M^{2/3}}\frac{1}{t_o - t}\right)^{3/8}} \quad . \tag{26.127}$$

Here t_o (an integration constant) is the time remaining until the two stars merge, if the stars are thought of as point masses so their surfaces do not collide sooner. This equation can be inverted to read off the time until merger as a function of gravitational-wave frequency.

These results for a binary's waves and radiation-reaction-induced inspiral are of great importance for gravitational-wave detection; see, e.g., Cutler and Thorne (2002).

As the stars spiral inward, $(M\Omega)^{2/3} = M/a = v^2$ grows larger, h_+ and $h_\times$ grow larger, and relavistic corrections to our Newtonian, quadrupole analysis grow larger. Those relativistic corrections (including current-quadrupole waves, mass-octupole waves, etc.) can be computed using a *post-Newtonian* expansion of the Einstein field equations, i.e. an expansion in $M/a \sim v^2$. The expected accuracies of the LIGO/VIRGO network require that, for neutron-star binaries, the expansion be carried to order v^6 beyond our Newtonian, quadrupole analysis!

At the end of the inspiral, the binary's stars (or black holes) come crashing together. To compute the waves from this final merger, with an accuracy comparable to the expected observations, it is necessary to solve the Einstein field equation on a computer. The techniques for this are called *numerical relativity*. Numerical relativity is currently in its infancy, but has great promise for producing new insights into general relativity.

****

EXERCISES

Exercise 26.7 *Example: Quadrupolar wave generation in linearized theory*
Derive the quadrupolar wave-generation formula (26.111) for a slow-motion, weak-gravity source in linearized theory, in Lorenz gauge, beginning with the retarded-integral formula

$$\bar{h}_{\mu\nu}(t, \mathbf{x}) = \int \frac{4T_{\mu\nu}(t - |\mathbf{x} - \mathbf{x}'|, \mathbf{x}'])}{|\mathbf{x} - \mathbf{x}'|} dV_{x'} \tag{26.128}$$

[Eq. (24.107)]. Your derivation might proceed as follows:

(a) Show that for a slow-motion source, the retarded integral gives for the $1/r \equiv 1/|\mathbf{x}|$ (radiative) part of $\bar{h}_{jk}$

$$\bar{h}_{jk}(t, \mathbf{x}) = \frac{4}{r}\int T_{jk}(t - r, \mathbf{x}')dV_{x'} \ . \tag{26.129}$$

(b) Show that in linearized theory in Lorenz gauge, the vacuum Einstein equations $-\bar{h}_{\mu\nu,\alpha}{}^{\alpha} = 16\pi T_{\mu\nu}$ [Eq. (24.106)] and the Lorenz gauge condition $\bar{h}_{\mu\nu,}{}^{\nu} = 0$ [Eq. (24.105)] together imply that the stress-energy tensor that generates the waves must have vanishing coordinate divergence, $T^{\mu\nu}{}_{,\nu} = 0$. This means that linearized theory is ignorant of the influence of self gravity on the gravitating $T^{\mu\nu}$!

(c) Show that this vanishing divergence implies $[T^{00}x^j x^k]_{,00} = [T^{lm}x^j x^k]_{,ml} - 2[T^{lj}x^k + T^{lk}x^j]_{,l} + 2T^{jk}$.

(d) By combining the results of (a) and (c), deduce that

$$\bar{h}_{jk}(t,\mathbf{x}) = \frac{2}{r}\frac{d^2 I_{jk}(t-r)}{dt^2} \ , \tag{26.130}$$

where I_{jk} is the second moment of the source's (Newtonian) mass-energy distribution $T^{00} = \rho$ [Eq. (26.110)].

(e) Noticing that the trace-reversed metric perturbation (26.130) has the "speed-of-light-propagation" form, deduce that the gravitational-wave field $h^{\rm TT}_{jk}$ can be computed from (26.130) by a transverse-traceless projection, Eq. (26.96).

Comment: Part (b) shows that this linearized-theory analysis is incapable of deducing the gravitational waves emitted by a source whose dynanmics is controlled by its self gravity, e.g., a nearly Newtonian binary star system. By contrast, the derivation of the quadrupole formula given in Sec. 26.4.2 is valid for any slow-motion source, regardless of the strength and roles of its internal gravity; see the discussion following Eq. (26.109).

Exercise 26.8 *Problem: Energy carried by gravitational waves*
Compute the net rate at which the quadrupolar waves (26.109) carry energy away from their source, by carrying out the surface integral (24.114) with T^{0j} being Isaacson's gravitational-wave energy flux (26.58). Your answer should be Eq. (26.113). [Hint: perform the TT projection in Cartesian coordinates using the projection tensor (26.94), and make use of the following integrals over solid angle on the unit sphere

$$\frac{1}{4\pi}\int n_i d\Omega = 0 \ , \quad \frac{1}{4\pi}\int n_i n_j d\Omega = \frac{1}{3}\delta_{ij} \ , \quad \frac{1}{4\pi}\int n_i n_j n_k d\Omega \ = 0; \ ,$$
$$\frac{1}{4\pi}\int n_i n_j n_k n_l d\Omega = \frac{1}{15}(\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}). \tag{26.131}$$

These relations should be obvious by symmetry, aside from the numerical factors out in front. Those factors are most easily deduced by computing the z components, i.e., by setting $i = j = k = l = z$ and using $n_z = \cos\theta$.]

Exercise 26.9 *Problem: Energy removed by gravitational radiation reaction*
Burke's radiation-reaction potential (26.115) produces a force per unit volume $-\rho\boldsymbol{\nabla}\Phi^{\rm react}$ on its nearly Newtonian source. If we multiply this force per unit volume by the velocity

$\mathbf{v} = d\mathbf{x}/dt$ of the source's material, we obtain thereby a rate of change of energy per unit volume. Correspondingly, the net rate of change of the system's mass-energy must be

$$\frac{dM}{dt} = -\int \rho \mathbf{v} \cdot \boldsymbol{\nabla} \Phi^{\rm react} dV_x \ . \tag{26.132}$$

Show that, when averaged over a few gravitational-wave periods, this formula agrees with the rate of change of mass (26.113) that we derived by integrating the outgoing waves' energy flux.

Exercise 26.10 *Problem: Propagation of waves through an expanding universe*
As we shall see in Chap. 27, the following line element is a possible model for the large-scale structure of our universe:

$$ds^2 = b^2[-d\eta^2 + d\chi^2 + \chi^2(d\theta^2 + \sin^2\theta d\phi^2)] \ , \quad \text{where} \quad b = b_o \eta^2 \tag{26.133}$$

and b_o is a constant with dimensions of length. This is an expanding universe with flat spatial slices $\eta =$ constant. Notice that the proper time measured by observers at rest in the spatial coordinate system is $t = b_o \int \eta^2 d\eta = (b_o/3)\eta^3$. A nearly Newtonian, circular binary is at rest at $\chi = 0$ in an epoch when $\eta \sim \eta_o$. The coordinates of the binary's local asymptotic rest frame are (t, r, θ, ϕ) where $r = a\chi$ and the coordinates cover only a tiny region of the universe, $\chi \lesssim \chi_o \ll \eta_o$. The gravitational waves in this local asymptotic rest frame are described by the Eqs. (26.123) and (26.124); see also Sec. 26.3.5. Use geometric optics (Sec. 26.3.6) to propagate these waves out through the expanding universe. In particular

(a) Show that the null rays are the curves of constant θ, ϕ, and $\eta - \chi$.

(b) Show that the orthonormal basis vectors $\vec{e}_{\hat{\theta}}$, $\vec{e}_{\hat{\phi}}$ associated with the $(\eta, \chi, \theta, \phi)$ coordinates are parallel transported along the rays. (This should be fairly obvious from symmetry.)

(c) Show that the wave fields have the form (26.124) with $t - r$ replaced by the retarded time $\tau_r = \frac{1}{3} b_o (\eta - \chi)^3$, and with $1/r$ being some function of χ and η (what is that function?).

Exercise 26.11 *Problem: Gravitational waves emitted by a linear oscillator*
Consider a mass m attached to a spring so it oscillates along the z axis of a Cartesian coordinate system, moving along the world line $z = a\cos\Omega t$, $y = x = 0$. Use the quadrupole moment formalism to compute the gravitational waves $h_+(t, r, \theta, \phi)$ and $h_\times(t, r, \theta, \phi)$ emitted by this oscillator, with the polarization tensors chosen as in Eqs. (26.123). Pattern your analysis after the computation of waves from a binary in Sec. 26.4.3 .

Exercise 26.12 *Problem: Gravitational waves from waving arms*
Wave your arms rapidly, and thereby try to generate gravitational waves.

(a) Compute in order of magnitude, using classical general relativity, the wavelength of the waves you generate and their dimensionless amplitude at a distance of one wavelength away from you.

(b) How many gravitons do you produce per second? Discuss the implications of your result.

****

26.5 The Detection of Gravitational Waves

Physicists and astronomers are searching for gravitational waves in four different frequency bands using four different techniques:

- In the extremely low frequency (ELF) band, $\sim 10^{-15}$ to $\sim 10^{-18}$ Hz, gravitational waves are sought via their imprint on the polarization of the cosmic microwave background (CMB) radiation. There is only one expected ELF source of gravitational waves, but it is a very interesting one: quantum fluctuations in the gravitational field (spacetime curvature) that emerge from the big bang's quantum-gravity regime, the *Planck era*, and that are subsequently amplified to classical, detectable sizes by the universe's early inflationary expansion. We shall study this amplification and the resulting ELF gravitational waves in Chap. 27 and shall see these waves' great potential for probing the physics of inflation.

- In the very low frequency (VLF) band, $\sim 10^{-7}$ to $\sim 10^{-9}$ Hz, gravitational waves are sought via their influence on the propagation of radio waves emitted by pulsars (spinning neutron stars) and by the resulting fluctuations in the arrival times of the pulsars' radio-wave pulses at earth. The expected VLF sources are violent processes in the first fraction of a second of the universe's life (Chap. 27), and the orbital motion of extremely massive pairs of black holes in the distant universe.

- In the low frequency (LF) band, $\sim 10^{-4}$ to ~ 0.1 Hz, gravitational waves are currently sought via their influence on the radio signals by which NASA tracks interplanetary spacecraft. In ~ 2012 this technique will be supplanted by LISA, the Laser Interferometer Space Antenna—three "drag-free" spacecraft in a triangular configuration with 5 kilometer long arms, that track each other via laser beams. LISA is likely to see waves from massive black-hole binaries (hole masses $\sim 10^5$ to $10^7 M_\odot$) out to cosmologiocal distances; from small holes, neutron stars, and white dwarfs spiraling into massive black holes out to cosmological distances; from the orbital motion of white-dwarf binaries, neutron-star binaries, and stellar-mass black-hole binaries in our own galaxy; and possibly from violent processes in the very early universe.

- The high frequency (HF) band, ~ 10 to $\sim 10^3$ Hz, is where earth-based detectors operate: laser interferometer gravitational wave detectors such as LIGO, and resonant-mass detectors in which a gravitational wave alters the amplitude and phase of vibrations of a normal mode of a large, cylindrical bar. These detectors are likely to see waves from spinning, slightly deformed neutron stars in our own galaxy, and from a variety of sources in the distant universe: the final inspiral and collisions of binaries made

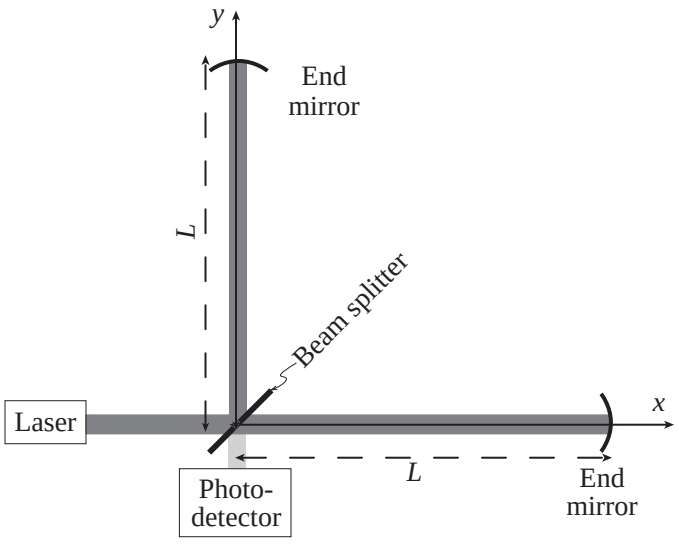


Fig. 26.3: An idealized gravitational-wave interferometer

from neutron stars and/or stellar-mass black holes (up to hundreds of solar masses); the tearing apart of a neutron star by the spacetime curvature of a companion black hole; supernovae and the triggers of gamma ray bursts; and possibly waves from violent processes in the very early universe.

For detailed discussions of these gravitational-wave sources in all four frequency bands, and of prospects for their detection, see e.g. Cutler and Thorne (2002) and references therein. It is likely that waves will be seen in all four bands within the next 20 years, and the first detection is likely to occur in the HF band using gravitational-wave interferometers such as LIGO.

We briefly discussed such interferometers in Sec. 8.5, focusing on optical interferometry issues. In this chapter we shall analyze the interaction of a gravitational wave with such an interferometer. That analysis will not only teach us much about gravitational waves, but will also illustrate some central issues in the physical interpretation of general relativity theory.

To get quickly to the essentials, we shall examine initially a rather idealized detector: A Michaelson interferometer (one without the input mirrors of Fig. 8.11) that floats freely in space, so there is no need to hang its mirrors by wires; see Fig. 26.3. At the end of this chapter, we shall briefly discuss more realistic interferometers and their analysis.

We shall use linearized theory to analyze the interaction of our idealized interferometer with a gravitational wave. We shall perform our analysis twice, using two different coordinate systems (two different gauges). Our two analyses will predict the same results for the interferometer output, but they will appear to attribute those results to two different mechanisms.

In our first analysis (performed in TT gauge; Sec. 26.5.1) the interferometer's test masses will remain always at rest in our chosen coordinate system, and the gravitational waves $h_+(t-z)$ will interact with the interferometer's light. The imprint that $h_+(t-z)$ leaves on the light will cause a fluctuating light intensity $I_{\rm out}(t) \propto h_+(t)$ to emerge from the

interferometer's output port and be measured by the photodiode.

In our second analysis (performed in the proper reference frame of the interferometer's beam splitter; Sec. 26.5.2) the gravitational waves will interact hardly at all with the light. Instead, they will push the end mirrors back and forth relative to the coordinate system, thereby lengthening one arm while shortening the other. These changing arm lengths will cause a changing interference of the light returning to the beam splitter from the two arms, and that changing interference will produce the fluctuating light intensity $I_{\rm out}(t) \propto h_+(t)$ measured by the photodiodes.

These differences of viewpoint are somewhat like the differences between the Heisenberg Picture and the Schrödinger Picture in quantum mechanics. The intuitive pictures associated with two viewpoints appear to be very different (Schrödinger's wave function vs. Heisenberg's matrices; gravitational waves interacting with light vs. gravitational waves pushing on mirrors). But whenever one computes the same physical observable from the two different viewpoints (probability for a quantum measurement outcome; light intensity measured by photodetector), the two viewpoints give the same answer.

26.5.1 Interferometer analyzed in TT gauge

For our first analysis, we place the interferometer at rest in the x-y plane of a TT coordinate system, with its arms along the x and y axes and its beam splitter at the origin as shown in Fig. 26.3. For simplicity, we assume that the gravitational wave propagates in the z direction and has + polarization, so the linearized spacetime metric has the TT-gauge form

$$\boxed{ds^2 = -dt^2 + [1 + h_+(t-z)]dx^2 + [1 - h_+(t-z)]dy^2 + dz^2} \tag{26.134}$$

[Eq. (26.89)]. For ease of notation, we shall omit the subscript + from h_+ in the remainder of this section.

The beam splitter and end mirrors move freely and thus travel along geodesics of the metric (26.134). The splitter and mirrors are at rest in the TT coordinate system before the wave arrives, so initially the spatial components of their 4-velocities vanish, $u_j = 0$. Because the metric coefficients $g_{\alpha\beta}$ are all independent of x and y, the geodesic equation dictates that the components u_x and u_y are conserved and thus remain zero as the wave passes, which implies (since the metric is diagonal) $u^x = dx/d\tau = 0$ and $u^y = dy/d\tau = 0$. One can also show (see Ex. 26.13) that $u^z = dz/d\tau = 0$ throughout the wave's passage. Thus, in terms of motion relative to the TT coordinate system, the gravitational wave has no influence at all on the beam splitter and mirrors; they all remain at rest (constant x, y and z) as the waves pass.

(Despite this lack of motion, the proper distances between the mirrors and the beam splitter—the interferometer's physically measured arm lengths—do change. If the unchanging coordinate lengths of the two arms are $\Delta x = \ell_x$ and $\Delta y = \ell_y$, then the metric (26.134) says that the physically measured arm lengths are

$$\boxed{L_x = \left[1 + \frac{1}{2}h(t)\right]\ell_x\ , \quad L_y = \left[1 - \frac{1}{2}h(t)\right]\ell_y\ .} \tag{26.135}$$

When h is positive, the x arm is lengthened and the y arm is shortened; when negative, L_x is shortened and L_y is lengthened.)

Turn, next, to the propagation of light in the interferometer. We assume, for simplicity, that the light beams have large enough transverse sizes that we can idealize them, on their optic axes, as plane electromagnetic waves. (In reality, they will be Gaussian beams, of the sort studied in Sec. 7.5.5). The light's vector potential satisfies the curved-spacetime vacuum wave equation $A^{\alpha;\mu}{}_{\mu} = 0$ [Eq. (24.71) with vanishing Ricci tensor]. We write the vector potential in geometric optics (eikonal-approximation) form as

$$A^{\alpha} = \Re(\mathcal{A}^{\alpha} e^{i\phi}) \ , \tag{26.136}$$

where $\mathcal{A}^{\alpha}$ is a slowly varying amplitude and ϕ is a rapidly varying phase; cf. Eq. (6.18). Because the wavefronts are (nearly) planar and the spacetime metric is nearly flat, the light's amplitude $\mathcal{A}^{\mu}$ will be very nearly constant as it propagates down the arms, and we can ignore its variations. Not so the phase. It oscillates at the laser frequency, $\omega_o \sim 3 \times 10^{14}$ Hz; i.e., $\phi^{\rm out}_{x\ {\rm arm}} \simeq \omega_o(x-t)$ for light propagating outward from the beam splitter along the x arm, and similarly for the returning light and the light in the y arm. The gravitational wave imprints onto the phase tiny deviations from this $\omega_o(x-t)$; we must compute those imprints.

In the spirit of geometric optics, we introduce the light's spacetime wave vector

$$\vec{k} \equiv \vec{\nabla}\phi \ , \tag{26.137}$$

and we assume that $\vec{k}$ varies extremely slowly compared to the variations of ϕ. Then the wave equation $A^{\alpha;\mu}{}_{\mu} = 0$ reduces to the statement that the wave vector is null, $\vec{k}\cdot\vec{k} = \phi_{,\alpha}\phi_{,\beta}g^{\alpha\beta} = 0$. For light in the x arm the phase depends only on x and t; for that in the y arm it depends only on y and t. Combining this with the TT metric (26.134) and noting that the interferometer lies in the $z = 0$ plane, we obtain

$$\begin{aligned} -\left(\frac{\partial\phi_{x\ {\rm arm}}}{\partial t}\right)^2 + [1-h(t)]\left(\frac{\partial\phi_{x\ {\rm arm}}}{\partial x}\right)^2 &= 0 \ , \\ -\left(\frac{\partial\phi_{y\ {\rm arm}}}{\partial t}\right)^2 + [1+h(t)]\left(\frac{\partial\phi_{y\ {\rm arm}}}{\partial y}\right)^2 &= 0 \ . \end{aligned} \tag{26.138}$$

We idealize the laser as perfectly monochromatic and we place it at rest in our TT coordinates, arbitrarily close to the beam splitter. Then the outgoing light frequency, as measured by the beam splitter, must be precisely ω_o and cannot vary with time. Since proper time, as measured by the beam splitter, is equal to coordinate time t [cf. the metric (26.134))], the frequency that the laser and beam splitter measure must be $\omega = -\partial\phi/\partial t = -k_t$. This dictates the following boundary conditions (initial conditions) on the phase of the light that travels outward from the beam splitter:

$$\frac{\partial\phi^{\rm out}_{x\ {\rm arm}}}{\partial t} = -\omega_o \text{ at } x = 0 \ , \quad \frac{\partial\phi^{\rm out}_{y\ {\rm arm}}}{\partial t} = -\omega_o \text{ at } y = 0 \ . \tag{26.139}$$

It is straightforward to verify that the solutions to Eq. (26.138) [and thence to the wave equation and thence to Maxwell's equation] that satisfy the boundary conditions (26.139)

are

$$\begin{aligned}\phi^{\rm out}_{x\ {\rm arm}} &= -\omega_o\left[t - x + \frac{1}{2}H(t-x) - \frac{1}{2}H(t)\right] ,\\ \phi^{\rm out}_{y\ {\rm arm}} &= -\omega_o\left[t - y - \frac{1}{2}H(t-y) + \frac{1}{2}H(t)\right] ,\end{aligned} \tag{26.140}$$

where $H(t)$ is the first time integral of the gravitational waveform,

$$H(t) \equiv \int_0^t h(t')dt' ; \tag{26.141}$$

cf. Ex. 26.14.

The outgoing light reflects off the mirrors, which are at rest in the TT coordinates at locations $x = \ell_x$ and $y = \ell_y$. As measured by observers at rest in these coordinates, there is no doppler shift of the light because the mirrors are not moving. Correspondingly, the phases of the reflected light, returning back along the two arms, have the following forms:

$$\begin{aligned}\phi^{\rm back}_{x\ {\rm arm}} &= -\omega_o\left[t + x - 2\ell_x + \frac{1}{2}H(t+x-2\ell_x) - \frac{1}{2}H(t)\right] ,\\ \phi^{\rm back}_{y\ {\rm arm}} &= -\omega_o\left[t + y - 2\ell_y - \frac{1}{2}H(t+y-2\ell_y) + \frac{1}{2}H(t)\right] .\end{aligned} \tag{26.142}$$

The difference of the phases of the returning light, at the beam splitter ($x = y = 0$), is

$$\begin{aligned}\Delta\phi &\equiv \phi^{\rm back}_{x\ {\rm arm}} - \phi^{\rm back}_{y\ {\rm arm}} = -\omega_o[-2(\ell_x - \ell_y) + \frac{1}{2}H(t-2\ell_x) + \frac{1}{2}H(t-2\ell_y) - H(t)]\\ &\simeq +2\omega_o[\ell_x - \ell_y + \ell h(t)] \quad \text{for earth-based interferometers.}\end{aligned} \tag{26.143}$$

In the second line we have used the fact that for earth-based interferometers operating in the high-frequency band, the gravitational wavelength $\lambda_{\rm GW} \sim c/(100{\rm Hz}) \sim 3000$ km is long compared to the interferometers' ~ 4 km arms, and the arms have nearly the same length, $\ell_y \simeq \ell_x \equiv \ell$.

The beam splitter sends a light field $\propto e^{i\phi^{\rm back}_{x\ {\rm arm}}} + e^{i\phi^{\rm back}_{y\ {\rm arm}}}$ back toward the laser, and a field $\propto e^{i\phi^{\rm back}_{x\ {\rm arm}}} - e^{i\phi^{\rm back}_{y\ {\rm arm}}} = e^{i\phi^{\rm back}_{y\ {\rm arm}}}(e^{i\Delta\phi} - 1)$ toward the photodetector. The intensity of the light entering the photodetector is proportional to the squared amplitude of the field, $I_{\rm PD} \propto |e^{i\Delta\phi} - 1|^2$. We adjust the interferometer's arm lengths so their difference $\ell_x - \ell_y$ is small compared to the light's reduced wavelength $1/\omega_o = c/\omega_o$ but large compared to $|\ell h(t)|$. Correspondingly, $|\Delta\phi| \ll 1$, so only a tiny fraction of the light goes toward the photodetector (it is the interferometer's "dark port"), and that dark-port light intensity is

$$\boxed{I_{\rm PD} \propto |e^{i\Delta\phi} - 1|^2 \simeq |\Delta\phi|^2 \simeq 4\omega_o^2(\ell_x - \ell_y)^2 + 8\omega_o^2(\ell_x - \ell_y)\ell h(t) .} \tag{26.144}$$

The time varying part of this intensity is proportional to the gravitational waveform $h(t)$, and it is this time varying part that the photodetector reports as the interferometer output.

26.5.2 Interferometer analyzed in proper reference frame of beam splitter

We shall now reanalyze our idealized interferometer in the proper reference frame of its beam splitter, denoting that frame's coordinates by $\hat{x}^\alpha$. Because the beam splitter is freely falling (moving along a geodesic through the gravitational-wave spacetime), its proper reference frame is locally Lorentz ("LL"), and its metric coefficients have the form $g_{\hat{\alpha}\hat{\beta}} = \eta_{\alpha\beta} + \mathrm{O}(\delta_{jk}\hat{x}^j\hat{x}^k/\mathcal{R}^2)$ [Eq. (24.15)]. Here $\mathcal{R}$ is the radius of curvature of spacetime, and $1/\mathcal{R}^2$ is of order the components of the Riemann tensor, which have magnitude $\ddot{h}(\hat{t}-\hat{z})$ [Eq. (26.41) with t and z equal to $\hat{t}$ and $\hat{z}$ aside from fractional corrections of order h]. Thus,

$$g_{\hat{\alpha}\hat{\beta}} = \eta_{\alpha\beta} + \mathrm{O}[\ddot{h}(\hat{t}-\hat{z})\delta_{jk}\hat{x}^j\hat{x}^k] \; . \tag{26.145}$$

The following coordinate transformation takes us from the TT coordinates x^α used in the previous section to the beam splitter's LL coordinates:

$$\begin{aligned} x &= \left[1 - \frac{1}{2}h(\hat{t}-\hat{z})\right]\hat{x} \; , \qquad y = \left[1 + \frac{1}{2}h(\hat{t}-\hat{z})\right]\hat{y} \; , \\ t &= \hat{t} - \frac{1}{4}\dot{h}(\hat{t}-\hat{z})(\hat{x}^2-\hat{y}^2) \; , \qquad z = \hat{z} - \frac{1}{4}\dot{h}(\hat{t}-\hat{z})(\hat{x}^2-\hat{y}^2) \; . \end{aligned} \tag{26.146}$$

It is straightforward to insert this coordinate transformation into the TT-gauge metric (26.134) and thereby obtain, to linear order in h,

$$\boxed{ds^2 = -d\hat{t}^2 + d\hat{x}^2 + d\hat{y}^2 + d\hat{z}^2 + \frac{1}{2}(\hat{x}^2-\hat{y}^2)\ddot{h}(t-z)(d\hat{t}-d\hat{z})^2 \; .} \tag{26.147}$$

This has the expected LL form (26.145) and, remarkably, it turns out not only to be a solution of the vacuum Einstein equations in linearized theory but also an exact solution to the full vacuum Einstein equations [cf. Ex. 35.8 of MTW].

Throughout our idealized interferometer, the magnitude of the metric perturbation in these LL coordinates is $|h_{\hat{\alpha}\hat{\beta}}| \lesssim (\ell/\lambda\!\!\!^{-}_{\rm GW})^2 h$, where $\lambda\!\!\!^{-}_{\rm GW} = \lambda_{\rm GW}/2\pi$ is the waves' reduced wavelength and h is the magnitude of $h(\hat{t}-\hat{z})$. For earth-based interferometers operating in the HF band (~ 10 to ~ 1000 Hz), $\lambda\!\!\!^{-}_{\rm GW}$ is of order 50 to 5000 km, and the arm lengths are $\ell \leq 4$ km, so $(L/\lambda\!\!\!^{-})^2 \lesssim 10^{-2}$ to 10^{-6}. Thus, the metric coefficients $h_{\hat{\alpha}\hat{\beta}}$ are no larger than $h/100$. This has a valuable consequence for the analysis of the interferometer: Up to fractional accuracy $\sim (\ell/\lambda\!\!\!^{-}_{\rm GW})^2 h \lesssim h/100$, the LL coordinates are globally Lorentz throughout the interferometer; i.e., $\hat{t}$ measures proper time, and $\hat{x}^j$ are Cartesian and measure proper distance. In the rest of this section, we shall restrict attention to such earth-based interfrometers, but shall continue to idealize them as freely falling.

The beam splitter, being initially at rest at the origin of these LL coordinates, remains always at rest, but the mirrors move. Not surprisingly, the geodesic equation for the mirrors in the metric (26.147) dictates that their coordinate positions are, up to fractional errors of order $(\ell/\lambda\!\!\!^{-}_{\rm GW})^2 h$,

$$\begin{aligned} \hat{x} = L_x &= \left[1 + \frac{1}{2}h(\hat{t})\right]\ell_x \; , \quad \hat{y} = \hat{z} = 0 \quad \text{for mirror in } x \text{ arm}, \\ \hat{y} = L_y &= \left[[1 - \frac{1}{2}h(\hat{t})\right]\ell_y \; , \quad \hat{x} = \hat{z} = 0 \quad \text{for mirror in } y \text{ arm}. \end{aligned} \tag{26.148}$$

(This can also be deduced from the gravitational-wave tidal acceleration $-R^{\rm GW}_{\hat{t}\hat{0}\hat{k}\hat{0}}\hat{x}^k$, as in Eq. (26.45), and from the fact that to good accuracy $\hat{x}$ and $\hat{y}$ measure proper distance from the beam splitter.) Thus, although the mirrors do not move in TT coordinates, they do move in LL coordinates. The two coordinate systems predict the same time-varying physical arm lengths (the same proper distances from beam splitter to mirrors), L_x and L_y [Eqs. (26.135) and (26.148)].

As in TT coordinates, so also in LL coordinates, we can analyze the light propagation in the geometric optics approximation, with $A^{\hat{\alpha}} = \Re(\mathcal{A}^{\hat{\alpha}} e^{i\phi})$. Just as the wave equation for the vector potential dictates, in TT coordinates, that the rapidly varying phase of the outward light in the x arm has the form $\phi^{\rm out}_{x\ {\rm arm}} = -\omega_o(t-x) + {\rm O}(\omega_o \ell h_{\mu\nu})$ [Eq. (26.140) with $x \sim \ell \ll \lambda\!\!\!^{-}_{\rm GW}$ so $H(t-x) - H(t) \simeq \dot{H}(t)x = h(t)x \sim hL \sim h_{\mu\nu}L$], so similarly the wave equation in LL coordinates turns out to dictate that

$$\phi^{\rm out}_{x\ {\rm arm}} = -\omega_o(\hat{t}-\hat{x}) + {\rm O}(\omega_o \ell h_{\hat{\mu}\hat{\nu}}) = -\omega_o(\hat{t}-\hat{x}) + {\rm O}\left(\omega_o \ell h \frac{\ell^2}{\lambda\!\!\!^{-2}_{\rm GW}}\right) , \tag{26.149}$$

and similarly for the returning light and the light in the y arm. The term ${\rm O}(\omega_o \ell h\ \ell^2/\lambda\!\!\!^{-2}_{\rm GW})$ is the influence of the direct interaction between the gravitational wave and the light. Aside from this term, the analysis of the interferometer proceeds in exactly the same way as in flat space (because $\hat{t}$ measures proper time and $\hat{x}$ and $\hat{y}$ proper distance): The light travels a round trip distance L_x in one arm and L_y in the other, and therefore acquires a phase difference, upon arriving back at the beam splitter, given by

$$\begin{aligned} \Delta\phi &= -\omega_o[-2(L_x - L_y)] + {\rm O}\left(\omega_o \ell h \frac{\ell^2}{\lambda\!\!\!^{-2}_{\rm GW}}\right) \\ &\simeq +2\omega_o[\ell_x - \ell_y + \ell h(\hat{t})] + {\rm O}\left(\omega_o \ell h \frac{\ell^2}{\lambda\!\!\!^{-2}_{\rm GW}}\right) . \end{aligned} \tag{26.150}$$

This net phase difference for the light returning from the two arms is the same as we deduced in TT coordinates [Eq. (26.143)], up to the negligible correction ${\rm O}(\omega_o \ell h\ \ell^2/\lambda\!\!\!^{-2}_{\rm GW})$, and therefore the time-varying intensity of the light into the photodiode will be the same [Eq. (26.144)].

In our TT analysis the phase shift $2\omega_o \ell h(t)$ arose from the interaction of the light with the gravitational waves. In the LL analysis, it is due to the displacements of the mirrors in the LL coordinates (i.e., the displacements as measured in terms of proper distance), which cause the light to travel different distances in the two arms. The direct LL interaction of the waves with the light produces only the tiny correction ${\rm O}(\omega_o \ell h\ \ell^2/\lambda\!\!\!^{-2}_{\rm GW})$ to the phase shift.

It should be evident that the LL description is much closer to elementary physics than the TT description. This is always the case, when one's apparatus is sufficiently small that one can regard $\hat{t}$ as measuring proper time and $\hat{x}^j$ as Cartesian coordinates that measure proper distance throughout the apparatus. But for a large apparatus (e.g. LISA, with its arm lengths $\ell \gtrsim \lambda\!\!\!^{-}_{\rm GW}$) the LL analysis becomes quite complicated, as one must pay close attention to the ${\rm O}(\omega_o \ell h\ \ell^2/\lambda\!\!\!^{-2}_{\rm GW})$ corrections. In such a case, the TT analysis is much simpler.

26.5.3 Realistic Interferometers

For realistic, earth-based interferometers, one must take account of the acceleration of gravity. Experimenters do this by hanging their beam splitters and test masses on wires or fibers. The simplest way to analyze such an interferometer is in the proper reference frame of the beam splitter, where the metric must now include the influence of the acceleration of gravity by adding a term $-2g_e\hat{z}$ to the metric coefficient $h_{\hat{0}\hat{0}}$ [cf. Eq. (23.87)]. The resulting analysis, like that in the LL frame of our freely falling interferometer, will be identical to what one would do in flat spacetime, so long as one takes account of the motion of the test masses as dictated by the gravitational-wave tidal acceleration $-R_{\hat{i}\hat{0}\hat{j}\hat{0}}\hat{x}^j$, and so long as one is willing to ignore the tiny effects of $\mathrm{O}(\omega_o \ell h\, \ell^2/\lambda_{\rm GW}^2)$.

To make the realistic interferometer achieve high sensitivity, the experimenters introduce a lot of clever complications, such as the input mirrors of Fig. 8.11 which turn the arms into Fabry-Perot cavities. All these complications can be analyzed, in the beam splitter's proper reference frame, using standard flat-spacetime techniques, so long as one makes sure to take account of the end-mirror motion as dictated by the gravitational-wave tidal acceleration. The direct coupling of the light to the gravitational waves can be neglected, as in our idealized interferometer.

****

EXERCISES

Exercise 26.13 *Derivation and Practice: Geodesic motion in TT coordinates*
Consider a particle that is at rest in the TT coordinate system of the gravitational-wave metric (26.134) before the gravitational wave arrives. In the text it is shown that the particle's 4-velocity has $u^x = u^y = 0$ as the wave passes. Show that $u^z = 0$ and $u^t = 1$ as the wave passes, so the components of the particle's 4-velocity are unaffected by the passing gravitational wave.

Exercise 26.14 *Example: Light in an interferometric gravitational wave detector in TT gauge*
Consider the light propagating outward from the beam splitter, along the x arm of an interferometric gravitational wave detector, as analyzed in TT gauge, so (suppressing the subscript "x arm" and superscript "out") the electromagnetic vector potential is $A^\alpha = \Re(\mathcal{A}^\alpha e^{i\phi(x,t)})$ with $\mathcal{A}^\alpha$ constant and with $\phi = -\omega_o\left[t - x + \frac{1}{2}H(t-x) - \frac{1}{2}H(t)\right]$ [Eqs. (26.140) and (26.141).]

(a) Show that this ϕ satisfies the nullness equation (26.138), as claimed in the text — which implies that $A^\alpha = \Re(\mathcal{A}^\alpha e^{i\phi(x,t)})$ satisfies Maxwell's equations in the geometric optics limit.

(b) Show that this ϕ satisfies the initial condition (26.139), as claimed in the text.

(c) Show, by an argument analogous to Eq. (26.76), that $\nabla_{\vec{k}}\vec{k} = 0$. Thus, the wave vector must be the tangent vector to geometric optics rays that are null geodesics

in the gravitational-wave metric. Photons travel along these null geodesics and have 4-momenta $\vec{p} = \hbar\vec{k}$.

(d) Because the gravitational-wave metric (26.134) is independent of x, the p_x component of a photon's 4-momentum must be conserved along its geodesic world line. Compute $p_x = k_x = \partial\phi/\partial x$, thereby verify this conservation law.

(e) Explain why the photon's frequency, as measured by observers at rest in our TT coordinate system, is $\omega = -k_t = -\partial\phi/\partial t$. Explain why the rate of change of this frequency, as computed moving with the photon, is $d\omega/dt \simeq (\partial/\partial t + \partial/\partial x)\omega$, and show that $d\omega/dt \simeq -\frac{1}{2}\omega_o dh/dt$.

****

Bibliographic Note

For an up-to-date, elementary introduction to experimental tests of general relativity in the solar system, we recommend Chap. 10 of Hartle (2003). For an enjoyable, popular-level book on experimental tests, see Will (1993a). For a very complete monograph on the theory underlying experimental tests, see Will (1993b), and for an up to date review of experimental tests, see Will (2006).

For an elementary, up to date, and fairly complete introduction to gravitational waves, we recommend Chaps. 16 and 23 of Hartle (2003). Also good at the elementary level, but less complete and less up to date, is Chap. 9 of Schutz (1980). For a more advanced treatment of the properties of gravitational waves, their propagation and their generation, we suggest MTW Sec. 18.2 and Chaps. 35 and 36; but Chap. 37 on gravitational wave detection is terribly out of date and not recommended. For a more advanced presentation of gravitational wave theory, see Thorne (1983), and for a fairly complete and nearly up to date review of gravitational wave sources for ground-based detectors (LIGO etc.) and space-based detectors (LISA etc.), see Cutler and Thorne (2002). For a lovely monograph on the physics of interferometric gravitational-wave detectors, see Saulson (1994). For a two-term course on gravitational waves (theory and experiment), including videos of lectures plus large numbers of readings, problems and solutions, see Thorne, Bondarescu and Chen (2002). For a compendium of educational materials about gravitational waves, see the website of the LIGO Academic Advisory Committee, http://www.ligo.caltech.edu/laac/ .

Bibliography

Bertotti, B., Iess, L. and Tortora, T., *Nature* **425**, 374 (2003).

Burke, William L., 1971. "Gravitational radiation damping of slowly moving systems calculated using matched asymptotic expansions," *Journal of Mathematical Physics*, **12**, 402–418.

Box 26.2
Important Concepts in Chapter 25

- Experimental Tests of general relativity, Sec. 26.2
  - Weak equivalence principle (universality of free fall), Sec. 26.2.1
  - Gravitational redshift, Sec. 26.2.1
  - Perihelion shift, Sec.26.2.2 and Ex. 26.2
  - Fermat's principle, gravitational lenses, and deflection of light, Sec. 26.2.3 and Ex. 26.3
  - Shapiro time delay, Sec. 26.2.4
  - Frame dragging, Sec. 26.2.5
- Gravitational Wave Properties
  - Definition via short-wavelength condition $\lambda\!\!\!^{-} \ll \{\mathcal{R}, \mathcal{L}\}$, Sec. 26.3.1
  - Background metric $g^{\rm B}_{\alpha\beta}$ and waves' metric perturbation $h_{\alpha\beta}$ and Riemann tensor $R^{\rm GW}_{\alpha\beta\gamma\delta}$, Sec. 26.3.1
  - Gravitational-wave fields h_+ and $h_\times$ and their stretch–squeeze physical manifestations, Sec. 26.3.2
  - Gravitons, Sec. 26.3.3 and end of Sec. 26.3.4
  - Gravitatiional-wave stress-energy tensor $T^{\rm GW}_{\alpha\beta}$ and its conservation laws, Sec. 26.3.4 and paragraph containing Eq. (26.113)
- Gravitational wave propagation via geometric optics, Sec. 26.3.6
- TT gauge and computation of $h^{\rm TT}_{jk}$ and correspondingly h_+ and $h_\times$ via TT projection, Sec. 26.3.7
- Gravitational-wave generation, Sec. 26.4
  - Multipolar expansion of h_+ and $h_\times$, Sec. 26.4.1
  - Slow-motion sources: quadrupole-moment formalism for computing wave generation, Sec. 26.4.2
  - Burke's radiation-reaction potential, Eq. (26.115)
  - Application: waves from a binary system, and binary inspiral due to radiation reaction, Sec. 26.4.3
- Gravitational-wave detection, Sec. 26.5
  - Frequency bands: ELF, VLF, LF and HF, Sec. 26.5
  - Interferometric gravitational-wave detector ("interferometer"), Sec. 26.5.1
  - How to analyze an interferometer in TT gauge (Sec. 26.5.1) and in the proper reference frame of the beam splitter (Sec. 26.5.2)

Cutler, Curt and Thorne, Kip S., 2002. "An overview of gravitational wave sources," in Proceedings of the GR16 Conference on General Relativity and Gravitation, ed. N. Bishop and S. D. Maharaj (World Scientific, 2002), 72–111; also available at http://xxx.lanl.gov/abs/gr-qc/0204090

Einstein, Albert, 1916. "Die Grundlage der allgemeinen Relativitätstheorie," *Annalen der Physik*, **49**, 769–822. English translation in Einstein *et al*. (1923).

Einstein, Albert, 1918. "Über Gravitationwellen," *Sitzungsberichte der Königlish Preussischen Akademie der Wissenschaften*, **1918 volume**, 154–167.

Fierz, M. and Pauli, Wolfgang, 1939. "On relativistic wave equations for particles of arbitary spin in an electromagnetic field," *Proceedings of the Royal Society A*, **173**, 211–232.

Hartle, J. B., 2003. *Gravity: An Introduction to Einstein's General Relativity*, San Francisco: Addison-Wesley.

Isaacson, R. A. 1968. *Physical Review* **166**, 1272.

MTW: Misner, Charles W., Thorne, Kip S. and Wheeler, John A., 1973. *Gravitation*, W. H. Freeman & Co., San Francisco.

Saulson, Peter, *Fundamentals of Interferometric Gravitational Wave Detectors*, World Scientific, Singapore.

Schutz, B. 1980. *Geometrical Methods of Mathematical Physics*, Cambridge: Cambridge University Press.

Thorne, Kip S., 1980. *Review of Modern Physics*, **52**, 299.

Thorne, Kip S., 1983. "The Theory of Gravitational Radiation: An Introductory Review," in *Gravitational Radiation*, eds. N. Dereulle and T. Piran, North Holland, Amsterdam, pp. 1–57.

Thorne, Kip S., Bondarescu, Mihai and Chen, Yanbei, 2002. *Gravitational Waves: A Web-Based Course*, http://elmer.tapir.caltech.edu/ph237/

Weissberg, J.M. and Taylor, J.H. in *Proceedings of the Aspen Conference on Binary Radio Pulsars*, eds. F.A. Rasio and I.H. Stairs (Astronomical Society of the Pacific Conference Series, in press); http://xxx.lanl.gov/abs/gr-qc/0407149 .

Will, Clifford M., 1993a. emphWas Einstein Right?: Putting General Relativity to the Test, New York: Basic Books.

Will, Clifford M., 1993b. *Theory and Experiment in Gravitational Physics*, Revised Edition, Cambridge University press, Cambridge, UK.

Will, Clifford M., 2006. "The Confrontation between General Relativity and Experiment," *Living Reviews in Relativity* **9**, URL (cited on cited on 16 May 2007): http://www.livingreviews.org/Irr-2006-3 .

Contents

27 Cosmology **1**
27.1 Overview . . . 1
27.2 Homogeneity and Isotropy of the Universe; Robertson-Walker Line Element . . . 2
27.3 The Stress-energy Tensor and the Einstein Field Equation . . . 9
27.4 Evolution of the Universe . . . 13
27.4.1 Constituents of the universe: Cold matter, radiation, and dark energy 13
27.4.2 The vacuum stress-energy tensor . . . 14
27.4.3 Evolution of the densities . . . 14
27.4.4 Evolution in time and redshift . . . 17
27.4.5 Physical processes in the expanding universe . . . 22
27.5 Observational Cosmology . . . 27
27.5.1 Parameters characterizing the universe . . . 27
27.5.2 Local Lorentz frame of homogeneous observers near Earth . . . 28
27.5.3 Hubble expansion rate . . . 29
27.5.4 Primordial nucleosynthesis . . . 30
27.5.5 Density of Cold Dark Matter . . . 30
27.5.6 Radiation Temperature and Density . . . 31
27.5.7 Anisotropy of the CMB: Measurements of the Doppler Peaks . . . 31
27.5.8 Age of the universe: Constraint on the dark energy . . . 36
27.5.9 Magnitude-Redshift relation for type Ia supernovae: Confirmation that the universe is accelerating . . . 37
27.6 The Big-Bang Singularity, Quantum Gravity, and the Initial Conditions of the Universe . . . 40
27.7 Inflationary Cosmology . . . 42
27.7.1 Amplification of Primordial Gravitational Waves by Inflation . . . 43
27.7.2 Search for Primordial Gravitational Waves by their Influence on the CMB; Probing the Inflationary Expansion Rate . . . 43

Chapter 27

Cosmology

Version 0627.1.K.pdf, 23 May 2007.

Please send comments, suggestions, and errata via email to kip@tapir.calte ch.edu or on paper to Kip Thorne, 130-33 Caltech, Pasadena CA 91125

[NOTE: I have done a quick and incomplete revision of this chapter in light of the observational data from the WMAP satellite (Bennett et. al. 2003). I have not yet had time to do this carefully. — Kip]

Box 27.1
Reader's Guide

- This chapter relies significantly on
  - The special relativity portions of Chap. 1.
  - Chapter 23, on the transition from special relativity to general relativity.
  - Chapter 24, on the fundamental concepts of general relativity.
  - Sec. 25.3.3 on local energy-momentum conservation for a perfect fluid and Sec. 25.6 on the many-fingered nature of time.
- In addition, Box 27.3 and Ex. 27.7 of this chapter rely on the Planckian distribution function for thermalized photons and its evolution (Liouville's theorem or collisionless Boltzmann equation), as presented in Secs. 2.2.4, 2.3, and sec:02EvolutionLaws of Chap. 2.

27.1 Overview

General Relativity is an indispensable foundation for understanding the large scale structure and evolution of the universe (*cosmology*), but it is only one foundation out of many. The crudest of understandings can be achieved with general relativity and little else; but more detailed and deeper understandings require combining general relativity with quantum field

theory, nuclear and atomic physics, thermodynamics, fluid mechanics, and large bodies of astrophysical lore.

In this chapter we shall explore aspects of cosmology which are sufficiently crude that general relativity, augmented by only bits and pieces of other physics, can provide an adequate foundation. Our exploration will simultaneously illustrate key aspects of general relativity and give the reader an overview of modern cosmology.

We shall begin in Sec. 27.2 by discussing the observational data that suggest our universe is homogeneous and isotropic when averaged over regions of space huge compared to clusters of galaxies, and we then shall construct a spacetime metric for an idealized homogeneous, isotropic model for the universe. In Sec. 27.3 we shall construct a stress-energy tensor that describes, approximately, the total, averaged energy and pressure of the universe's matter and radiation; and we shall insert that stress-energy tensor and the metric of Sec. 27.2 into the Einstein field equation, thereby deducing a set of equations that govern the evolution of the universe. In Sec. 27.4 we shall study the predictions that those evolution equations make for the rate of expansion of the universe and the manner in which the expansion changes with time, and we shall describe the most important physical processes that have occured in the universe during its evolution into its present state. As we shall see, the details of the expansion are determined by the values of seven parameters that can be measured today—with the caveat that there may be some big surprises associated with the so-called *dark energy*. In Sec. 27.5 we shall describe the astronomical observations by which the universe's seven parameters are being measured, and the multifaceted evidence for dark energy. In Sec. 27.6 we shall discuss the big-bang singularity in which the universe probably began, and shall discuss the fact that this singularity, like singularities inside black holes, is a signal that general relativity breaks down and must be replaced by a quantum theory of gravity which (hopefully) will not predict singular behavior. We shall also examine a few features that the quantum theory of gravity is likely to exhibit. Finally, in Sec. 27.7 we shall discuss the "inflationary" epoch that the universe appears to have undergone immediately after the quantum gravity, big-bang epoch.

27.2 Homogeneity and Isotropy of the Universe; Robertson-Walker Line Element

The universe obviously is not homogeneous or isotropic in our neighborhood: In our solar system (size $\sim 10^{14}$ cm) almost all the mass is concentrated in the sun and planets, with a great void in between. Looking beyond the solar system, one sees the Milky Way Galaxy (size $\sim 10^{23}$ cm $\sim 10^5$ light years, or equivalently 3×10^4 parsecs),[1] with its mass concentrated toward the center and its density falling off roughly as $1/(\text{distance})^2$ as one moves out past the sun and into the Galaxy's outer reaches. Beyond the Galaxy is the emptiness of intergalactic space; then other galaxies congregated into our own "local group" (size $\sim 10^6$ parsecs). The local group is in the outer reaches of a cluster of several thousand galaxies called the Virgo

[1] One parsec is 3.262 light years, i.e. 3.086×10^{18} cm. It is defined as the distance of a star whose apparent motion on the sky, induced by the Earth's orbital motion, is a circle with radius one arc second.

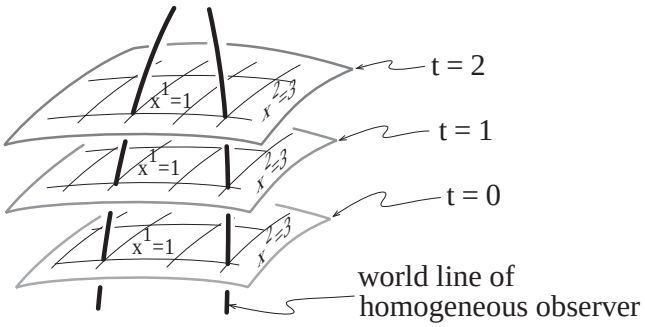


Fig. 27.1: The synchronous coordinate system for a homogeneous, isotropic model of the universe.

cluster (size $\sim 10^7$ parsecs), beyond which is the void of intercluster space, and then other clusters at distances $\gtrsim 10^8$ parsecs.

Despite all this structure, the universe appears to be nearly homogeneous and isotropic on scales $\gtrsim 10^8$ parsecs, i.e., $\gtrsim 3 \times 10^8$ light years: On such scales one can regard galaxies and clusters of galaxies as "atoms" of a homogeneous, isotropic "gas" that fills the universe.

On scales far larger than clusters of galaxies, our best information about homogeneity and isotropy comes from the cosmic microwave background radiation ("CMB"). As we shall see in Secs. 27.4 and 27.5 below, this radiation, emitted by hot, primordial gas long before galaxies formed, comes to us from distances of order 3×10^9 parsecs (1×10^{10} light years)—a scale 100 times larger than a rich cluster of galaxies (i.e., than a "supercluster"), and the largest scale on which observations can be made. Remarkably, this microwave radiation has a black body spectrum with a temperature that is the same, in all directions on the sky, to within about three parts in 10^5. This means that the temperature of the primordial gas was homogeneous on large scales to within this impressive accuracy.

These observational data justify a procedure in modeling the universe which was adopted by Einstein (1917) and others, in the early days of relativistic cosmology, with little more than philosophical justification: Like Einstein, we shall assume, as a zero-order approximation, that the universe is precisely homogeneous and isotropic. Later we shall briefly discuss galaxies and clusters of galaxies as first-order corrections to the homogeneous and isotropic structure.

Our assumption of homogeneity and isotropy can be stated more precisely as follows: *There exists a family of slices of simultaneity (3-dimensional spacelike hypersurfaces), which completely covers spacetime, with the special property that on a given slice of simultaneity* (i) *no two points are distinguishable from each other ("homogeneity"), and* (ii) *at a given point no one spatial direction is distinguishable from any other ("isotropy").*

Whenever, as here, the physical (geometrical) structure of a system has special symmetries, it is useful to introduce a coordinate system which shares and exhibits those symmetries. In the case of a spherical black hole, we introduced spherical coordinate systems. Here we shall introduce a coordinate system in which the homogeneous and isotropic hypersurfaces are slices of constant coordinate time t [Fig. (27.1)]:

Recall (Sec. 25.6) the special role of observers whose world lines are orthogonal to the homogeneous and isotropic hypersurfaces: if they define simultaneity locally (on small scales)

by the Einstein light-ray synchronization process, they will regard the homogeneous hypersurfaces as their own slices of simultaneity. Correspondingly, we shall call them the "homogeneous observers."

We shall define our time coordinate t to be equal to the proper time τ as measured by these homogeneous observers, with the arbitrary additive constant in t so adjusted that one of the homogeneous hypersurfaces (the "initial" hypersurface) has $t = 0$ everywhere on it. Stated differently, but equivalently, we select arbitrarily the initial hypersurface and set $t = 0$ throughout it; and we then define t along the world line of a homogeneous observer to be the proper time that the observer's clock has ticked since the observer passed through the initial hypersurface.

This definition of t has an important consequence: Since the points at which each of the observers pass through the initial hypersurface are all equivalent (all indistinguishable; "homogeneity"), the observers' subsequent explorations of the homogeneous universe must be indistinguishable; and, correspondingly, they must all reach any specific homogeneous hypersurface at the same proper time τ, and thence at the same coordinate time $t = \tau$. Thus, the hypersurfaces of constant coordinate time t are the same as the homogeneous hypersurfaces.

Turn, next, to the three spatial coordinates x^j. We shall define them in an arbitrary manner on the initial hypersurface, but shall insist that the homogeneous observers carry them forward (and backward) in time along their world lines, so that each homogeneous observer's world line is a curve of constant x^1, x^2, and x^3; cf. Fig. 27.1.

In this $\{t, x^j\}$ coordinate system the spacetime metric, described as a line element, will take the generic form

$$ds^2 = g_{tt}dt^2 + 2g_{tj}dt\,dx^j + g_{jk}dx^j dx^k \ . \tag{27.1}$$

Since x^j are constant along a homogeneous observer's world line, the basis vector $(\partial/\partial t)_{x^j}$ is tangent to the world line; and since t is constant in a homogeneous hypersurface, the basis vector $(\partial/\partial x^j)_t$ lies in the hypersurface. These facts, plus the orthogonality of the homogeneous observer's world line to the homogeneous hypersurface, imply that

$$g_{tj} \equiv \mathbf{g}\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial x^j}\right) = \frac{\partial}{\partial t}\cdot\frac{\partial}{\partial x^j} = 0 \ . \tag{27.2}$$

Moreover, since the proper time along a homogeneous observer's world line (line of constant x^j) is $d\tau = \sqrt{-g_{tt}dt^2}$, and since by construction dt is equal to that proper time, it must be that

$$g_{tt} = -1 \ . \tag{27.3}$$

By combining Eqs. (27.1)–(27.3) we obtain for the line element in our very special coordinate system

$$ds^2 = -dt^2 + g_{jk}dx^j dx^k \ . \tag{27.4}$$

Because our spatial coordinates, thus far, are arbitrary (i.e., they do not yet mold themselves in any special way to the homogeneous hypersurfaces), the spatial metric coefficients g_{jk} must be functions of the spatial coordinates x^i as well as of time t.

[*Side Remark*: Any coordinate system in which the line element takes the form (27.4) is called a *synchronous coordinate system*. This is true whether the hypersurfaces $t =$ const

are homogeneous and isotropic or not. The key features of synchronous coordinates are that they mold themselves to the world lines of a special family of observers in such a way that t is proper time along the family's world lines, and the slices of constant t are orthogonal to those world lines (and thus are light-ray-synchronization-defined simultaneities for those observers). Since the introduction of synchronous coordinates involves a specialization of precisely four metric coefficients ($g_{tt} = -1$, $g_{tj} = 0$), by a careful specialization of the four coordinates one can construct a synchronous coordinate system in any and every spacetime. On the other hand, one cannot pick an arbitrary family of observers and use them as the basis of synchronous coordinates: The observers must move freely; i.e., their world lines must be geodesics. This one can see by computing $u^{\alpha}{}_{;\beta}u^{\beta}$ for the vector field $\vec{u} \equiv \partial/\partial t$, which represents the 4-velocities of the synchronous coordinate system's special observers; a straightforward calculation [Exercise 27.1] gives $u^{\alpha}{}_{;\beta}u^{\beta} = 0$, in accord with geodesic motion. Thus it is that the static observers (observers with constant r, θ, ϕ) outside a black hole cannot be used as the foundation for synchronous coordinates. Those observers must accelerate to prevent themselves from falling into the hole; and correspondingly, the closest thing to a synchronous coordinate system that one can achieve, using for $x^j =$ const the world lines of the static observers, is the Schwarzschild coordinate system, which has $g_{tj} = 0$ (the slices of constant t are simultaneities as measured by the static observers), but $g_{tt} = -(1 - 2M/r) \neq -1$ (the proper time between two adjacent simultaneities depends on the radius at which the static observer resides).]

Returning to cosmology, we shall now specialize our spatial coordinates so they mold themselves nicely to the homogeneity and isotropy of the slices of constant t. One might have hoped this could be done in such a way that the metric coefficients are independent of all three coordinates x^j. Not so. The surface of a sphere is a good example in one lower dimension: it is homogeneous and isotropic, but the most symmetric coordinates one can find for it, spherical polar coordinates, produce a line element ${}^{(2)}ds^2 = a^2 d\theta^2 + a^2 \sin^2\theta d\phi^2$ with a metric coefficient $g_{\phi\phi} = a^2 \sin^2\theta$ that depends on θ. The deep, underlying reason is that the vector field $(\partial/\partial\phi)_\theta$ that "generates" rotations about the polar axis (z-axis) does not commute with the vector field that generates rotations about any other axis; and, correspondingly, those two vector fields cannot simultaneously be made the basis vectors of any coordinate system, and the metric coefficients cannot be made independent of two angular coordinates simultaneously. (For further detail see Secs. 25.2 and 25.3 of MTW, and especially Exercise 25.8.)

Similarly, on our cosmological homogeneous hypersurfaces the most symmetric coordinate system possible entails metric coefficients that are independent of only one coordinate, not all three. In order to construct that most-symmetric coordinate system, we choose arbitrarily on the hypersurface $t =$ const an origin of coordinates. Isotropy about that origin (all directions indistinguishable) is equivalent to spherical symmetry, which motivates our introducing spherical polar coordinates θ, ϕ, and a radial coordinate that we shall call χ. In this coordinate system the line element of the hypersurface will take the form

$$a^2[d\chi^2 + \Sigma^2(d\theta^2 + \sin^2\theta d\phi^2)] \,, \tag{27.5}$$

where a multiplicative constant (*scale factor*) a has been factored out for future convenience (it could equally well have been absorbed into χ and Σ), and where Σ is an unknown

function of the radial coordinate χ. Correspondingly, the 4-dimensional line element of spacetime (27.4) will take the form

$$\boxed{ds^2 = -dt^2 + a^2[d\chi^2 + \Sigma^2(d\theta^2 + \sin^2\theta d\phi^2)] \ ,} \tag{27.6}$$

where a is now a function of time t (i.e., it varies from hypersurface to hypersurface).

Our next task is to figure out what functions $\Sigma(\chi)$ are compatible with homogeneity and isotropy of the hypersurfaces. There are elegant, group-theoretic ways to figure this out; see, e.g., Ryan and Shepley (1975). A more straightforward but tedious way is to note that, because the 3-dimensional Riemann curvature tensor of the hypersurface must be homogeneous and isotropic, it must be algebraically expressible in terms of (*i*) constants, and the only tensors that pick out no preferred locations or directions: (*ii*) the metric tensor g_{jk} and (*iii*) the Levi-Civita tensor ϵ_{ijk}. Trial and error shows that the only combination of these three quantities which has the correct number of slots and the correct symmetries is

$$R_{ijkl} = K(g_{ik}g_{jl} - g_{il}g_{jk}) \ , \tag{27.7}$$

where K is a constant. By computing, for the 3-dimensional metric (27.5), the components of the 3-dimensional Riemann tensor and comparing with (27.7), one can show that there are three possibilities for the function $\Sigma(\chi)$ in the metric, and three corresponding possibilities for the constant K in the three-dimensional Riemann tensor. These three possibilities are nicely parametrized by a quantity k which takes on the values $+1$, 0, and -1:

$$\boxed{k = +1 \quad : \quad \Sigma = \sin\chi \ , \qquad K = \frac{k}{a^2} = +\frac{1}{a^2} \ ,} \tag{27.8a}$$

$$\boxed{k = 0 \quad : \quad \Sigma = \chi \ , \qquad\qquad K = \frac{k}{a^2} = 0 \ ,} \tag{27.8b}$$

$$\boxed{k = -1 \quad : \quad \Sigma = \sinh\chi \ , \quad K = \frac{k}{a^2} = -\frac{1}{a^2} \ .} \tag{27.8c}$$

We shall discuss each of these three possibilities in turn:

Closed universe [$k = +1$]: For $k = +1$ the geometry of the homogeneous hypersurfaces,

$${}^{(3)}ds^2 = a^2[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta d\phi^2)] \ , \tag{27.9}$$

is that of a 3-sphere, i.e., an ordinary sphere generalized to one higher dimension. One can verify this, for example, by showing (Ex. 27.2) that in a 4-dimensional Euclidean space with Cartesian coordinates (w, x, y, z) and line element

$${}^{(4)}ds^2 = dw^2 + dx^2 + dy^2 + dz^2 \ , \tag{27.10}$$

the 3-sphere

$$w^2 + x^2 + y^2 + z^2 = a^2 \tag{27.11}$$

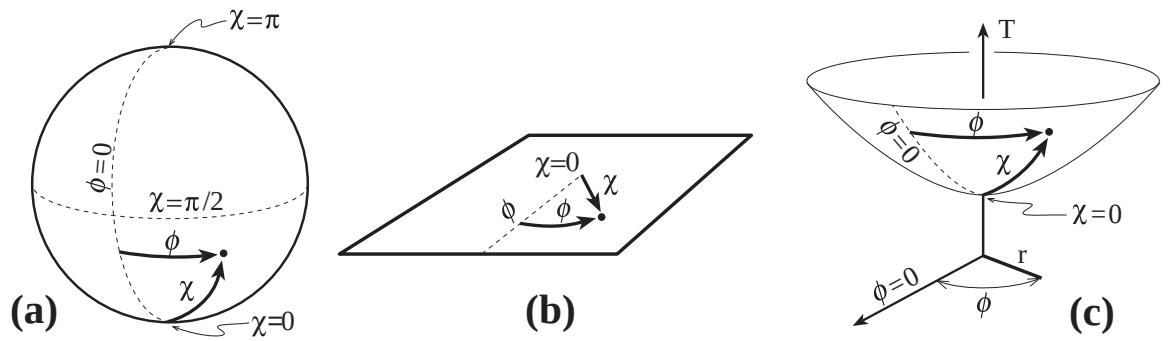


Fig. 27.2: Embedding diagrams for the homogeneous hypersurfaces of (a) a closed, $k = +1$, cosmological model; (b) a flat, $k = 0$, model; and (c) an open, $k = -1$ model.

has the same metric (27.9) as our cosmological, homogeneous hypersurface. Figure 27.2(a) is an embedding diagram for an equatorial slice, $\theta = \pi/2$, through the homogeneous hypersurface [2-geometry ${}^{(2)}ds^2 = a^2(d\chi^2 + \sin^2\chi d\phi^2)$; cf. Eq. (27.9)]. Of course, the embedded surface is a 2-sphere. As radius χ increases, the circumference $2\pi a \sin\chi$ around the spatial origin at first increases, then reaches a maximum at $\chi = \pi/2$, then decreases again to zero at $\chi = \pi$. Clearly, the homogeneous hypersurface is topologically "closed" and has a finite volume. For this reason a $k = +1$ cosmological model is often called a "closed universe". The universe's 3-volume, $V = 2\pi^2 a^3$ (Ex. 27.2).

Flat universe [$k = 0$]: For $k = 0$ the geometry of the homogeneous hypersurfaces,

$$ {}^{(3)}ds^2 = a^2[d\chi^2 + \chi^2(d\theta^2 + \sin^2\theta d\phi^2)] \,, \tag{27.12}$$

is that of a flat, 3-dimensional Euclidean space—as one can easily see by setting $r = a\chi$ and thereby converting (27.12) into the standard spherical-polar line element for Euclidean space. Correspondingly, this cosmological model is said to represent a "flat universe." Note, however, that this universe is only spatially flat: the Riemann curvature tensor of its 3-dimensional homogeneous hypersurfaces vanishes; but, as we shall discuss below, because of the time evolution of the expansion factor a, the Riemann curvature of the full 4-dimensional spacetime does not vanish. The volumes of the homogeneous hypersurfaces are infinite, so one cannot talk of the universe's total volume changing with time. However, the volume ΔV of a box in which resides a specific set of homogeneous observers will change as the expansion factor a evolves. For example, the volume could be a box with sides $\Delta\chi$, $\Delta\theta$, $\Delta\phi$, so

$$ \Delta V = \epsilon_{\chi\theta\phi}\Delta\chi\Delta\theta\Delta\phi = a^3\chi^2\sin\theta\Delta\chi\Delta\theta\Delta\phi \,, \tag{27.13}$$

where $\epsilon_{\chi\theta\phi}$ is a component of the Levi Civita tensor.

Open universe [$k = -1$]: For $k = -1$ the geometry of the homogeneous hypersurfaces,

$$ {}^{(3)}ds^2 = a^2[d\chi^2 + \sinh^2\chi(d\theta^2 + \sin^2\theta d\phi^2)] \,, \tag{27.14}$$

is different from geometries with which we have ordinary experience: The equatorial plane $\theta = \pi/2$ is a 2-surface whose circumference $2\pi a \sinh\chi$ increases with growing radius $a\chi$ faster than is permitted for any 2-surface that can ever reside in a 3-dimensional Euclidean space:

$$ \frac{d(\text{circumference})}{d(\text{radius})} = 2\pi\cosh[(\text{radius})/a] > 2\pi \,. \tag{27.15}$$

Correspondingly, any attempt to embed that equatorial plane in a Euclidean 3-space is doomed to failure. As an alternative, we can embed it in a flat, Minkowski 3-space with line element

$$^{(3)}ds^2 = -dT^2 + dr^2 + r^2 d\phi^2 \ . \tag{27.16}$$

The result is the hyperboloid of revolution,

$$T^2 - r^2 = a^2 \ , \tag{27.17}$$

which is shown pictorially in Fig. 27.2(c). By analogy it is reasonable to expect, and one easily can verify, that the full homogeneous hypersurface [metric (27.14)] has the same 3-geometry as the 3-dimensional hyperboloid

$$T^2 - r^2 = a^2 \tag{27.18}$$

in the 4-dimensional Minkowski space

$$^{(4)}ds^2 = -dT^2 + dr^2 + r^2(d\theta^2 + \sin^2 d\phi^2) \ . \tag{27.19}$$

That this hyperboloid is, indeed, homogeneous and isotropic one can show by verifying that Lorentz transformations in the T, r, θ, ϕ 4-space can move any desired point on the hyperboloid into the origin, and can rotate the hyperboloid about the origin by an arbitrary angle. Note that the T, r, θ, ϕ space has no relationship whatsoever to the physical spacetime of our homogeneous, isotropic universe. It merely happens that both spaces possess 3-dimensional hypersurfaces with the same 3-geometry (27.14). Because these hypersurfaces are topologically open, with infinite volume, the $k = -1$ cosmological model is often called an "open universe."

[*Side remark*: Although homogeneity and isotropy force the cosmological model's hypersurfaces to have one of the three metrics (27.9), (27.12), (27.14), the topologies of those hypersurfaces need not be the obvious ones adopted and described above. For example, a flat model could have a closed topology with finite volume rather than an open topology with infinite volume. This could come about if, for example, in a Cartesian coordinate system $\{x = \chi \sin\theta \cos\phi,\ y = \chi \sin\theta \sin\phi,\ z = \chi \cos\theta\}$ the 2-surface $x = -L/2$ were identical to $x = +L/2$ (so x, like ϕ in spherical polar coordinates, is periodic), and if similarly $y = -L/2$ were identical to $y = +L/2$ and $z = -L/2$ were identical to $z = +L/2$. The resulting universe would have volume a^3L^3; and if one were to travel outward far enough, one would find oneself (as on the surface of the Earth) returning to where one started. This and other unconventional choices for the topology of the standard cosmological models are kept in mind by cosmologists, just in case observational data someday should give evidence for them; but in the absence of such evidence, cosmologists assume the simplest choices of topology: those made above.]

Historically, the three possible choices for the geometry of a homogeneous, isotropic cosmological model were discovered by Alexander Alexandrovich Friedmann (1922), a Russian mathematician in Saint Petersburg; and, correspondingly, the specific solutions to the Einstein field equations which Friedmann constructed using those geometries are called *Friedmann cosmological models*. The first proof that these three choices are the only possibilities

for the geometry of a homogeneous, isotropic cosmological model was given independently by Howard Percy Robertson (1935), who was a professor at Caltech, and by Arthur Geoffrey Walker (1936), who was a young researcher at the Royal College of Science in London; and, correspondingly, the general line element (27.6) with $\Sigma = \sin\chi$, χ, or $\sinh\chi$ is called the *Robertson-Walker line element*.

****

EXERCISES

Exercise 27.1 *Example: The Observers of a Synchronous Coordinate System*
Show that any observer who is at rest in a synchronous coordinate system [Eq. (27.4)] is freely falling, i.e., moves along a geodesic of spacetime.

Exercise 27.2 *Example: The 3-Sphere Geometry of a Closed Universe*

(a) Show, by construction, that there exist coordinates χ, θ, ϕ on the 3-sphere (27.11) [which resides in the the 4-dimensional Euclidean space of Eq. (27.10)] such that the 3-sphere's line element assumes the same form (27.9) as that of a homogeneous hypersurface in a closed, $k = +1$, universe.

(b) Show that the total 3-volume of this 3-sphere is $V = 2\pi^2 a^3$.

****

27.3 The Stress-energy Tensor and the Einstein Field Equation

The expansion factor, $a(t)$, of our zero-order, homogeneous, isotropic cosmological model is governed by the Einstein field equation $\mathbf{G} = 8\pi\mathbf{T}$. In order to evaluate that equation we shall need a mathematical expression for the stress-energy tensor, $\mathbf{T}$.

We shall deduce an expression for $\mathbf{T}$ in two different ways: by mathematical arguments, and by physical considerations; and the two ways will give the same answer. Mathematically, we note that because the spacetime geometry is homogeneous and isotropic, the Einstein curvature tensor must be homogeneous and isotropic, and thence the Einstein equation forces the stress-energy tensor to be homogeneous and isotropic. In the local Lorentz frame of a homogeneous observer, which has basis vectors

$$\boxed{\vec{e}_{\hat{0}} = \frac{\partial}{\partial t}\ , \quad \vec{e}_{\hat{\chi}} = \frac{1}{a}\frac{\partial}{\partial \chi}\ , \quad \vec{e}_{\hat{\theta}} = \frac{1}{a\Sigma}\frac{\partial}{\partial \theta}\ , \quad \vec{e}_{\hat{\phi}} = \frac{1}{a\Sigma\sin\theta}\frac{\partial}{\partial \phi}\ ,} \tag{27.20}$$

the components of the stress-energy tensor are $T^{\hat{0}\hat{0}}$ =(energy density measured by homogeneous observer), $T^{\hat{0}\hat{j}}$ =(momentum density), $T^{\hat{j}\hat{k}}$ =(stress). Isotropy requires that the

momentum density (a 3-dimensional vector in the homogeneous hypersurface) vanish; if it did not vanish, its direction would break the isotropy. Isotropy also requires that the stress, a symmetric-second rank 3-tensor residing in the homogeneous hypersurface, not pick out any preferred directions; and this is possible if and only if the stress is proportional to the metric tensor of the hypersurface. Thus, the components of the stress-energy tensor in the observer's local Lorentz frame must have the form

$$\boxed{T^{\hat{0}\hat{0}} \equiv \rho \ , \quad T^{\hat{0}\hat{j}} = 0, \quad T^{\hat{j}\hat{k}}} = P\delta^{jk} \ , \tag{27.21}$$

where ρ is just a new notation for the energy density, and P is the isotropic pressure. This is precisely the stress-energy tensor of a *perfect fluid* which is at rest with respect to the homogeneous observer. Reexpressed in geometric, frame-independent form, this stress-energy tensor is

$$\mathbf{T} = (\rho + P)\vec{u} \otimes \vec{u} + P\mathbf{g} \ , \tag{27.22}$$

where $\vec{u}$ is the common 4-velocity of the fluid and of the homogeneous observers

$$\boxed{\vec{u} = \vec{e}_{\hat{0}} = \frac{\partial}{\partial t} \ .} \tag{27.23}$$

Physical considerations lead to this same stress-energy tensor: The desired stress-energy tensor must be that of our own universe, coarse-grain-averaged over scales large compared to a cluster of galaxies, i.e., averaged over scales $\sim 10^8$ parsecs. The contributors to that stress-energy tensor will be (*i*) the galaxy clusters themselves, which like the atoms of a gas will produce a perfect-fluid stress-energy with ρ equal to their smeared-out mass density and P equal to 1/3 times ρ times their mean square velocity relative to the homogeneous observers; (*ii*) the intercluster gas, which (one can convince oneself by astrophysical and observational arguments) is a perfect fluid, nearly at rest in the frame of the homogeneous observers; (*iii*) the cosmic microwave radiation, which, being highly isotropic, has the stress-energy tensor of a perfect fluid with rest frame the same as that of the homogeneous observers; (*iv*) as-yet undetected cosmological backgrounds of other fundamental particles such as neutrinos, gravitons, axions, neutralinos, ..., which are expected on theoretical grounds to be homogeneous and isotropic when coarse-grain averaged, with the same rest frame as the homogeneous observers; and (v) a possibly nonzero stress-energy tensor of the vacuum, which we shall discuss in Sec. 27.4 below, and which also has the perfect-fluid form . Thus, all the contributors are perfect fluids, and their energy densities and pressures add up to give a stress-energy tensor of the form (27.22).

As in our analysis of relativistic stars (Sec. 25.3), so also here, before evaluating the Einstein field equation we shall study the local law of conservation of 4-momentum, $\vec{\nabla}\cdot\mathbf{T} = 0$. (That conservation law is always easier to evaluate than the field equation, and by virtue of the contracted Bianchi identity it is equivalent to some combination of components of the field equation.)

The quantity $\vec{\nabla} \cdot \mathbf{T}$, which appears in the law of 4-momentum conservation, is a vector. Since $\mathbf{T}$ has already been forced to be spatially isotropic, the spatial, 3-vector part of $\vec{\nabla} \cdot \mathbf{T}$, i.e., the projection of this quantity into a homogeneous hypersurface, is guaranteed already to

vanish. Thus, only the projection orthogonal to the hypersurface, i.e., along $\vec{e}_{\hat{0}} = \vec{u} = \partial/\partial t$, will give us any information. This projection is viewed by a homogeneous observer, or equivalently by the perfect fluid, as the law of energy conservation. Evaluation of it, i.e., computation of $T_{\hat{0}}{}^{\hat{\mu}}{}_{;\hat{\mu}} = 0$ with $\mathbf{T}$ given by (27.21) and the metric given by (27.6), yields (Exercise 27.3)

$$\boxed{\frac{d(\rho a^3)}{dt} = -P\frac{da^3}{dt}\ .} \tag{27.24}$$

This is precisely the first law of thermodynamics for a perfect fluid, as one can see by the following calculation: Imagine a rectangular parallelopiped of fluid contained in the spatial region between χ and $\chi + \Delta\chi$, between θ and $\theta + \Delta\theta$, and between ϕ and $\phi + \Delta\phi$. As time passes the "walls" of this parallelopiped remain fixed relative to the homogeneous observers (since the walls and the observers both keep x^j fixed as t passes), and correspondingly the walls remain fixed in the fluid's rest frame. The volume of this fluid element is $\Delta V = a^3\Sigma^2 \sin\theta \Delta\chi\Delta\theta\Delta\phi$, and the total mass-energy contained in it is $E = \rho V$. Correspondingly, the first law of thermodynamics for the fluid element, $dE/dt = -PdV/dt$ says

$$\frac{\partial(\rho a^3\Sigma^2 \sin\theta\Delta\chi\Delta\theta\Delta\phi)}{\partial t} = -P\frac{\partial(a^3\Sigma^2 \sin\theta\Delta\chi\Delta\theta\Delta\phi)}{\partial t}\ . \tag{27.25}$$

By dividing out the coordinate volume $\Sigma^2 \sin\theta\Delta\chi\Delta\theta\Delta\phi$ (which is time independent), and then replacing the partial derivative by an ordinary derivative (because ρ and a depend only on t), we obtain the local law of energy conservation (27.24).

The fact that the local law of energy conservation, $T_{\hat{0}}{}^{\hat{\mu}}{}_{;\hat{\mu}}$ is identical to the first law of thermodynamics should not be surprising. Into our stress-energy tensor we put only the contribution of a perfect fluid, so energy conservation for it, in the fluid's local rest frame, must reduce to energy conservation for a perfect fluid, which *is* the first law of thermodynamics. If we had put other contributions into the stress-energy tensor, we would have obtained from energy conservation corresponding contributions to the first law; for example (as we saw in Part IV, when we studied fluid mechanics), if we had put viscous stresses into the stress-energy tensor, we would have obtained the first law in the form $d(\rho V) = -PdV + TdS$, including an explicit expression for the entropy increase dS due to viscous heating.

Turn, next, to the components of the Einstein equation $\mathbf{G} = 8\pi\mathbf{T}$. Because the metric has already been forced to be homogeneous and isotropic, the Einstein tensor is guaranteed already to have the homogeneous, isotropic form $G^{\hat{0}\hat{0}} \neq 0$, $G^{\hat{0}\hat{j}} = 0$, $G^{\hat{j}\hat{k}} \propto \delta^{jk}$, i.e., the same form as the stress-energy tensor (27.21). Correspondingly, there are only two nontrivial components of the Einstein field equation, the time-time component and the isotropic (proportional to δ^{jk}) space-space component. Moreover, the contracted Bianchi identity guarantees that some combination of these two components will be equivalent to our nontrivial law of energy conservation, thereby leaving only one new piece of information to be extracted from the Einstein equation. We shall extract that information from the time-time component, $G^{\hat{0}\hat{0}} = 8\pi T^{\hat{0}\hat{0}}$. A straightforward but tedious evaluation of $G^{\hat{0}\hat{0}} = G^{tt}$ for the Robertson-Walker line element (27.6), and insertion into the field equation along with

$T^{\hat{0}\hat{0}} = T^{tt} = \rho$ gives

$$\boxed{\left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2} = \frac{8\pi}{3}\rho\ ,} \tag{27.26}$$

where the dot represents a derivative with respect to the homogeneous observers' proper time t.

To verify that no errors have been made, one can evaluate the remaining nontrivial component of the field equation, $G^{\hat{x}\hat{x}} = 8\pi T^{\hat{x}\hat{x}}$ (or the $\hat{\theta}\hat{\theta}$ or $\hat{\phi}\hat{\phi}$ component; they are all equivalent since $G^{\hat{j}\hat{k}}$ and $T^{\hat{j}\hat{k}}$ are both proportional to δ^{jk}). The result,

$$2\frac{\ddot{a}}{a} + \left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2} = -8\pi P\ , \tag{27.27}$$

is, as expected, a consequence of the first of the Einstein components (27.26) together with the law of energy conservation (27.24): by differentiating (27.26) and then using (27.24) to eliminate $\dot{\rho}$, one obtains (27.27).

The task of computing the time evolution of our zero-order cosmological model now takes the following form: (*i*) Specify an equation of state

$$P = P(\rho) \tag{27.28}$$

for the cosmological perfect fluid; (*ii*) integrate the first law of thermodynamics

$$\frac{d\rho}{da} = -3\frac{(\rho + P)}{a} \tag{27.29}$$

[Eq. (27.24), rearranged] to obtain the density ρ and [via Eq. (27.28)] the pressure P as functions of the expansion factor a; (*iii*) evolve the expansion factor forward in time using the field equation

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2} = \frac{8\pi}{3}\rho \tag{27.30}$$

[Eq. (27.26)].

****

EXERCISES

Exercise 27.3 *Practice: Energy Conservation for a Perfect Fluid*

Consider a perfect fluid, with the standard stress-energy tensor $T^{\alpha\beta} = (\rho + P)u^\alpha u^\beta + Pg^{\alpha\beta}$. Assume that the fluid resides in an arbitrary spacetime, not necessarily our homogeneous, isotropic cosmological model.

(a) Explain why the law of energy conservation, as seen by the fluid, is given by $u_\alpha T^{\alpha\beta}{}_{;\beta} = 0$.

(b) Show that this law of energy conservation reduces to

$$\frac{d\rho}{d\tau} = -(\rho + P)\vec{\nabla} \cdot \vec{u} \,, \tag{27.31}$$

where τ is proper time as measured by the fluid.

(c) Show that for a fluid at rest with respect to the homogeneous observers in our homogeneous, isotropic cosmological model, (27.31) reduces to the first law of thermodynamics (27.24). *Note:* as a tool in this calculation, you might want to derive and use the following formulas, which are valid in any *coordinate* basis:

$$\Gamma^{\alpha}{}_{\mu\alpha} = \frac{1}{\sqrt{-g}}(\sqrt{-g})_{,\mu} \,, \quad A^{\alpha}{}_{;\alpha} = \frac{1}{\sqrt{-g}}(\sqrt{-g}A^{\alpha})_{,\alpha} \,. \tag{27.32}$$

Here, g denotes the determinant of the covariant components of the metric.

$$g \equiv \det||g_{ij}|| \,. \tag{27.33}$$

****

27.4 Evolution of the Universe

27.4.1 Constituents of the universe: Cold matter, radiation, and dark energy

The evolution of our zero-order cosmological model is highly dependent on the equation of state $P(\rho)$; and that equation of state, in turn, depends on the types of matter and fields that fill the universe—i.e., the universe's *constituents*.

The constituents can be divided into three classes: (i) *Cold matter*, i.e. material whose pressure is neglible compared to its total density of mass-energy so the equation of state can be idealized as $P_M = 0$ (subscript M for "matter"). The cold matter includes the *baryonic matter* of which people, planets, stars, galaxies, and intergalactic gas are made, as well as so-called *cold, dark matter* which is known to exist in profusion and might be predominantly fundamental particles (e.g. axions or neutralinos). (ii) *Radiation*, i.e. material with equation of state $P_R = \rho_R/3$. This includes the CMB (primordial photons), primordial gravitons, primordial neutrinos when their temperatures exceed their rest masses, and other finite-rest-mass particles when the temperature is sufficiently high (i.e., very early in the universe). (iii) *Dark energy* (denoted by a subscript Λ for historical reasons described in Box 27.2), with very negative pressure, $P_\Lambda \lesssim -\frac{1}{2}\rho_\Lambda$. As we shall see in Sec. 27.5, observations give strong evidence that such matter is present today in profusion. We do not yet know for sure its nature or its equation of state, but the most likely candidate is a nonzero stress-energy tensor associated with the vacuum, for which the equation of state is $P_\Lambda = -\rho_\Lambda$.

27.4.2 The vacuum stress-energy tensor

Let us digress, briefly, to discuss the vacuum: The stress-energy tensors of quantum fields are formally divergent, and must be renormalized to make them finite. In the early decades of quantum field theory, it was assumed that (in the absence of boundaries such as those of highly electrically conducting plates) the renormalized vacuum stress-energy tensor $\mathbf{T}_{\rm vac}$ would vanish. In 1968 Yakov Borisovich Zel'dovich initiated speculations that $\mathbf{T}_{\rm vac}$ might, in fact, be nonzero, and those speculations became fashionable in the 1980s in connection with inflationary models for the very early universe (Sec. 27.7). It was presumed in the 1980s and 90s that a phase transition in the early universe had driven quantum fields into a new vacuum state, for which $\mathbf{T}_{\rm vac}$ vanishes; but in the late 1990s, much to physicists' amazement, observational evidence began to mount that our universe today is filled with a profusion of "dark energy", perhaps in the form of a nonzero $\mathbf{T}_{\rm vac}$; and by 2000 that evidence was compellingly strong.

If $\mathbf{T}_{\rm vac}$ is nonzero, what form can it take? It must be a second-rank symmetric tensor, and it would be very surprising if that tensor broke the local homogeneity and isotropy of spacetime or picked out any preferred reference frames. In order not to break those local symmetries, $\mathbf{T}_{\rm vac}$ must be proportional to the metric tensor, with its proportionality factor independent of location in spacetime:

$$\boxed{\mathbf{T}_{\rm vac} = -\rho_\Lambda \mathbf{g} \ , \quad \text{i.e.,} \quad T^{\hat{0}\hat{0}}_{\rm vac} = \rho_\Lambda \ , \quad T^{\hat{j}\hat{k}}_{\rm vac} = -\rho_\Lambda \ .} \tag{27.34}$$

This is a perfect-fluid equation of state with $P_\Lambda = -\rho_\Lambda$.

If there is no significant transfer of energy or momentum between the vacuum and other constituents of the universe, then energy-momentum conservation requires that $\mathbf{T}_{\rm vac}$ be divergence free. This, together with the vanishing divergence of the metric tensor, implies that ρ_Λ is constant, despite the expansion of the universe! This constancy can be understood in terms of the first law of thermodynamics (27.24): As the universe expands, the expansion does work against the vacuum's tension $-P_\Lambda = \rho_\Lambda$ at just the right rate as to replenish the vacuum's otherwise-decreasing energy density. For further insight into $\mathbf{T}_{\rm vac}$, see Box 27.2.

27.4.3 Evolution of the densities

In order to integrate the Einstein equation backward in time and thereby deduce the universe's past evolution, we need to know how much radiation, cold matter, and dark energy the universe contains today. Those amounts are generally expressed as fractions of the *critical energy density* that marks the dividing line between a closed universe and an open one. By asking that k/a^2 be zero, we find from the Einstein equation (27.26) that

$$\boxed{\rho_{\rm crit} = \frac{3}{8\pi}\left(\frac{\dot{a}_o}{a_o}\right)^2 \simeq 9 \times 10^{-30} {\rm g/cm}^3 \ .} \tag{27.35}$$

Here we have used a numerical value of $\dot{a}_o/a_o$ (the value of $\dot{a}/a$ today) that is discussed in Sec. 27.5 below. The energy density today in units of the critical density is denoted

$$\boxed{\Omega \equiv \frac{\rho_o}{\rho_{\rm crit}} \ ,} \tag{27.36}$$

Box 27.2
The Cosmological Constant

Soon after formulating general relativity, Einstein discovered that his field equation, together with then plausible equations of state $P(\rho)$, demanded that the universe be either expanding or contracting; it could not be static. Firmly gripped by the mind-set of his era, Einstein regarded a nonstatic universe as implausible, and thus thought his field equation incompatible with the way the universe ought to behave; and so he modified his field equation. There were very few possibilities for the modification, since (*i*) it seemed clear that the source of curvature should still be the stress-energy tensor, and accordingly the field equations should say $\mathbf{E} = 8\pi\mathbf{T}$ where $\mathbf{E}$ is a tensor (evidently *not* the Einstein tensor) which characterizes gravity; and (*ii*) in order that the field equation leave four of the metric coefficients arbitrary (so they could be adjusted by coordinate freedom) the tensor $\mathbf{E}$ should have an automatically vanishing divergence. Of the various possibilities for $\mathbf{E}$, one stood out as far simpler than all the rest: $\mathbf{E} = \mathbf{G} + \Lambda\mathbf{g}$, where Λ is a "cosmological constant." To Einstein's great satisfaction, by choosing Λ negative he was able, from his modified field equation

$$\mathbf{G} + \Lambda\mathbf{g} = 8\pi\mathbf{T} \ , \tag{1}$$

to obtain a forever-static, homogeneous and isotropic cosmological model; see Ex. 27.5.

In 1929 Edwin Powell Hubble (1929), at the Mount Wilson Observatory, discovered that the universe was expanding. What a shock this was to Einstein! After visiting Mount Wilson and discussing Hubble's observations with him, Einstein (1931) formally renounced the cosmological constant and returned to his original, 1915, field equation $\mathbf{G} = 8\pi\mathbf{T}$. In his later years, Einstein described the cosmological constant as the greatest mistake of his life. Had he stuck to his original field equation, the expansion of the universe might have been regarded as the greatest of all the predictions made by his general relativity.

Remarkably, the cosmological-constant term $\Lambda\mathbf{g}$ in Einstein's modified field equation is identical to the modern vacuum contribution to the stress-energy tensor. More specifically, if we define $\rho_\Lambda \equiv \Lambda/8\pi$ so $T_{\rm vac} = -\rho_\Lambda\mathbf{g} = -(\Lambda/8\pi)\mathbf{g}$, then $\mathbf{G} + \Lambda\mathbf{g} = 8\pi\mathbf{T}$ becomes $\mathbf{G} = 8\pi(\mathbf{T} + \mathbf{T}_{\rm vac})$. Thus, the modern conclusion that there might be a nonzero vacuum stress-energy tensor is actually a return to Einstein's modified field equation.

It is not at all clear whether the universe's dark energy has the equation of state $P_\Lambda = -\rho_\Lambda$ and thus is the vacuum stress-energy. Cosmologists' prejudice that it may be vacuum is built into their adoption of Einstein's cosmological constant notation Λ to denote the dark energy.

and observations give values

$$\boxed{\Omega_R \sim 10^{-4} \ , \quad \Omega_M \simeq 0.3 \ , \quad \Omega_\Lambda \simeq 0.7 \ , \quad \Omega \equiv \Omega_R + \Omega_M + \Omega_\Lambda \simeq 1 \ .} \tag{27.37}$$

We shall discuss these numbers and the observational error bars on them in Sec. 27.5.

The evolution of the universe could be influenced by energy transfer among its three constituents. However, that transfer was small during the epoch from $a/a_o \sim 10^{-9}$ to today; see Box 27.3. This means that the first law of thermodynamics (27.24) must hold true for each of the three constituents individually: $d(\rho a^3) = -P da^3$. By combining this law with the constituents' equations of state, $P_M = 0$, $P_R = \rho_R/3$, and (assuming the dark energy is vacuum) $P_\Lambda = -\rho_\Lambda$, we obtain

$$\boxed{\rho_R = \rho_{Ro}\frac{a_o^4}{a^4}, \quad \rho_M = \rho_{Mo}\frac{a_o^3}{a^3}\ , \quad \rho_\Lambda = \text{const}\ .} \tag{27.38}$$

These relations are plotted in Fig. 27.4 below, which we shall discuss later.

The qualitative evolution of our zero-order cosmological model is easily deduced by inserting Eqs. (27.38) into Einstein's equation (27.26) and rewriting the result in the standard form for the motion of a particle in a potential well:

$$\boxed{\frac{1}{2}\dot{a}^2 + V(a) = \frac{-k}{2}\ ,} \tag{27.39}$$

where

$$\boxed{V(a) = -\frac{4\pi}{3}a^2\rho = -\frac{4\pi}{3}\rho_{\rm crit}a_o^2\left(\Omega_R\frac{a_o^2}{a^2} + \Omega_M\frac{a_o}{a} + \Omega_\Lambda\frac{a^2}{a_o^2}\right)\ .} \tag{27.40}$$

Note that a/a_o is the ratio of the linear size of the universe at some time in the past, to the size of the universe today. Each volume element, comoving with the homogeneous observers, expands in length by a_o/a from then until now, and expands in volume by $(a_o/a)^3$.

The shape of the effective potential $V(a)$ is shown in Fig. 27.3: It increases monotonically from $-\infty$ at $a = 0$ to about $-4\rho_{\rm crit}a_o^2$ at $a/a_o \simeq 0.7$, and then, as the universe nears our own era, it begins decreasing. The universe is radiation dominated at $a/a_o \lesssim 10^{-4}$ (Fig. 27.4), it is cold-matter dominated between $a/a_o \sim 10^{-4}$ and $a/a_o \sim 1$, and the maxing-out of the effective potential and reversal to plunge is triggered by a modern-era ($a/a_o \sim 1$) transition to dark-energy dominance.

The implications of this effective potential for the past evolution of our universe should be clear from one's experience with particle-in-potential problems: The universe must have expanded at an ever decreasing rate $\dot{a}$ from an age $\sim$ a small fraction of second, when our equations of state became valid, until nearly the present epoch, $a/a_o \sim 0.4$, and then the universe's vacuum tension must have triggered an acceleration of the expansion.

It seems strange that the universe should switch over to acceleration in just the epoch that we are alive, rather than far earlier or far later or not at all. The reasons for this are unknown. It is a big surprise, revealed by recent observations.

If P_Λ/ρ_Λ is independent of time, then the universe's past evolution is *not* very sensitive to the precise value of P_Λ/ρ_Λ. For $P_\Lambda/\rho_\Lambda = -\frac{1}{2}$ (the least negative pressure allowed by current observations), as for $P_\Lambda/\rho_\Lambda = -1$ (vacuum), the dark energy begins to influence $V(a)$ significantly only in the modern era, and its influence is to accelerate the expansion in accord with observation. It is of no dynamical importance earlier.

However, nothing requires that P_Λ/ρ_Λ be constant. It is possible, in principle for P_Λ/ρ_Λ to evolve in a wide variety of ways that, in principle, could have had a strong influence on

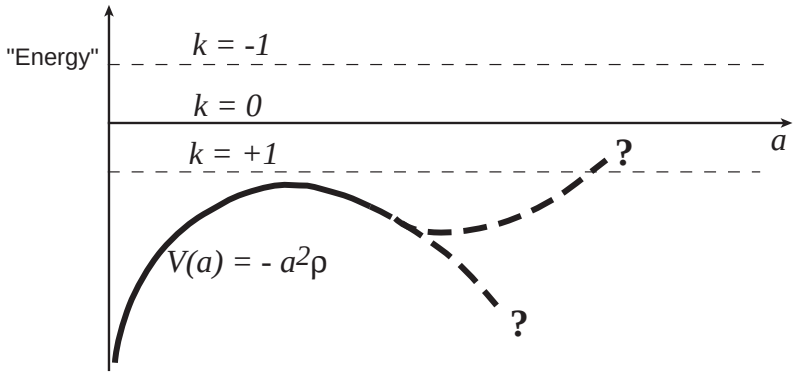


Fig. 27.3: The "particle-in-a-potential" depiction of the evolutionary equation (27.39) for the expansion factor a of the universe. Plotted horizontally is the expansion factor, which plays the role of the position of a particle. Plotted vertically is the "particle's" potential energy $V(a)$ (thick curve) and its total energy $-k/2$ (thin dotted line). The difference in height between the dotted line and the thick curve is the particle's kinetic energy $\frac{1}{2}\dot{a}^2$, which must be positive. The form of $V(a)$ in the past ($a \leq a_o$) is shown solid. The form in the future is unknown because we do not know the nature of the dark energy. If the form is that of the upper thick dashed curve, the universe may reach a maximum size and then recontract. If the form is that of the lower dashed curve, the univeral expansion will continue to accelerate forever.

the universe's early evolution. That this probably did *not* occur we know from observational data which show that the dark energy cannot have had a very significant influence on the universal expansion at several key epochs in the past: (i) during the nucleosynthesis of light elements when the universe was about 1 minute old, (ii) during recombination of the primordial plasma (conversion from ionization to neutrality) when the universe was about 10^6 years old, and (iii) during early stages of galaxy formation when the universe was about 1 billion years old. Nevertheless, we are so ignorant, today, of the precise nature of the dark energy, that we must be prepared for new surprises.

By contrast, the evolution of the dark energy in the future and the resulting evolution of the universe are unconstrained by observation and are unknown. Until we learn for sure the nature and dynamics of the dark energy, we cannot predict the universe's future evolution.

27.4.4 Evolution in time and redshift

Since the dark energy cannot have had a very important dynamical role in the past, we shall ignore it in the remainder of this section and shall idealize the universe as containing only cold matter, with density $\rho_M = \rho_{Mo}(a_o/a)^3$ and radiation with density $\rho_R = \rho_{Ro}(a_o/a)^4$.

The radiation includes the cosmic background photons (which today are in the microwave frequency band), plus gravitons, and plus those neutrinos whose rest masses are much less than their thermal energies. In order for the observed abundances of the light elements to agree with the theory of their nucleosynthesis, it is necessary that the neutrino and graviton contributions to ρ_R be less than or of order the photon contributions.

The photons were in thermal equilibrium with other forms of matter in the early universe and thus had a Planckian spectrum. Since black-body radiation has energy density $\rho_R \propto T_R^4$

Box 27.3

Interaction of Radiation and Matter

In the present epoch there is negligible radiation/matter interaction: the radiation propagates freely, with negligible absorption by interstellar or intergalactic gas. However, much earlier, when the matter was much denser and the radiation much hotter than today and galaxies had not yet formed, the interaction must have been so strong as to keep the photons and matter in thermodynamic equilibrium with each other. In this early epoch the matter temperature, left to its own devices, would have liked to drop as $1/\text{Volume}^{2/3}$, i.e., as $1/a^2$, while the radiation temperature, left to its own devices would have dropped as $1/\text{Volume}^{1/3}$, i.e., as $1/a$. To keep their temperatures equal, the photons had to feed energy into matter. This feeding was not at all a serious drain on the photons' energy supply, however: Today the ratio of the number density of background photons to the number density of baryons (i.e., protons and neutrons) is

$$\frac{n_{Ro}}{n_{Mo}} = \frac{a_R {T_{Ro}}^4/(2.8kT_{Ro})}{\rho_{mo}/m_p} \sim 10^8 \ , \tag{1}$$

where m_p is the proton mass, k is Boltzman's constant and $2.8kT_{Ro}$ is the average energy of a black-body photon at the CMB temperature $T_{Ro} = 2.728$ K. Because, aside from a factor of order unity, this ratio is the entropy per baryon in the universe (Chaps. 3 and 4), and because the entropy per baryon was (nearly) conserved during the (nearly) adiabatic expansion of the universe, this ratio was about the same in the early era of thermal equilibrium as it is today. Since the specific heat per photon and that per baryon are both of order Boltzman's constant k, the specific heat of a unit volume of background radiation in the early era exceeded that of a unit volume of matter by eight orders of magnitude. Consequently, the radiation could keep the matter's temperature equal to its own with little effort; and accordingly, despite their interaction, the radiation by itself satisfied energy conservation to high accuracy.

This remained true, going backward in time, until the temperature reached $T \sim \frac{1}{5}m_e/k \sim 10^9$ K, at which point electron-positron pairs formed in profusion and sucked roughly half the photon energy density out of the photons. This pair formation, going backward in time, or pair annihilation going forward, occurred when the universe was several 10's of seconds old (Fig. 27.4) and can be regarded as converting one form of radiation (photons) into another (relativistic pairs). Going further backward in time, at $T \sim m_p/k \sim 10^{13}$ K, the neutrons and protons (baryons) became relativistic, so cold matter ceased to exist—which means that, going forward in time, cold matter formed at $T \sim m_p/k \sim 10^{13}$ K.

As is shown in the text, the dark energy only became significant in the modern era; so its interaction with cold matter and radiation (if any, and there presumably is very little) cannot have been important during the universe's past evolution.

and the density decreases with expansion as $\rho_R \propto 1/a^4$, the photon temperature must have been redshifted during this early era as $T_R \propto 1/a$. When the temperature dropped below $\sim 10^4$ K, the electrons dropped into bound states around atomic nuclei, making the matter neutral, and its opacity negligible, so the photons were liberated from interaction with the matter and began to propagate freely. Kinetic theory (Box 27.4) tells us that during this free propagation, the photons retained their Planckian spectrum, and their temperature continued to be redshifted as $T_R \propto 1/a$. In accord with this prediction, the spectrum of the photons today is measured to be Planckian to very high accuracy; its temperature is $T_{Ro} = 2.728$ K, corresponding to a photon energy density today $\rho_{\gamma o} \sim 5 \times 10^{-34}$ g cm^{-3}. Adding to this the neutrino and graviton energy densities, we conclude that $\rho_{Ro} \sim 10^{-33}$ g cm^{-3}. By contrast, the matter density today is $\rho_{Mo} \simeq 3 \times 10^{-30}$ g cm^{-3}.

To recapitulate: the matter and radiation densities and temperatures must have evolved as

$$\rho_M = \rho_{Mo}\frac{a_o^3}{a^3}\ , \quad \rho_R = \rho_{Ro}\frac{a_o^4}{a^4}\ , \quad \boxed{T_R = T_{Ro}\frac{a_o}{a}\ ,} \tag{27.41}$$

throughout the entire epoch from $a/a_o \sim 3 \times 10^{-13}$ until today.

The density and temperature evolutions (27.41) are depicted as functions of the universe's expansion factor a/a_o in Fig. 27.4. A second way to express the evolution is in terms of *cosmological redshift*: Imagine photons emitted in some chosen spectral line of some chosen type of atom (e.g., the Lyman alpha line of atomic hydrogen), at some chosen epoch during the universe's evolution. Let the atoms be at rest in the mean rest frame of the matter and radiation, i.e., in the rest frame of a homogeneous observer, so they move orthogonally to the homogeneous hypersurfaces. Focus attention on specific photons that manage to propagate to Earth without any interaction whatsover with matter. Then they will arrive with a wavelength, as measured on Earth today, which is much larger than that with which they were emitted: The expansion of the universe has increased their wavelength, i.e., has redshifted them. As is shown in Exercise 27.6 below, if the expansion factor was a at the time of their emission, and if their wavelength at emission as measured by the emitter was λ, then at reception on Earth as measured by an astronomer, they will have wavelength λ_o given by

$$\frac{\lambda_o}{\lambda} = \frac{a_o}{a}\ ; \tag{27.42}$$

i.e., *the photons' wavelength is redshifted in direct proportion to the expansion factor of the universe*. It is conventional to speak of the redshift z as not the ratio of the wavelength today to that when emitted, but rather as the fractional change in wavelength, so

$$\boxed{z \equiv \frac{\lambda_o - \lambda}{\lambda} = \frac{a_o}{a} - 1\ .} \tag{27.43}$$

In Fig. 27.4's depiction of the density evolution of the universe, the horizontal axis at the bottom is marked off in units of z.

It is also instructive to examine the density evolution in terms of proper time, t, as measured in the mean rest frame of the matter and radiation; i.e., as measured by clocks carried by the homogeneous observers. At redshifts $z > \rho_{Mo}/\rho_{Ro} \sim 5000$, when the energy

Box 27.4

Kinetic Theory of Photons in General Relativity

The kinetic theory of photons and other particles (Chap. 2) can be lifted from special relativity into general relativity using the equivalence principle:

In any local Lorentz frame in curved spacetime, the number density in phase space is given by the special relativity expression (2.3): $\mathcal{N}(\mathcal{P},\vec{p}) = dN/d\mathcal{V}_x d\mathcal{V}_p$. Here $\mathcal{P}$ is the location of the observer in spacetime, $\vec{p}$ is the momentum of some chosen "fiducial" photon, $d\mathcal{V}_x$ is a small 3-volume at $\mathcal{P}$ in the physical space of the observer's local Lorentz frame, $d\mathcal{V}_p$ is a small 3-volume in the momentum space of the observer's local Lorentz frame, centered on $\vec{p}$, and dN is the number of photons in $d\mathcal{V}_x$ and $d\mathcal{V}_p$. For a homogeneous observer, we can choose $d\mathcal{V}_x = a^3\Sigma^2 \sin\theta\, d\chi\, d\theta\, d\phi$, $d\mathcal{V}_p = dp^{\hat{\chi}} dp^{\hat{\theta}} dp^{\hat{\phi}}$, where the hats denote components on the unit vectors $\vec{e}_{\hat{\chi}}$, $\vec{e}_{\hat{\theta}}$, $\vec{e}_{\hat{\phi}}$.

The equivalence principle guarantees that, just as in flat spacetime, so also in curved spacetime, (i) the number density in phase space, $\mathcal{N}$, is independent of the velocity of the local Lorentz frame in which it is measured (with all the frames presumed to be passing through the event $\mathcal{P}$); and (ii) if the photons do not interact with matter, then $\mathcal{N}$ is constant along the world line of any chosen (fiducial) photon as it moves through spacetime and as its 4-momentum $\vec{p}$ evolves in the free-particle (geodesic) manner. [In asserting this constancy of $\mathcal{N}$, one must examine carefully the issue of curvature coupling; Sec. 24.7. Because the volume element $d\mathcal{V}_x$ involved in the definition of $\mathcal{N}$ has some finite, though tiny size, spacetime curvature will produce geodesic deviation between photons on opposite sides of $d\mathcal{V}_x$. One can show fairly easily, however, that this geodesic deviation merely skews the phase-space volume element along its momentum directions in a manner analogous to Fig. 2.6(b), while leaving the product $d\mathcal{V}_x\, d\mathcal{V}_p$ fixed and thereby leaving $\mathcal{N}$ unchanged; cf. Sec. 2.7.]

The equivalence principle also guarantees that in curved spacetime, as in flat, the number density in phase space can be expressed in terms of the specific intensity I_ν and the frequency of the chosen photon ν (as measured in any local Lorentz frame): $\mathcal{N} = h^{-4} I_\nu/\nu^3$ [Eq. (2.18)]. If the spectrum is Planckian with temperature T as measured in this Lorentz frame, then $\mathcal{N}$ will have the form

$$\mathcal{N} = \frac{2}{h^3}\frac{1}{e^{h\nu/kT}-1}\ . \tag{1}$$

The Lorentz-invariance and conservation of $\mathcal{N}$, together with the fact that this $\mathcal{N}$ depends only on the ratio ν/T, implies that, (i) a spectrum that is Planckian in one reference frame will be Planckian in all reference frames, with the temperature T getting Doppler shifted in precisely the same manner as the photon frequency ν; and (ii) an initially Planckian spectrum will remain always Planckian (under free propagation), with its temperature experiencing the same cosmological redshift, gravitational redshift, or other redshift as the frequencies of its individual photons.

For the CMB as measured by homogeneous observers, the frequencies of individual photons get redshifted by the expansion as $\nu \propto 1/a$, so the photon temperature also gets redshifted as $T \propto 1/a$.

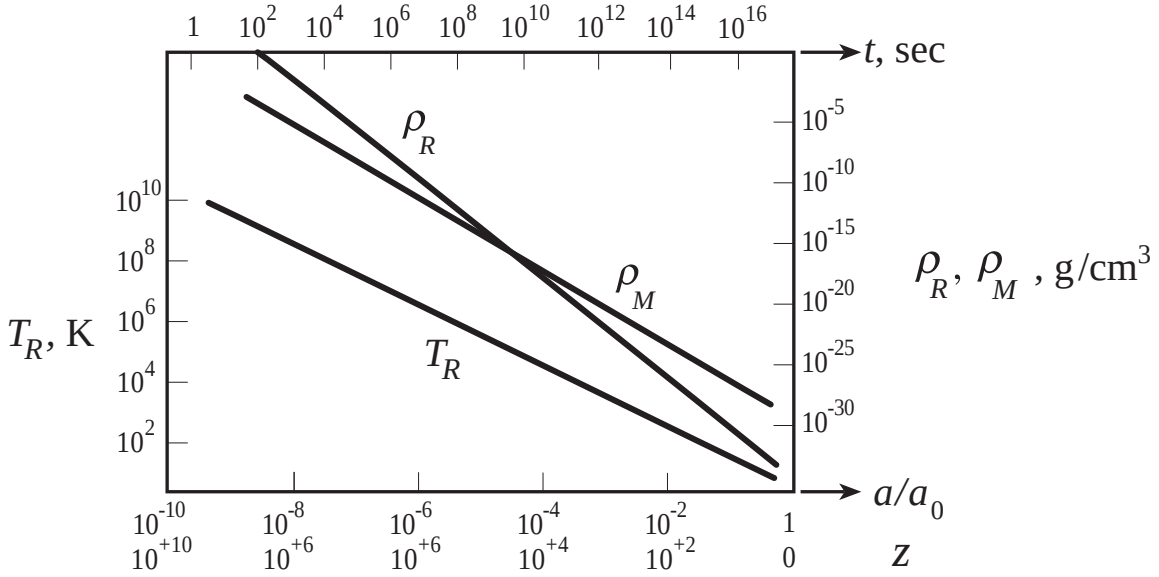


Fig. 27.4: The evolution of the total mass-energy densities ρ_M and ρ_R in matter and in radiation and the radiation's photon temperature T_R, as functions of the expansion factor a of the universe, the cosmological redshift z and the proper time t (in the mean rest frame of the matter and radiation) since the "big bang."

in radiation dominated over that in matter, a as a function of time was governed by the Einstein field equation

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2} = \frac{8\pi}{3}\rho_R = \frac{8\pi}{3}\rho_{Ro}\left(\frac{a_o}{a}\right)^4 \ . \tag{27.44}$$

[Eq. (27.30)]. As we shall see in Sec. 27.5 below, the present epoch of the universe's expansion is an early enough one that, if it is closed or open, the evolution has only recently begun to depart significantly from that associated with a flat, $k = 0$ model. Correspondingly, in the early, radiation-dominated era, the evolution was independent of k to high precision, i.e., the factor k/a^2 in the evolution equation was negligible. Ignoring that factor and integrating Eq. (27.30), then setting $\rho \simeq \rho_R = \rho_{Ro}(a_o/a)^4$, we obtain

$$\boxed{\rho \simeq \rho_R = \frac{3}{32\pi t^2} \ , \quad \frac{a}{a_o} = \left(\frac{32\pi}{3}\rho_{Ro}t^2\right)^{1/4} \quad \text{when } \frac{a}{a_o} < \frac{\rho_{Ro}}{\rho_{Mo}} \sim 3\times 10^{-4} \ .} \tag{27.45}$$

Here the origin of time, $t = 0$, is taken to be at the moment when the expansion of the universe began: the "big-bang."

This early, *radiation-dominated era* ended at a *cross-over time*

$$t_c = \left[\frac{3}{32\pi\rho_{Ro}}\left(\frac{\rho_{Ro}}{\rho_{Mo}}\right)^4\right]^{1/2} \sim \left[\frac{3}{32\pi\times 10^{-33}\mathrm{g/cm}^3}(3\times 10^{-4})^4\right]^{1/2} \times$$
$$\times \left[\frac{1\,\mathrm{g}}{0.742\times 10^{-28}\mathrm{cm}}\right]^{1/2} \times \frac{1\,\mathrm{year}}{0.946\times 10^{18}\mathrm{cm}} \sim 70,000\,\mathrm{years} \ . \tag{27.46}$$

In this calculation the first two factors on the second line are introduced to convert from geometrized units to conventional units. After the crossover time the solution to the Einstein equation is that for pressure-free matter. The precise details of the time evolution will depend on whether the universe is open, closed, or flat; but the three possibilities will agree up to the present epoch to within a few tens of per cent (see Sec. 27.5). Ignoring the differences between open, closed, and flat, we can adopt the $k = 0$, pressure-free evolution as given by Friedmann's flat model, Eqs. (27.48), (27.49), (27.50), and (27.51) —but with the origin of time adjusted to match onto the radiation-dominated solution (27.45) at the cross-over time:

$$\boxed{\rho \simeq \rho_M = \frac{1}{6\pi(t\,+\,t_c/3)^2}\ , \quad \frac{a}{a_o} = \left[6\pi\rho_{Mo}\left(t+\frac{t_c}{3}\right)^2\right]^{1/3} \quad \text{when} \quad \frac{a}{a_o} > \frac{\rho_{Ro}}{\rho_{Mo}} \sim 3\times 10^{-4}\ .} \tag{27.47}$$

The present age of the universe, as evaluated by setting $\rho_M = \rho_{Mo}$ in this formula and converting to cgs units, is of order 10^{10} years. We shall evaluate the age with higher precision in section 27.5.8 below. In Fig. 27.4's depiction of the evolution, the time t since the big bang, as computed from Eqs. (27.45), (27.46), and (27.47), is marked along the top axis.

27.4.5 Physical processes in the expanding universe

The evolution laws (27.41) and (27.45), (27.46), and (27.47) for ρ_M and ρ_R are a powerful foundation for deducing the main features of the physical evolution of the universe from the very early epoch, $a/a_o \sim 3\times 10^{-13}$, i.e. $z \sim 3\times 10^{12}$ and $t \sim 10^{-5}$ sec up to the present. For detailed, pedagogical presentations of those features see, e.g., Peebles (1971) and Zel'dovich and Novikov (1983). Here we shall just present a very short summary.

Some key physical events that one deduces during the evolution from $z = 3\times 10^{12}$ to the present, $z = 0$, are these:

(*i*) At redshift $z \sim 3\times 10^{12}$, baryon-antibaryon pairs annihilated and the thermal energies of neutrons and protons became much smaller than their rest-mass energies. This was the epoch of formation of baryonic cold matter. (*ii*) At redshifts $z \sim 10^9$ when the universe was of order a second old, the photons ceased being energetic enough to make electron-positron pairs; the pairs, which earlier had been plentiful, annihilated, feeding their energy into photons; and with their annihilation the primordial gas suddenly became transparent to neutrinos. Since then the neutrinos, born in thermodynamic equilibrium at $z > 10^9$, should have propagated freely.

(*iii*) At redshifts $z \sim 3\times 10^8$, when the universe was a few minutes old, ρ_R was roughly 1 g/cm^3, and the temperature was $T_R = T_M \sim 10^9$ K, nuclear burning took place. Going into this epoch of *primordial nucleosynthesis* the matter consisted of equal numbers of protons, neutrons, and electrons all in thermodynamic equilibrium with each other. Coming out, according to evolutionary calculations for the relevant nuclear reactions, it consisted of about 75 per cent protons (by mass), 25 per cent alpha particles (^{4}He nuclei), and tiny ($< 10^{-6}$), but observationally important amounts of deuterium, ^{3}He, lithium, beryllium, and boron. [The agreement of these predictions with observation constitutes strong evidence that cosmologists are on the right track in their deductions about the early universe.] All the elements heavier than boron were almost certainly made in stars when the universe was billions of years old.

(*iv*) At the redshift $z \sim 3000$, when the universe was about $70,000$ years old and $T_R \sim T_M$ was about 10^4 K, came the cross-over from radiation dominance to matter dominance; i.e., $\rho_R = \rho_M \sim 10^{-17}\text{g/cm}^3$.

(*v*) At the redshift $z \simeq 1090$, when the universe was $\simeq 380,000$ years old and its temperature had dropped to roughly 3000 K and its density to $\rho_M \sim 10^{-20}\text{g/cm}^3$, the electrons in the primordial plasma were captured by the protons and alpha particles to form neutral hydrogen and helium. Before this *epoch of recombination* the matter was highly ionized and opaque to radiation; afterward it was highly neutral and transparent.

(*vi*) Before recombination, if any matter tried to condense into stars or galaxies, it would get adiabatically heated as it condensed and the rising pressure of radiation trapped inside it would prevent the condensation from proceeding. After recombination, radiation was no longer trapped inside a growing condensation. Now, for the first time, stars, galaxies, and clusters of galaxies could begin to form. Measured anisotopies of the CMB, however, tells us that the size of the density fluctuations at recombination was $\Delta\rho_M/\rho_M \sim 10^{-4}$, which is just the right size to grow, by gravitational condensation, to $\Delta\rho/\rho \sim 1$ at $z \sim 10$. Thus it is that the *epoch of galaxy formation* probably began around a redshift $z \sim 10$, when the universe was already about two billion years old compared to its present age of roughly 14 billion years.

(*vi*) In galaxies such as ours there has been, since formation, a continuing history of stellar births, nucleosynthesis, and deaths. The unravelling of our Galaxy's nucleosynthesis history and the resulting understanding of the origin of all the elements heavier than boron, was achieved, in large measure in the 1950s, 60s and 70s, by nuclear astrophysicists under the leadership of Caltech's William A. Fowler. Our own sun is of the second generation or later: By measuring in meteorites the relative abundances of unstable atomic nuclei and their decay products, Caltech's Gerald J. Wasserburg and his colleagues have deduced that the solar system is only 4.58 billion years old—i.e., it formed when our Milky Way Galaxy was already ~ 5 billion years old.

Before recombination the radiation was kept thermalized by interactions with matter. However, since recombination the radiation has propagated freely, interacting only with the gravitation (spacetime curvature) of our universe.

****

EXERCISES

Exercise 27.4 *Example: Friedmann's Cosmological Models*
Consider a model universe of the type envisioned by Alexander Friedmann (1922)—one with zero pressure (i.e., containing only cold matter), so its density is $\rho_M \propto 1/a^3$; cf. Eq. (27.38). Write this density in the form

$$\rho = \frac{3}{8\pi}\frac{a_m}{a^3} \ , \tag{27.48}$$

where a_m is a constant whose normalization (the factor $3/8\pi$) is chosen for later convenience.

(a) Draw the effective potential $V(a)$ for this model universe, and from it discuss qualitatively the evolution in the three cases $k = 0, \pm 1$.

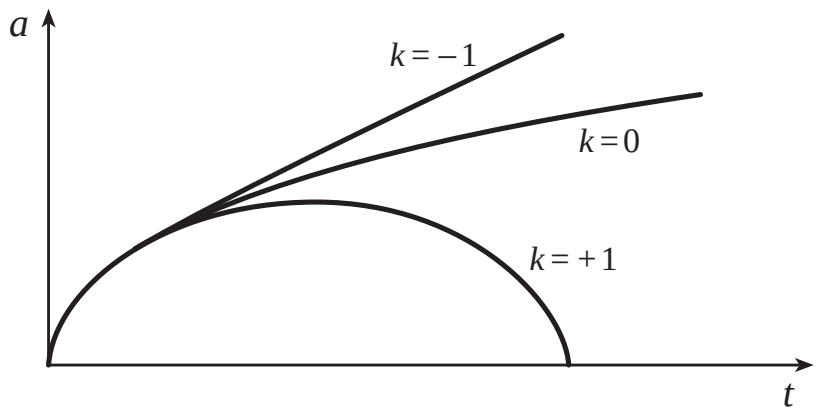


Fig. 27.5: Time evolution of the expansion factor $a(t)$ for the zero-pressure, Friedmann cosmological models (Exercise 27.4). The three curves correspond to the closed, $k = +1$ model; the flat, $k = 0$, model; and the open, $k = -1$ model.

(b) Show that for a closed, $k = +1$ universe with zero pressure, the expansion factor a evolves as follows:

$$\boxed{a = \frac{a_m}{2}(1 - \cos\eta) \ , \quad t = \frac{a_m}{2}(\eta - \sin\eta) \ .} \tag{27.49}$$

Here η is a parameter which we shall use as a time coordinate in Sec. 27.5 [Eq. (27.79)]below. This $a(t)$ [depicted in Fig. 27.5], is a *cycloid*.

(c) Show that for a flat, $k = 0$ universe, the evolution is given by

$$\boxed{a = \left(\frac{9a_m}{4}\right)^{1/3} t^{2/3} \ , \quad \rho = \frac{1}{6\pi t^2} \ .} \tag{27.50}$$

(d) Show that for an open, $k = -1$ universe, the evolution is given by

$$\boxed{a = \frac{a_m}{2}(\cosh\eta - 1) \ , \quad t = \frac{a_m}{2}(\sinh\eta - \eta) \ .} \tag{27.51}$$

Note, as shown in Fig. 27.5, that for small expansion factors, $a \ll a_m$, the evolutions of the three models are almost identical.

Exercise 27.5 *Problem: Einstein's Static Universe*
Consider a model universe of the sort that Einstein (1917) envisioned: one with a nonzero, positive cosmological constant and containing matter with negligible pressure, $P = 0$. Reinterpret this in modern language as a universe with cold matter and a nonzero vacuum stress-energy. Einstein believed that (when averaged over the motions of all the stars), the universe must be static — i.e., neither expanding nor contracting: $a =$ constant independent of time.

(a) Show that Einstein's equations do admit a solution of this form, and deduce from it (i) the spatial geometry of the universe (spherical, flat, or hyperboloidal), and (ii) relationships between the universe's "radius" a, its matter density ρ_M, and its vacuum energy density ρ_Λ.

(b) Show that Einstein's static cosmological model is unstable against small perturbations of its "radius": if a is reduced slightly from its static, equilibrium value, the universe will begin to collapse; if a is increased slightly, the universe will begin to expand. Einstein seems not to have noticed this instability.

For a historical discussion of Einstein's ideas about cosmology, see Sec. 15e of Pais (1982).

Exercise 27.6 *Example: Cosmological Redshift*

Consider a particle, with finite rest mass or zero, that travels freely through a homogeneous, isotropic universe. Let the particle have energy E as measured by a homogeneous observer side-by-side with it, when it starts its travels, at some early epoch in the universe's evolution; and denote by E_o its energy as measured by a homogeneous observer at its location, near Earth, today. Denote by

$$p = \sqrt{E^2 - m^2}\ , \quad p_o = \sqrt{{E_o}^2 - m^2} \tag{27.52}$$

the momentum of the particle as measured in the early epoch and today. In this problem you will evaluate the ratio of the momentum today to the momentum in the early epoch, p_o/p, and will deduce some consequences of that ratio.

(a) Place the spatial origin, $\chi = 0$, of the spatial coordinates of a Robertson-Walker coordinate system [Eq. (27.6)] at the point where the particle started its travel. (Homogeneity guarantees we can put the spatial origin anywhere we wish.) Orient the coordinates so the particle starts out moving along the "equatorial plane" of the coordinate system, $\theta = \pi/2$ and along $\phi = 0$. (Isotropy guarantees we can orient our spherical coordinates about their origin in any way we wish.) Then spherical symmetry about the origin guarantees the particle will continue always to travel radially, with $\theta = \pi/2$ and $\phi = 0$ all along its world line; in other words, the only nonvanishing contravariant components of its 4-momentum are $p^t = dt/d\zeta$ and $p^\chi = d\chi/d\zeta$; and, since the metric is diagonal, the lowering of components shows that the only nonvanishing covariant components are p_t and p_χ. Show that the quantity p_χ is conserved along the particle's world line.

(b) Express the momentum p measured by the homogeneous observer at the starting point and the momentum p_o measured near Earth today in terms of p_χ. [Hint: The local Lorentz frame of a homogeneous observer has the basis vectors (27.20).] Show that

$$\frac{p_o}{p} = \frac{a}{a_o} = \frac{1}{1+z}\ , \tag{27.53}$$

where z is the cosmological redshift at the starting point.

(c) Show that if the particle is a photon, then its wavelength is redshifted in accord with Eqs. (27.42) and (27.43).

(d) Show that if the particle has finite rest mass and has speed $v \ll 1$ at its starting point, as measured by a homogeneous observer there, then its velocity today as measured by the near-Earth homogeneous observer will be

$$v_o = v\frac{a}{a_o} = \frac{v}{1+z} \ . \tag{27.54}$$

Exercise 27.7 *Practice: Cosmic Microwave Radiation in an Anisotropic Cosmological Model*

Consider a cosmological model with the spacetime metric

$$ds^2 = -dt^2 + a^2(t)dx^2 + b^2(t)dy^2 + c^2(t)dz^2 \ . \tag{27.55}$$

The quantities a, b, and c (not to be confused with the speed of light which is unity in this chapter) are expansion factors for the evolution of the universe along its x, y, and z axes. The Einstein field equation governs the time evolution of these expansion factors; but the details of that evolution will not be important to us in this problem.

(a) Show that the space slices $t =$ const in this model have Euclidean geometry, so the model is spatially flat and homogeneous. Show that the observers who see these slices as hypersurfaces of simultaneity, i.e., the homogeneous observers, have world lines of constant x, y, and z, and their proper time is equal to the coordinate time t.

(b) At time t_e when the expansion factors were a_e, b_e, and c_e the universe was filled with isotropic black-body photons with temperature T_e, as measured by homogeneous observers. Define $p_x \equiv \vec{p}\cdot\partial/\partial x$, $p_y \equiv \vec{p}\cdot\partial/\partial y$, $p_z \equiv \vec{p}\cdot\partial/\partial z$ for each photon. Show that in terms of these quantities the photon distribution function at time t_e is

$$\mathcal{N} = \frac{2}{h^3}\frac{1}{e^{E/kT_e}-1} \ , \quad \text{where } E = \left[\left(\frac{p_x}{a_e}\right)^2 + \left(\frac{p_y}{b_e}\right)^2 + \left(\frac{p_z}{c_e}\right)^2\right]^{1/2} \ . \tag{27.56}$$

(c) After time t_e each photon moves freely through spacetime (no emission, absorption, or scattering). Explain why p_x, p_y, and p_z are constants of the motion along the phase-space trajectory of each photon.

(d) Explain why $\mathcal{N}$, expressed in terms of p_x, p_y, p_z, retains precisely the form (27.56) for all times $t > t_e$.

(e) At time $t_o > t_e$, when the expansion factors are a_o, b_o, c_o, what are the basis vectors $\vec{e}_{\hat{0}}$, $\vec{e}_{\hat{x}}$, $\vec{e}_{\hat{y}}$, $\vec{e}_{\hat{z}}$ of the local Lorentz frame of a homogeneous observer?

(f) Suppose that such an observer looks at the photons coming in from a direction $\mathbf{n} = n_{\hat{x}}\vec{e}_{\hat{x}} + n_{\hat{y}}\vec{e}_{\hat{y}} + n_{\hat{z}}\vec{e}_{\hat{z}}$ on the sky. Show that she sees a precisely Planck frequency distribution with temperature T_o that depends on the direction $\mathbf{n}$ that she looks:

$$T_o = T_e\left[\left(\frac{a_o}{a_e}n_{\hat{x}}\right)^2 + \left(\frac{b_o}{b_e}n_{\hat{y}}\right)^2 + \left(\frac{c_o}{c_e}n_{\hat{z}}\right)^2\right]^{-1/2} \ . \tag{27.57}$$

(g) In the case of isotropic expansion, $a = b = c$, show that T_o is isotropic and is redshifted by the same factor, $1 + z$, as the frequency of each photon [Eqs. (27.42) and (27.43)]:

$$\frac{T_o}{T_e} = \frac{1}{1+z} = \frac{a_e}{a_o} \,. \tag{27.58}$$

[The redshift z must not be confused with the coordinate z of Eq. (27.55).]

****

27.5 Observational Cosmology

27.5.1 Parameters characterizing the universe

Our zero-order (homogeneous and isotropic) model of the universe is characterized, today, by the following parameters: (i) The quantity

$$\boxed{H_o \equiv \dot{a}_o/a_o \,,} \tag{27.59}$$

which is called the Hubble expansion rate, and which determines the critical density $\rho_{\rm crit} = (3/8\pi)H_o^2$ [Eq. (27.35)]. (ii) The density of cold matter measured in units of the critical density, $\Omega_M = \rho_{Mo}/\rho_{\rm crit}$ [Eq. (27.36)]. (iii) The split of Ω_M into two parts, $\Omega_M = \Omega_B + \Omega_{\rm CDM}$. Here Ω_B is that portion due to "baryonic matter," the type of matter (protons, neutrons, electrons, and atoms and molecules made from them) of which stars, galaxies, and interstellar gas are made; and $\Omega_{\rm CDM}$ is the portion due to non-baryonic, "cold, dark matter" (probably axions and/or neutralinos and/or other types of weakly interacting, massive particles produced in the big bang). (iv) The temperature T_{Ro} of the CMB. (v) The density of radiation in units of the critical density, $\Omega_R = \rho_{Ro}/\rho_{\rm crit}$. (vi) Ω_Λ, the density of dark energy in units of the critical density. (vii) P_Λ/ρ_Λ, the ratio of the dark energy's pressure to its density (equal to -1 if the dark energy is a nonzero stress-energy of the vacuum).

The time-time component of the Einstein field equation, Eq. (27.26), translated into the notation of our seven parameters, says

$$\boxed{\frac{k}{a_o^2} = H_o^2(\Omega - 1) \,,} \quad \text{where} \quad \Omega = \Omega_M + \Omega_R + \Omega_\Lambda \simeq \Omega_M + \Omega_\Lambda \tag{27.60}$$

is the total density in units of the critical density. In most of the older literature (before ~ 1995), much attention is paid to the dimensionless *deceleration parameter* of the universe, defined as

$$\boxed{q_o \equiv \frac{-\ddot{a}_o/a_o}{H_o^2} = \frac{\Omega_M}{2} + \frac{P_\Lambda}{\rho_\Lambda}\Omega_\Lambda \,.} \tag{27.61}$$

Here the second equality follows from the space-space component of the Einstein field equation, Eq. (27.27), translated into the language of our parameters, together with the fact that

today the only significant pressure is that of dark energy, P_Λ. We shall not use q_o in this book.

Remarkably, the values of our seven independent parameters H_o, Ω_B, Ω_M, T_{Ro}, Ω_R, Ω_Λ, and P_Λ/ρ_Λ are all fairly well known today (spring 2003), thanks largely to major observational progress in the past several years. In this section we shall discuss the observations that have been most effective in determining these parameters. For greater detail see, e.g., the review article by Turner (1999), and the recent WMAP results presented by Bennett et. al. (2003).

27.5.2 Local Lorentz frame of homogeneous observers near Earth

As a foundation for discussing some of the observations, we shall construct the local Lorentz frame of a homogeneous observer near Earth. (The Earth moves relative to this frame with a speed $v = 630 \pm 20$ km s^{-1}, as revealed by a dipole anisotropy in the temperature distribution of the CMB on the Earth's sky.)

Homogeneous observers can have local Lorentz frames because they moves freely (along timelike geodesics) through spacetime; cf. Ex. 27.1. For ease of analysis, we place the spatial origin of the Robertson-Walker $\{t, \chi, \theta, \phi\}$ coordinate system at the location of a near-Earth homogeneous observer. Then that observer's Lorentz coordinates are

$$\hat{t} = t + \frac{1}{2}\chi^2 a\dot{a} \ , \quad \hat{x} \equiv a\chi \sin\theta\cos\phi \ , \quad \hat{y} \equiv a\chi\sin\theta\sin\phi \ , \quad \hat{z} \equiv a\chi\cos\theta \ ; \tag{27.62}$$

cf. Eq. (24.12) and associated discussion. Note that only at second-order in the distance away from the origin of the local Lorentz frame does the Lorentz time $\hat{t}$ differ from the Robertson-Walker time coordinate t. This second-order difference will never be important for anything we compute, so henceforth we will ignore it and set $\hat{t} = t$.

The near-Earth local Lorentz frame, like any local Lorentz frame, must be kept spatially small compared to the radius of curvature of spacetime. That radius of curvature is related to the Lorentz-frame components of spacetime's Riemann curvature tensor by

$$\mathcal{R} \sim \frac{1}{|R_{\hat{\alpha}\hat{\beta}\hat{\gamma}\hat{\delta}}|^{1/2}} \ . \tag{27.63}$$

More precisely, since it is the largest components of Riemann that have the biggest physical effects, we should use the largest components of Riemann when evaluating $\mathcal{R}$. These largest components of Riemann today turn out to be $\sim \dot{a}_o^2/a_o^2 = H_o^2$ and $\sim k/a_o^2$. Observations discussed below reveal that $\Omega \sim 1$, so $k/a_o^2 \lesssim H_o^2$ [Eq. (27.60)]. Therefore, the universe's coarse-grain-averaged radius of spacetime curvature today is

$$\boxed{\mathcal{R} \sim \frac{1}{H_o} \ ;} \tag{27.64}$$

and the demand that the local Lorentz frame be small compared to this radius of curvature is equivalent to the demand that we confine the local-Lorentz spatial coordinates to the region

$$H_o r \ll 1 \ , \quad \text{where } r \equiv \sqrt{\hat{x}^2 + \hat{y}^2 + \hat{z}^2} = a\chi + \mathrm{O}(a\chi^3) \ . \tag{27.65}$$

(Below we shall neglect the tiny $a\chi^3$ correction.)

27.5.3 Hubble expansion rate

Consider a homogeneous observer near enough to Earth to be in the near-Earth local Lorentz frame, but not at its origin. Such an observer has fixed Robertson-Walker radius χ and local-Lorentz radius $r = a\chi$, and thus moves away from the origin of the local Lorentz frame with a velocity, as measured by the frame's rods and clocks, given by $v = dr/dt = \dot{a}\chi$; i.e., evaluating that velocity today, $v = \dot{a}_o\chi$. Correspondingly, special relativity insists that light emitted by this homogeneous observer at local Lorentz radius r and received today by the homogeneous observer at $r = 0$ should be Doppler shifted by an amount $\Delta\lambda/\lambda \cong v = \dot{a}_o\chi$. Note that this Doppler shift is proportional to the distance between the homogeneous observers, with the proportionality factor equal to the Hubble constant:

$$\boxed{z \equiv \frac{\Delta\lambda}{\lambda} = v = H_o r \ .} \tag{27.66}$$

This Doppler shift is actually nothing but the cosmological redshift, looked at from a new viewpoint: When specialized to emitters and receivers that are near each other, so they can be covered by a single local Lorentz frame, the cosmological redshift formula (27.43) reduces to

$$z = \frac{a_o}{a} - 1 = \frac{1}{a}(a_o - a) \cong \frac{1}{a_o}\dot{a}_o\Delta t = H_o\Delta t \ , \tag{27.67}$$

where Δt is the time required for the light to travel from emitter to receiver. Since the light travels at unit speed as measured in the local Lorentz frame, Δt is equal to the distance r between emitter and receiver, and the cosmological redshift becomes $z = H_o r$, in agreement with the Doppler shift (27.66).

To the extent that the galaxies which astronomers study are at rest with respect to homogeneous observers, they should exhibit the *distance-redshift relation* (27.66). In reality, because of the gravitational attractions of other, nearby galaxies, typical galaxies are not at rest relative to homogeneous observers, i.e., not at rest relative to *the "Hubble flow"*. Their velocities relative to local homogeneous observers are called *peculiar velocities* and have magnitudes that are typically $v_{\rm pec} \sim 300$ km/sec $\sim 10^{-3}$, and can be as large as $v_{\rm pec} \sim 1000$ km/sec. In order to extract the Hubble constant from measurements of galactic distances and redshifts, astronomers must determine and correct for these peculiar motions. That correction task is rather difficult when one is dealing with fairly nearby galaxies, say with $z \lesssim 0.003$ so $v \lesssim 1000$ km/sec. On the other hand, when one is dealing with more distant galaxies, the task of determining the distance r is difficult. As a result, the measurement of the Hubble constant has been a long, arduous task, involving hundreds of astronomers working for 2/3 of a century. Today this effort has finally paid off, with a number of somewhat independent measurements that give

$$\boxed{H_o = (70 \pm 5)\,{\rm km\ sec^{-1}Mpc^{-1}} \ ,} \tag{27.68}$$

where the unit of velocity is km s $^{-1}$ and the unit of distance is 1 Mpc = 1 megaparsec = 10^6 pc. Converting into units of length and time (using $c = 1$), the inverse Hubble constant is

$$\boxed{\frac{1}{H_o} = (4.3 \pm 0.4)\ {\rm Gpc} = (14.0 \pm 1.3) \times 10^9\ {\rm years} \ .} \tag{27.69}$$

Correspondingly, the critical density to close the universe, Eq. (27.35), is

$$\boxed{\rho_{\rm crit} = (9.2 \pm 1.2) \times 10^{-30} {\rm g/cm}^3 \ .} \tag{27.70}$$

In the cosmology literature one often meets the "Hubble parameter," whose definition and measured value are

$$\boxed{h \equiv \frac{H_o}{100 {\rm km\ s^{-1} Mpc^{-1}}} = 0.70 \pm 0.05 \ .} \tag{27.71}$$

27.5.4 Primordial nucleosynthesis

When the universe was about a minute old and had temperature $T_R \sim 10^9$ K, nuclear burning converted free protons and neutrons into the light isotopes deuterium $\equiv^2$H, ^{3}He, ^{4}He, and ^{7}Li. Over the past four decades astronomers have worked hard to achieve precision measurements of the primordial abundances of these isotopes. Those measurements, when compared with nucleosynthesis calculations based on models for the universal expansion, produce remarkably good agreement—but only when (i) the number of species of neutrinos (which contribute to the radiation density and via the Einstein equation to the expansion rate during the burning) is no greater than three (electron, muon, and tau neutrinos); and (ii) dark energy has negligible influence on the universe's expansion except in and near the modern era, and possibly before nucleosynthesis; and (iii) the normalized baryon density is

$$\boxed{\Omega_B = (0.040 \pm 0.006) \left(\frac{0.70}{h}\right)^2 \ .} \tag{27.72}$$

Here 0.006 is the 95 per cent confidence limit. This is a remarkably accurate measurement of the density of baryonic matter—and it shows that ρ_B is only about 5 per cent of the critical density.

27.5.5 Density of Cold Dark Matter

The only kind of matter that can condense, gravitationally, is that with pressure $P \ll \rho$, i.e., cold matter. The pressures of the universe's other constituents (radiation and dark energy) prevent them from condensing significantly; they must be spread rather smoothly throughout the universe. The total density of cold matter, Ω_M, can be inferred from the gravitationally-measured masses M of large clusters of galaxies. Those masses are measured in four ways: (i) by applying the virial theorem to the motions of their individual galaxies, (ii) by applying the equation of hydrostatic equilibrium to the distributions of hot, X-ray emitting gas in the clusters, (iii) by studying the influence of the clusters as gravitational lenses for more distant objects, and (iv) from the positions and shapes of the Doppler peaks in the CMB (Sec. 27.5.7). The results of the four methods agree reasonably well and yield a total density of cold matter

$$\boxed{\Omega_M = 0.3 \pm 0.1 \ ;} \tag{27.73}$$

see Turner (1999) for details and references; and for the Doppler peak measurements see Bennett et. al. (2003).

This $\Omega_M \simeq 0.3$ is much larger than the density of baryonic matter $\Omega_B \simeq 0.05$. Their difference,

$$\boxed{\Omega_{\rm CDM} = \Omega_M - \Omega_B = 0.26 \pm 0.1} \tag{27.74}$$

is the density of cold, dark matter.

27.5.6 Radiation Temperature and Density

The temperature of the CMB has been measured, from its Planckian spectrum, to be

$$\boxed{T_R = 2.728 \pm 0.002{\rm K}} \tag{27.75}$$

This temperature tells us with excellent accuracy the contribution of photons to the radiation density

$$\boxed{\Omega_\gamma = (0.5040 \pm 0.005) \left(\frac{h}{0.70}\right)^2 \times 10^{-4} \ .} \tag{27.76}$$

The radiation also includes primordial gravitational waves (gravitons), whose energy density is predicted by inflationary arguments to be small compared to Ω_γ, though this prediction could be wrong. It can be no larger than $\Omega_g \sim \Omega_\gamma$, as otherwise the gravitons would have exerted an unacceptably large influence on the expansion rate of the universe during primordial nucleosynthesis and thereby would have distorted the nuclear abundances measureably. The same is true of other hypothetical forms of radiation. Primordial neutrinos must have been in statistical equilibrium with photons and other forms of matter and radiation in the very early universe. Statistical arguments about that equilibrium predict an energy density for each neutrino species of $\Omega_\nu = (7/8)(4/11)^{4/2}\Omega_\gamma$, so long as $kT_R \gg m_\nu c^2$ (the neutrino rest mass-energy). Recent measurements of neutrino oscillations tell us that the neutrinos have rest masses $\gtrsim 0.01$ eV, which implies that they behaved like radiation until some transition temperature $T_R \gtrsim 100$ K (at a redshift $\gtrsim 30$) and then became nonrelativistic, with negligible pressure.

Combining all these considerations, we see that the total radiation density must be

$$\boxed{\Omega_R \sim 1 \times 10^{-4}} \tag{27.77}$$

to within a factor of order 2.

27.5.7 Anisotropy of the CMB: Measurements of the Doppler Peaks

Consider an object with physical diameter D that resides at a distance r from Earth, and neglect the Earth's motion and object's motion relative to homogeneous observers. Then

the object's angular diameter Θ as observed from Earth will be $\Theta = D/r$, if $r \ll 1/H_o$ so the effects of spacetime curvature are negligible. For greater distances, $r \sim 1/H_o$, the ratio

$$\boxed{r_{\rm AD} \equiv \frac{D}{\Theta}} \tag{27.78}$$

(called the object's *angular-diameter distance*) will be strongly influenced by the spacetime curvature—and thence by the cosmological parameters H_o, Ω_M, Ω_Λ, P_Λ/ρ_Λ that influence the curvature significantly. In Ex. 27.8 formulas are derived for $r_{\rm AD}$ as a function of these parameters and the object's cosmological redshift z (a measure of its distance).

Astronomers searched for many decades for objects on the sky (*standard yardsticks*), whose physical diameter D could be known with high confidence. By combining the known D's with the yardsticks' measured angular diameters Θ to get their $r_{\rm AD} = D/\Theta$ and by measuring the redshifts z of their spectral lines, the astronomers hoped thereby to infer the cosmological parameters from the theoretical relation $r_{\rm AD}(z,$ cosmological parameters).

This effort produced little of value in the era ∼1930 to ∼1990, when astronomers were focusing on familiar astronomical objects such as galaxies. No way could be found to determine, reliably, the physical diameter D of any such object.

Finally, in 1994, Marc Kamionkowski, David Spergel and Naoshi Sugiyama (1994) identified an object of a very different sort, whose physical diameter D could be known with high confidence: the *cosmological horizon* in the era when the primordial plasma was recombining and matter and radiation were decoupling from each other. This was the long-sought standard yardstick.

This cosmological horizon is *not* the same thing as the horizon of a black hole, but it is analogous. It is the distance between objects that are just barely able to communicate with each other via light signals. To discuss this concept quantitatively, it is useful to introduce a new time coordinate η for the Robertson-Walker line element (27.6)

$$\boxed{\eta = \int \frac{dt}{a}\ ; \quad \text{so} \quad d\eta = \frac{dt}{a}\ .} \tag{27.79}$$

Then the line element becomes

$$\boxed{ds^2 = a^2[-d\eta^2 + d\chi^2 + \Sigma^2(d\theta^2 + \sin^2\theta d\phi^2)]\ .} \tag{27.80}$$

By setting $\eta = 0$ at the beginning of the expansion and $\eta = \eta_{\rm rec}$ at the era of recombination, and noting that light travels in the χ direction with coordinate speed $d\chi/d\eta = 0$, we see that the diameter of the horizon at recombination is

$$D_{\rm rec} = \eta_{\rm rec}\, a_{\rm rec}\ , \quad \text{where} \quad \eta_{\rm rec} = \int_0^{t_{\rm rec}} \frac{dt}{a} \tag{27.81}$$

and $a_{\rm rec}$ is the value of a at recombination. Two objects separated by a distance greater than $D_{\rm rec}$ were unable to communicate with each other at recombination, because there had not been sufficient time since the birth of the universe for light to travel from one to the other. In this sense, they were outside each others' cosmological horizon. Objects with

separations less than $D_{\rm rec}$ could communicate, at recombination; i.e., they were inside each others' cosmological horizon.

As the universe expands, the cosmological horizon expands; objects that are outside each others' horizons in the early universe come inside those horizons at some later time, and can then begin to communicate.

Kamionkowski, Spergel and Sugiyama realized that the universe provides us with markers on the sky that delineate the horizon diameter at recombination, $D_{\rm rec}$. These markers are anisotropies of the CMB, produced by the same density and temperature inhomogeneities as would later grow to form galaxies.

The inhomogeneities are known, observationally, to have been perturbations of the density with fixed, homogeneous entropy per baryon, i.e. with fixed T_R^3/ρ_M, and with amplitudes, as they came inside the horizon,

$$3\frac{\Delta T_R}{T_R} = \frac{\Delta\rho_M}{\rho_M} \sim 1\times 10^{-4}\ . \tag{27.82}$$

We can resolve the perturbations $\Delta T_R/T_R$ at recombination into spatial Fourier components characterized by wave number k, or equivalently by reduced wavelength $\lambda\!\!\!^{-} = 1/k$. Observers on Earth find it more convenient to resolve the perturbations into spherical harmonics on the sky. Since order $\ell = 1$ corresponds to a perturbation with angular wavelength 360 degrees $= 2\pi$ radians, order ℓ must be a perturbation with angular wavelength $2\pi/\ell$ and thence angular reduced wavelength $\Theta = 1/\ell$. The ratio $\lambda\!\!\!^{-}/\Theta$ of physical reduced wavelength to angular reduced wavelength is the angular-diameter distance over which the CMB photons have traveled since recombination:

$$r_{\rm AD}^{\rm rec} = \frac{\lambda\!\!\!^{-}}{\Theta} = \frac{\ell}{k}\ . \tag{27.83}$$

Now, consider perturbations with spatial scale small enough that a reduced wavelength $\lambda\!\!\!^{-}$ came inside the horizon ("crossed the horizon") somewhat earlier than recombination. Before $\lambda\!\!\!^{-}$ crossed the horizon, each high temperature region was unaware of a neighboring low temperature region, so the two evolved independently, in such a way that their fractional temperature difference grew as

$$\frac{\Delta T_R}{T_R} \propto a \propto t^{2/3} \tag{27.84}$$

[Ex. 27.10]. When $\lambda\!\!\!^{-}$ crossed the horizon (i.e., when the horizon expanded to become larger than $\lambda\!\!\!^{-}$), the neighboring regions began to communicate. The high-T_R, high-ρ_M region pushed out against its low-T_R, low-ρ_M neighbor, generating sound waves; and correspondingly, the growth of $\Delta T_R/T_R$ changed into acoustic oscillations.

For perturbations with some specific physical size $\lambda\!\!\!^{-}_{1/4}$ (angular size $\Theta_{1/4}$), the acoustic oscillations had completed one quarter cyle at the time of recombination, so their temperature contrast was reduced, at recombination, to zero. For pertubations a little smaller, $\Theta_{1/2}$, the oscillations had completed a half cycle at recombination, so the hot regions and cold regions were reversed, and the temperature contrast was roughly as large as at horizon crossing. Perturbations still smaller, $\Theta_{3/4}$, had completed 3/4 of a cycle at recombination, so their

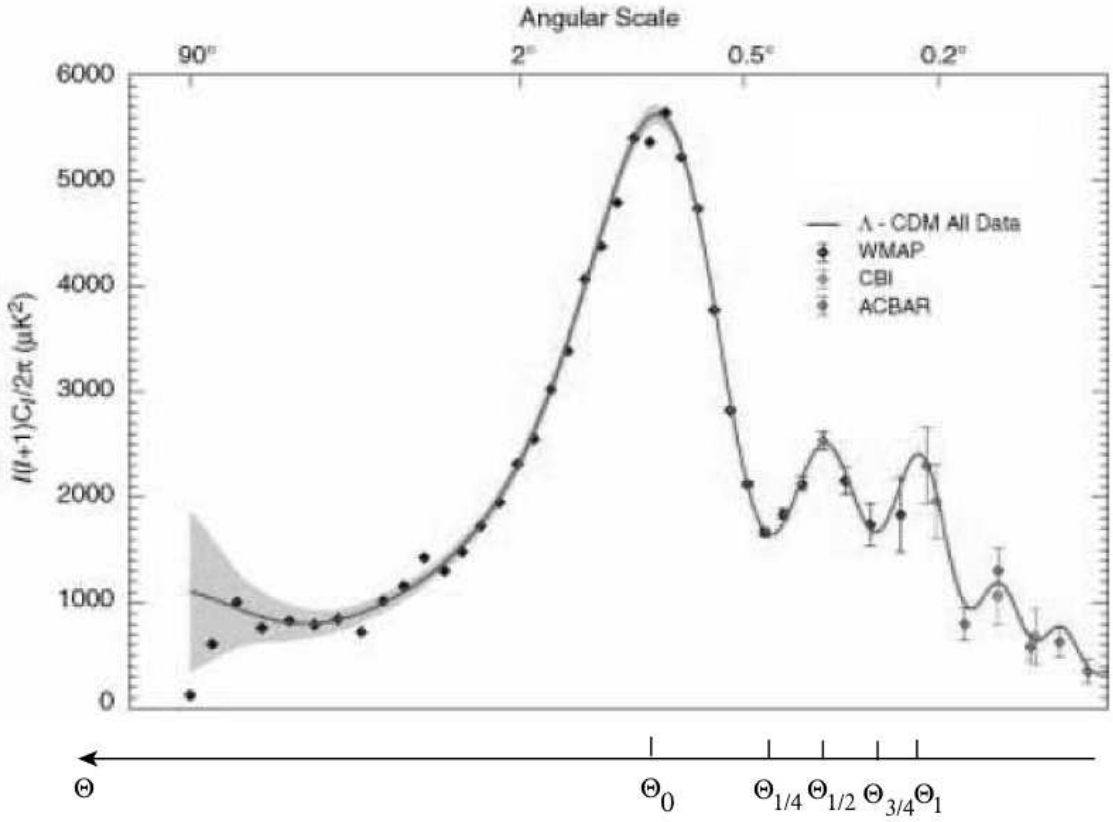


Fig. 27.6: Anisotropy of the CMB as measured by WMAP (the first two Doppler peaks; most error bars smaller than the dots; Bennett et. al. 2003) and by CBI and ACBAR (the last three peaks; Pearson et. al. 2002, and Kuo et. al. 2002). Plotted vertically is the mean square temperature fluctuation; plotted horizontally is the angular scale Θ. The solid curve is the theoretical prediction when one inserts the best-fit values of the cosmological parameters. The grey shading is the range of statistical fluctuations one would expect in an ensemble of universes all with these same cosmological parameter values. This figure is adapted from Bennett et. al. (2003).

density contrast was momentarily zero. Perturbations smaller still, Θ_1, had completed a full cycle of oscillation at recombination and so had a large density contrast; and so forth.

The result is the pattern of temperature anisotropy as a function of Θ or equivalently $\ell = 1/\Theta$ shown in Fig. 27.6. The first peak in the pattern is for perturbations whose reduced wavelength λ_0 had only recently come inside the horizon at recombination, so $\lambda_0 = r_{\rm AD}\Theta_0$ is equal to the diameter $D_{\rm rec}$ of the horizon at recombination, aside from a small difference that can be computed with confidence. This is the standard yardstick that astronomers had sought for decades.

The basic structure of the pattern of anisotropy oscillations shown in Fig. 27.6 is in accord with the above description of acoustic oscillations, but the precise details are modestly different because the initial distribution of inhomogeneities is statistical (i.e. is a random process in the sense of Chap. 5), and the physics of the oscillations is somewhat complex. Examples of the complexities are: (i) ΔT_R does not go to zero at the minima $\Theta_{1/3}$ and

$\Theta_{3/4}$ because the emitting matter has acquired inhomogeneous velocities relative to Earth by falling into the oscillations' gravitational potential wells, and these velocities produce Doppler shifts that smear out the minima. (ii) This same infall makes the density and temperature contrasts smaller at the half-cycle point $\Theta_{1/2}$ than at the full-cycle points Θ_0 and Θ_1.

Despite these complexities and statistical effects, the shapes of the acoustic oscillations can be computed with high confidence, once one has chosen values for the cosmological parameters. The reason for the confidence is that the amplitude of the oscillations is very small, so nonlinear effects are negligible. The pattern of the temperature oscillations at recombination is computed as a function of physical length $\lambda\!\!\!^{-}$, with results that depend modestly on some of the cosmological parameters; and then the physical pattern is converted into a pattern as seen on Earth's sky, using the angular-diameter distance $r_{AD} = \lambda\!\!\!^{-}/\Theta$ that the CMB photons have traveled since recombination, which depends very strongly on the cosmological parameters.

Remarkably, the positions Θ_0, $\Theta_{1/2}$, Θ_1, ... of the oscillation peaks (called *Doppler peaks* for no good reason[2]), depend more strongly on the total density $\Omega \simeq \Omega_M + \Omega_\Lambda$ than on other parameters.

The first quantitative studies of the Doppler peaks, by the Boomerang project's balloon-borne instruments (Lange et. al. 2000) and soon thereafter by MAXIMA (Balbi et. al. 2000) revealed that $\Omega = 1.0 \pm 0.2$ — a great triumph: the universe's total density is approximately critical, and therefore its spatial geometry is approximately flat. A variety of other balloon-based and ground-based measurements in 2000–2003 led to increasing confidence in this conclusion and in a variety of other Boomerang/MAXIMA cosmological discoveries. Recently the WMAP satellite-borne instruments have produced a great leap in accuracy (Bennett et. al. 2003):

$$\Omega = 1.02 \pm 0.02 \tag{27.85}$$

see Fig. 27.6. This near-unity value of Ω implies that the universe is very close to being spatially flat; see Eq. (27.60).

The WMAP measurements also reveal that before the sound waves began producing the oscillations, the spectral density of the temperature perturbations decayed as $S_{T_R}(k) \propto k^{-0.93\pm0.03}$, so the rms amplitude of the fluctuations, $\Delta T_R^{\rm rms} = \sqrt{(k/2\pi)S_{T_R}(k)}$ [Eq. (5.64)] was nearly independent of wave number k, i.e. independent of Θ. This is in accord with predictions from "inflationary" models for the production of the perturbations (Sec. 27.7 below.)

[2]Yakov B. Zel'dovich and Rashid A. Sunyaev (1970), who first predicted the existence of these peaks, later gave them a name that has a little more justification: They called them *Sakharov oscillations* because Andrei D. Sakharov (1965) was the first to predict the sound waves that give rise to the peaks. Zel'dovich and Sunyaev introduced this name at a time when their close friend Andrei Sakharov was being attacked by the Soviet government; they hoped that this would call attention to Sakharov's international eminence and help protect him. It seemed not to help.

27.5.8 Age of the universe: Constraint on the dark energy

The total mass density $\Omega = 1.0 \pm 0.2$ from CMB anisotropy and the cold-matter mass density $\Omega_M = 0.4 \pm 0.1$ leave a *missing mass density* $\Omega_\Lambda = 0.6 \pm 0.2$, which must be in some exotic form (dark energy) that does not condense, gravitationally, along with the cold matter, and that therefore must have a pressure $|P_\Lambda| \sim \rho_\Lambda$. This dark energy must have had a negligible density at the time of recombination and at the time of nucleosynthesis; otherwise, it would have disturbed the shapes of the Doppler peaks and distorted the nuclear abundances. In order that it be significant now and small earlier, compared to cold matter, it must have a negative pressure $P_\Lambda < 0$. One handle on how negative comes from the age of the universe.

Assuming, that $P_\Lambda/\rho_\Lambda \equiv w_\Lambda$ was constant or approximately so during most of the age of the universe (i.e., back to redshift $z \sim 10$) and that the dark energy did not exchange significant energy with other constituents of the universe during that recent epoch, then the first law of thermodynamics implies that $\rho_\Lambda = \rho_{\rm crit}\Omega_\Lambda (a_o/a)^{3(1+w_\Lambda)}$. Inserting this, $\rho_M = \rho_{\rm crit}\Omega_M (a_o/a)^3$, and $\rho = \rho_M + \rho_\Lambda$ into the Einstein equation (27.26), solving for $dt = da/\dot{a}$, and integrating, we obtain for the product of the current age of the universe t_o and the Hubble expansion rate H_o:

$$\boxed{H_o t_o = \int_0^1 \frac{dv}{\sqrt{1 - \Omega_M - \Omega_\Lambda + \frac{1}{v}\left(\Omega_M + \Omega_\Lambda v^{-3w_\Lambda}\right)}}\,.} \tag{27.86}$$

The more negative is $P_\Lambda/\rho_\Lambda = w_\Lambda$, the larger is the integral, and thus the larger is $H_o t_o$.

By comparing the observed properties of the oldest stars in our galaxy with the theory of stellar evolution, and estimating the age of the universe at galaxy formation, astronomers arrive at an estimate

$$t_o = (14 \pm 1.5) \times 10^9 {\rm yr} \tag{27.87}$$

for the age of the universe. Combining this with the measured value of the Hubble constant, Eq. (27.69), we obtain

$$H_o t_o = 1.00 \pm 0.13\,. \tag{27.88}$$

For comparison, Eq. (27.86) with $\Omega_M = 0.3$, $\Omega_\Lambda = 0.7$ (in accord with the measurements) gives $H_o t_o = 0.96$ for $w_\Lambda \equiv P_\Lambda/\rho_\Lambda = -1$ (vacuum), 0.86 for $w_\Lambda = -\frac{1}{2}$, 0.67 for $w_\Lambda = 0$, and even smaller values for positive w_Λ. Thus, *the measured ages of the oldest stars imply that*

$$w_\Lambda \equiv \frac{P_\Lambda}{\rho_\Lambda} \lesssim -\frac{1}{2}\,; \tag{27.89}$$

and those ages are nicely compatible with the dark energy being vacuum stress-energy, $w_\Lambda = -1$.

WMAP by itself is able to determine the age of the universe to somewhat higher accuracy than astronomers infer from the ages of the oldest stars:

$$\boxed{t_o = (13.7 \pm 0.2) \times 10^9 {\rm yr},} \tag{27.90}$$

in very good agreement with the stellar-based age (27.87). WMAP also places a tighter limit on w_Λ:

$$\boxed{w_\Lambda < -0.78} \tag{27.91}$$

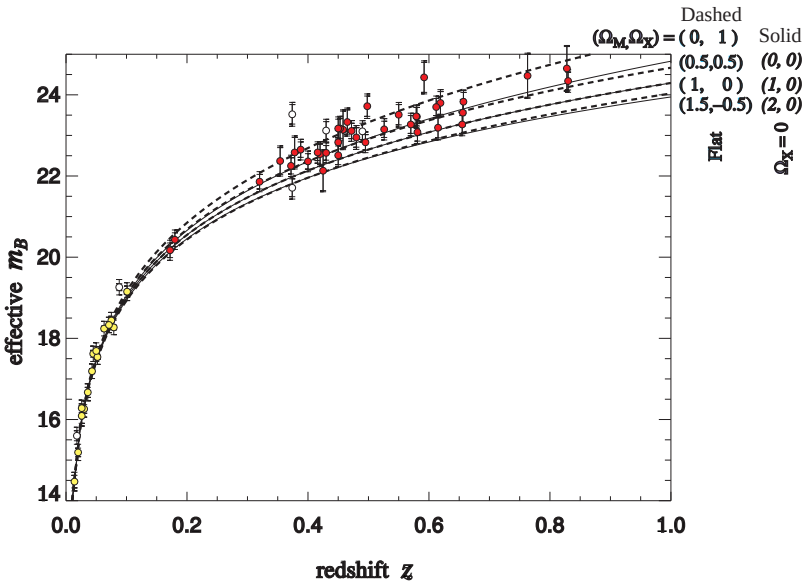


Fig. 27.7: Magnitude-redshift diagram for type Ia supernovae based on observations by Perlmutter et. al. (1999) and others. [Adapted from Perlmutter et. al. (1999).]

at the 95 per cent confidence level.

27.5.9 Magnitude-Redshift relation for type Ia supernovae: Confirmation that the universe is accelerating

The constraint $P_\Lambda/\rho_\Lambda < -0.78$ on the dark energy has a profound consequence for the expansion rate of the universe. In the "particle-in-a-potential" analysis [Eqs. (27.39), (27.40) and Fig. 27.3] the contribution of the dark energy to the potential is $V_\Lambda(a) = -(4\pi/3)a^2\rho_\Lambda \propto a^n$ where $n > 1.3$, which grows stronger with increasing a. Correspondingly, in the present era, the "potential energy" is becoming more negative, which means that the universe's "kinetic energy" $\frac{1}{2}\dot{a}^2$ must be increasing: the universe has recently made the transition from a decelerating expansion to an *accelerating expansion*.

In 1998 two independent groups of astronomers reported the first direct observational evidence of this acceleration (Riess et. al. 1998, Perlmutter et. al. 1999). Their evidence was based on systematic observations of the apparent brightness of *type Ia supernovae* as a function of the supernovae's redshift-measured distances. If the universal expansion is, indeed, accelerating, then distant objects (including supernovae), which we see when the universe was much younger than today, would have experienced a slower universal expansion than we experience today, so their observed redshifts z should be lower than in a universe with constant or decelerating expansion rate. These lowered redshifts should show up as a leftward displacement of the supernovae's locations in a diagram plotting the supernovae's redshift horizontally and their brightness (a measure of their distance) vertically—a so-called *magnitude-redshift diagram*.

Such a diagram is shown in Fig. 27.7. The measure of brightness used in this diagram is

the supernova's *apparent magnitude*

$$\boxed{m \equiv -2.5 \log_{10}(\mathcal{F}/2.5 \times 10^{-8}\,\mathrm{WM}^{-2})\ ,} \tag{27.92}$$

where $\mathcal{F}$ is the flux received at Earth. The sign is chosen so that the dimmer the supernova, the larger the magnitude. A series of theoretical curves is plotted in the diagram, based on assumed values for Ω_M and Ω_Λ, and on the presumption that the dark energy is vacuum stress-energy so $P_\Lambda/\rho_\Lambda = -1$. The formulae for these curves are derived in Ex. 27.9. The solid curves are for *no* dark energy, $\Omega_\Lambda = 0$. The dark energy, which converts the universal deceleration $\ddot{a} < 0$ into acceleration $\ddot{a} > 0$, pushes the curves leftward for distant supernovae (upper right-hand region), as described above. The dashed curves are for a mixture of dark energy and cold matter that sums to the critical density, $\Omega_\Lambda + \Omega_M = 1$.

A detailed analysis of the data by Perlmutter et. al. (1999) gives (assuming $P_\Lambda/\rho_\Lambda = -1$)

$$\Omega_\Lambda = \frac{1}{3}(4\Omega_M + 1) \pm \frac{1}{6}\ . \tag{27.93}$$

Combining with $\Omega_M = 0.3 \pm 0.1$, this implies

$$\Omega_\Lambda = 0.73 \pm 0.2\ , \tag{27.94}$$

in excellent agreement with the CMB measurements and deductions from the ages of the oldest stars.

To recapitulate: A variety of observations all point in the same direction. They agree that our universe is close to spatially flat, with $\Omega_\Lambda \simeq 0.7$, $\Omega_M \simeq 0.3$, and $\Omega_R \sim 10^{-4}$.

****

EXERCISES

Exercise 27.8 *Example: Angular-Diameter Distance*
Consider an electromagnetic emitter at rest in the cosmological fluid (i.e. at rest relative to homogeneous observers), and let the emitter's radiation be observed at Earth. Neglect the Earth's motion relative to homogeneous observers. Let the cosmological redshift of the emitted radiation be $z = \Delta\lambda/\lambda$.

(a) Show that the emitter's angular-diameter distance is

$$\boxed{r_{\rm AD} = \frac{R}{1+z}\ ,} \tag{27.95}$$

where

$$\boxed{R = a_o \Sigma(\Delta\chi) = \frac{\Sigma(\Delta\chi)}{H_o\sqrt{|1-\Omega|}}\ .} \tag{27.96}$$

Here a_o is the Universe's expansion factor today, Σ is the function defined in Eq. (27.8), $\Delta\chi$ is the coordinate distance that light must travel in going from emitter to Earth if its path has constant θ and ϕ, and for simplicity we have assumed $\Omega \neq 1$. [*Hint:* Place the Earth at $\chi = 0$ of the Robertson-Walker coordinate system, and the emitter at $\chi = \Delta\chi$, and use the line element (27.80).]

(b) Assuming that the dark energy is vacuum stress-energy so $P_\Lambda = -\rho_\Lambda$, show that in the limit $\Omega \to 1$ so $k = 0$, the quantity $\Sigma(\Delta\chi)/\sqrt{|1-\Omega|}$ appearing in Eq. (27.96) becomes

$$\frac{\Sigma(\Delta\chi)}{\sqrt{|1-\Omega|}} = \frac{\Delta\chi}{\sqrt{|1-\Omega|}} = \int_1^{1+z} \frac{du}{\sqrt{\Omega_R u^4 + \Omega_M u^3 + (1-\Omega)u^2 + \Omega_\Lambda}} \,. \tag{27.97}$$

[*Hint:* Use Eqs. (27.79) and (27.80) to deduce that

$$\Delta\chi = \int_{t_e}^{t_o} \frac{dt}{a} = \int_1^{1+z} \frac{a}{a_o}\frac{dt}{da} d\frac{a_o}{a}$$

and use the Einstein equation for da/dt.] Note that for a spatially flat universe [$k = 0$, $\Omega = 1$, $\Sigma(\Delta\chi) = \Delta\chi$], Eqs. (27.95), (27.96) and (27.97) imply

$$\boxed{r_{\rm AD} = \frac{1}{H_o(1+z)} \int_1^{1+z} \frac{du}{\sqrt{\Omega_R u^4 + \Omega_M u^3 + \Omega_\Lambda}} \,.} \tag{27.98}$$

(c) Plot $r_{\rm AD}(z)$ for the measured values $\Omega_\Lambda = 0.7$, $\Omega_M = 0.3$, $\Omega_R \simeq 0$, $H_o = 70$ km s^{-1} Mpc^{-1}. Explore graphically how $r_{\rm AD}(z)$ changes as Ω and Ω_Λ/Ω_M change.

Exercise 27.9 *Example: Magnitude-Redshift Relation and Luminosity Distance*
Consider a supernova which emits a total luminosity L as measured in its own local Lorentz frame in an epoch when the expansion factor is a and the cosmological redshift is $z = a_o/a - 1$.

(a) Assume that the supernova and the Earth are both at rest relative to homogeneous observers at their locations. Place the origin $\chi = 0$ of a Robertson-Walker coordinate system at the supernova's location, orient the coordinate axes so the Earth lies at $\theta = \pi/2$ and $\phi = 0$, and denote by $\Delta\chi$ the Earth's radial coordinate location. Show that the flux of energy received from the supernova at Earth today is given by

$$\boxed{\mathcal{F} = \frac{L}{4\pi R^2 (1+z)^2} \,,} \tag{27.99}$$

where R is the same function as appears in the angular-diameter distance, Eqs. (27.96), (27.97).

(b) It is conventional to define the source's *luminosity distance* r_L in such a manner that the flux is $\mathcal{F} = L/4\pi r_L^2$. Eq. (27.99) then implies that

$$\boxed{r_L = (1+z)R = (1+z)^2 r_{\rm AD} \,.} \tag{27.100}$$

Plot $r_L(z)$ for the measured values $\Omega_\Lambda = 0.7$, $\Omega_M = 0.3$, $\Omega_R \simeq 0$, $H_o = 70$ km s^{-1} Mpc^{-1}. Explore graphically how $r_L(z)$ changes as Ω and Ω_Λ/Ω_M change.

Exercise 27.10 *Challenge: Growth of Perturbations Before They Cross the Horizon*
Show that in a matter-dominated universe, $\rho_M \gg \rho_R$ and $\rho_M \gg \rho_\Lambda$, the fractional density difference of two neighboring regions that are outside each others' cosmological horizons grows as $\Delta\rho/\rho \propto a \propto t^{2/3}$. [Hint: the difference of spatial curvature between the two regions is of importance. For a solution, see, e.g., pp. 108–110 of Peebles (1993).]

****

27.6 The Big-Bang Singularity, Quantum Gravity, and the Initial Conditions of the Universe

Although we do not know for sure the correct equation of state at redshifts $z \gg 3 \times 10^{12}$, where the thermal energy of each baryon exceeds its rest mass-energy, the "particle-in-a-potential" form (27.39) of the evolution equation tells us that, so long as $dV/da \geq 0$ at small a, the universe must have begun its expansion in a state of vanishing expansion factor $a = 0$, nonzero $\dot{a}$, and infinite $\dot{a}/a$. Since some of the components of the Riemann curvature tensor in the local Lorentz frame of a homogeneous observer are of order $\dot{a}^2/a^2$, this means the expansion began in a state of infinite spacetime curvature, i.e., infinite tidal gravity, i.e., in a *"big-bang" singularity*. From the form $V = -(4\pi/3)a^2\rho$ for the effective potential [Eq. (27.39)] and the first law of thermodynamics (27.29), we see that the sufficient condition $dV/da \geq 0$ for the universe to have begun with a singularity is

$$\rho + 3P > 0 \; . \tag{27.101}$$

Cold matter and radiation satisfy this condition, but dark energy violates it. As we have seen, dark energy seems to have become important in our universe only recently, so it might not have been important at the universe's beginning. In this section we shall assume that it was not, and that the energy condition $\rho + 3P > 0$ was satisfied in the early universe. In the next section we shall discuss some consequences of a possible early-universe violation of $\rho + 3P > 0$.

The conclusion that the universe, if homogeneous and isotropic (and if $\rho + 3P > 0$), must have begun in a big-bang singularity, drove Yevgeny Lifshitz and Isaak Khalatnikov, students of Lev Landau in Moscow, to begin pondering in the late 1930s the issue of whether deviations from homogeneity and isotropy might have permitted the universe to avoid the singularity. A few events (the imprisonment of Landau for a year during Stalin's purges, then World War II, then the effort to rebuild Moscow, a nuclear weapons race with the United States, and other more urgent physics research) intervened, preventing the Lifshitz-Khalatnikov studies from reaching fruition until the early 1960s. However, after a great push in 1959–1961, Lifshitz and Khalatnikov reached the preliminary conclusion that early anisotropy and inhomogeneity could have saved the universe from an initial singularity: Perhaps the universe contracted from an earlier, large-scale state, then rebounded at finite size and finite curvature as a result of inhomogeneities and anisotropies. For a pedagogical presentation of the analysis which produced this conclusion see Landau and Lifshitz (1962).

The Lifshitz-Khalatnikov analysis was based on the mathematics of tensor analysis (differential geometry). In 1964 Roger Penrose (1965), a young faculty member at Kings College in London, introduced into general relativity an entire new body of mathematical techniques, those of differential topology, and used them to prove a remarkable theorem: that no matter how inhomogeneous and anisotropic an imploding star may be, if it implodes so far as to form a horizon, then it necessarily will produce a singularity of infinite spacetime curvature inside that horizon. Stephen Hawking and George Ellis (1968), at first graduate students and then research fellows at Cambridge University, by picking up Penrose's techniques and applying them to the early universe, proved that Lifshitz and Khalatnikov had to be wrong: The presently observed state of the universe plus reasonable constraints on its early equation of state imply that, regardless of any inhomogeneity or anisotropy, there must have been a singularity of infinite curvature. In response to these differential-topology analyses, Lifshitz, Khalatnikov, and their student Vladimir Belinsky reexamined their differential-geometry analyses, found an error, and discovered a possible structure for generic spacetime singularities. In this so-called *mixmaster structure*, as a freely moving observer approaches the singularity, inhomogeneities and anisotropies drive the tidal gravity (spacetime curvature) to oscillate in such a way that the observer feels an infinite, chaotic sequence of oscillations with ever growing amplitude, ever shortening period, and finite total proper-time duration. This is an example of the chaotic behavior which occurs frequently in nonlinear physics.

John Archibald Wheeler, a professor at Princeton University, realized in the mid 1950s that the singularities which began the big bang and terminate the implosion of a star cannot be classical: as one nears them, one must encounter a breakdown in classical general relativity, and new physics governed by the laws of *quantum gravity* (Wheeler 1957). Wheeler devised a simple argument to show that this is so, and to determine the critical radius of curvature of spacetime at which the transition to quantum gravity occurs:

Quantum theory insists that every field possess a half quantum of flucutational zero-point energy in each of its modes. Moreover, if one wishes to measure the average value of the field in a spacetime region with 4-volume L^4 (a region with side L along each of its 4 dimensions), one's measurements will be sensitive to the zero-point fluctuations of the modes that have wavelengths $\sim L$, but not to any others.

Now, so long as gravity is weak over the scale L, one can introduce a nearly Lorentz frame in the region L^4 and regard the deviations $\delta g_{\mu\nu} \equiv g_{\mu\nu} - \eta_{\mu\nu}$ of the metric coefficients $g_{\mu\nu}$ from the flat metric $\eta_{\mu\nu}$ as a nearly linear field that lives in nearly flat spacetime. This field must be just as subject to the laws of quantum mechanics as any other field. Its gravitational-wave modes with wavelength L have an energy density of order the square of the gradient of $\delta g_{\mu\nu}$, i.e., $\sim (\delta g_{\mu\nu}/L)^2$, and thus for these modes to contain a half quantum of unpredictable, fluctuational energy, they must have unpredictable fluctuations $\delta g_{\mu\nu}$ of the metric given by

$$\left(\frac{\delta g_{\mu\nu}}{L}\right)^2 L^3 \sim \frac{\hbar}{2L} \,. \tag{27.102}$$

Here the first term is the fluctuational energy density, L^3 is the 3-dimensional volume of the mode, and $\hbar/2L$ is its total fluctuational energy. Correspondingly, the mode's metric fluctuations are

$$\delta g_{\mu\nu} \sim \frac{\sqrt{\hbar}}{L} \,. \tag{27.103}$$

These fluctuations (which we have evaluated in the closest thing there is to a local Lorentz frame in our region L^4) guarantee that, whenever we try to measure a length L, we will make unavoidable errors with magnitude

$$\boxed{\frac{\delta L}{L} \sim \delta g_{\mu\nu} \sim \frac{\sqrt{\hbar}}{L} .} \tag{27.104}$$

The smaller is L, the larger are these fractional errors. When L is made smaller than $\sqrt{\hbar}$, the fractional errors exceed unity, there is no hope of measuring L at all (and our analysis also breaks down because we cannot introduce a nearly Lorentz frame throughout the region L^4). Thus, for a lengthscale L to be measureable, it must lie in the regime

$$\boxed{L \gtrsim L_{PW} , \quad \text{where } L_{PW} \equiv \sqrt{\hbar} = \left(\frac{G\hbar}{c^3}\right)^{1/2} = 1.616 \times 10^{-33} \text{ cm} .} \tag{27.105}$$

The critical lengthscale L_{PW} is called the Planck-Wheeler length. It is the shortest length that can possibly be measured with any accuracy. Thus, it is the smallest length that can be subjected to the classical laws of physics. Since gravity is characterized, classically, by the geometry of spacetime, classical gravity (i.e., general relativity) must break down on lengthscales shorter than L_{PW}. This should be true in the small in ordinary, nearly flat spacetime; and it also should be true near singularities: Near a singularity, when the radius of curvature of spacetime as predicted by classical general relativity becomes shorter than L_{PW}, general relativity must break down and be replaced by the correct quantum theory of gravity. And when quantum gravity comes into play, it may very well convert the singularity into something nonsingular.

Thus, to understand the true outcome of the gravitational implosion of a star, deep inside the horizon, one must understand quantum gravity; and to understand the initial conditions of the universe, one must understand quantum gravity.

The attempt to construct a quantum theory of gravity which unifies gravity with the strong, electromagnetic, and weak forces in an elegant and mutually consistent way is one of the "holy grails" of current theoretical physics research.

27.7 Inflationary Cosmology

If $\rho + 3P > 0$, then the universe is guaranteed to have cosmological horizons of the sort that we met when discussing acoustic oscillations in the era of recombination (Sec. 27.5.7).

The background radiation received at Earth today last interacted with matter (at a redshift $z \sim 10^3$) quite near our cosmological horizon. Two observers at the locations of that last interaction, one on our north celestial pole (i.e., directly above the north pole of the Earth) and the other on our south celestial pole (i.e., directly above the south pole of the Earth), are today far outside each others' cosmological horizons; and at the moment of that last interaction, they were enormously far outside each others' horizons. It is a great mystery how two regions of the universe, so far outside each others' horizons (i.e. with no

possibility for causal contact since the big bang) could have the same temperatures at the time of that last interaction, to within the measured accuracy of $\Delta T/T \sim 10^{-4}$.

One solution to this mystery is to assume that the universe emerged from the *Planck-Wheeler era* of quantum gravity in a very nearly homogeneous and isotropic state (but one with enough inhomogeneities to seed galaxy formation). This "solution" leaves to a future theory of quantum gravity the task of explaining why this state was nearly homogeneous and isotropic. An alternative solution, proposed by Alan Guth (1981), then a postdoctoral fellow at Stanford University, is *inflation*.

Suppose, Guth suggests, that the universe emerged from the Planck-Wheeler, quantum-gravity era, with its fields in a vacuum state for which $\mathbf{T}_{\rm vac} = -\rho_{\rm vac}\mathbf{g}$ was nonzero and perhaps was even as large in magnitude as $\rho_{\rm vac} \sim \hbar^{-2} \sim 10^{93}{\rm g/cm}^3$.

The expansion factor a presumably will have been of order L_{PW} when the universe emerged from the Planck-Wheeler era; and the evolution equation (27.30) predicts that it subsequently will expand classically in accord with the law

$$\boxed{a = L_{PW} \exp\left[\left(\frac{4\pi}{3}\rho_\Lambda\right)^{1/2} t\right] = L_{PW} \exp\left(\frac{t}{\mu L_{PW}}\right)} \ , \tag{27.106}$$

where μ is a dimensionless constant that might be of order unity. This exponential expansion under the action of vacuum stress-energy is called "inflation;" and if it lasted long enough, that means our entire universe was so small in the early stages of inflation that it could easily communicate with itself, producing homogeneity and isotropy.

Of course, inflation at this enormous rate could not have lasted forever; it surely is not continuing today. If it occurred at all, it must have turned off at some point as a result of the fields undergoing a phase transition from the original vacuum state (sometimes called the "false vacuum") to a new vacuum state in which $\rho_{\rm vac}$ is zero, or perhaps equal to the tiny ρ_Λ that we observe today.

Although these ideas seem speculative, they have been made quite plausible by two factors: (*i*) they fit naturally into present ideas about the physics of the grand unification of all forces; and (*ii*) they successfully explain a number of mysterious features of the universe in which we live, including its spatial flatness, the high degree of isotropy of the background radiation (Ex. 27.11), and the flat (wavelength-independent) spectrum of rms density fluctuations that ultimately condensed into galaxies. For details see, e.g., Kolb and Turner (1994).

27.7.1 Amplification of Primordial Gravitational Waves by Inflation

This section is not yet written.

27.7.2 Search for Primordial Gravitational Waves by their Influence on the CMB; Probing the Inflationary Expansion Rate

This section is not yet written.

****

EXERCISES

Exercise 27.11 *Practice: Inflationary explanation for the isotropy of the cosmological background radiation*
Consider an inflationary cosmological model in which (i) the expansion factor inflates as $a = L_{PW} \exp(t/\mu L_{PW})$ until it has e-folded $N \gg 1$ times, i.e., from time $t = 0$ (when it emerges from the Planck-Wheeler era) to time $t = N\mu L_{PW}$, and then (ii) a phase transition drives a into the standard expansion produced by radiation with $P = \rho/3$: an expansion with $a \propto t^{1/2}$ [Eq. (27.45)]. Show that in this model, if the number of e-folding times during inflation is $N \gg 70$, then the north-celestial-pole and south-celestial-pole regions which emit the background radiation we see today are actually inside each others' cosmological horizons: They were able to communicate with each other (and thereby, the inflationary scenario suggests, were able to homogenize to the same temperature) during the inflationary era. *Hint*: the number of e-foldings required is given analytically by

$$N \gg \ln \left[\frac{{H_o}^{-1}}{L_{PW}} \left(\frac{\rho_o}{\Lambda} \right)^{1/4} \right] \simeq 70 \; . \tag{27.107}$$

****

Bibliographic Note

For an elementary introduction to cosmology, we recommend Chaps. 17, 18, 19 of Hartle (2003); and at an intermediate level, similar to this Chap. 27, we recommend Chap. 8 of Carroll (2004). Textbook treatments of cosmology written before about 1995 are rather out of date, so one should only consult the standard old relativity texts such as MTW (1973) and Weinberg (1972) for the most basic ideas.

For physical processes in the early universe such as dark matter, inflation and phase transitions, we recommend Kolb and Turner (1994). Peebles (1993) is an excellent, but a bit out of date, treatise on all aspects of cosmology. More up to date treatises include Dodelson (2003) and Ryden (2002).

Bibliography

Balbi, A., et. al., 2000. "Constraints on cosmological parameters from MAXIMA-1," *Astrophysical Journal Letters*, submitted. astro-ph/0005124.

Bennet, C.L., et. al., 2003. "First Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Preliminary Maps and Basic Results", *Astrophysical Journal Letters*, submitted. astro-ph/0302207

Box 27.5
Important Concepts in Chapter 27

- Homogeneity and isotropy of universe, and its mathematical description via hypersurfaces, synchronous coordinates, Robertson-Walker line element, and three spatial geometries (closed, flat and open), Sec. 27.2
  - Homogeneous observers and their local Lorentz frame, Secs. 27.2, 27.3, 27.5.2
- Functions describing evolution of universe: expansion factor $a(t)$ and total density of mass-energy $\rho(t)$, Secs. 27.2, 27.3
  - Evolution laws for $\rho(t)$ and $a(t)$: first law of themodynamics, and Einstein equation for expansion rate, Sec. 27.3
  - Critical density to close the universe, $\rho_{\rm crit}$, Eq. (27.35)
  - Effective potential for expansion of universe and qualitative and quantitative forms of $a(t)$, Secs. 27.4.3, and 27.4.4
- Constituents of the universe: baryonic matter, cold dark matter, radiation, and dark energy; and their evolution as functions of the universe's expansion factor a, Secs. 27.4.1, 27.4.3, 27.5.8
  - $\Omega \equiv \rho/\rho_{\rm crit}$ and its measured values for constituents, Secs. 27.4.3, 27.5.4–27.5.7
  - Stress-energy tensor for the vacuum, cosmological constant, and their possible role as the dark energy, Sec. 27.4.1, Box 27.2
  - Radiation temperature and cosmological redshift as functions of a, Sec. 27.4.4
  - Preservation of Planckian spectrum during evolution, Box 27.4
- Physical processes during expansion: baryon-antibaryon annihilation, electron-positron annihilation, primordial nucleosynthesis, radiation-matter equality, plasma recombination, galaxy formation, Secs. 27.4.5, 27.5.4
- Observational parameters: Hubble expansion rate H_o, Ω for constituents, spatial curvature k/a_o^2, deceleration parameter q_o, age of universe, t_o, Secs. 27.5.1, 27.5.3
  - Measured values and methods of measurement, Secs. 27.5.3–27.5.9
  - Distance-redshift relation, Sec. 27.5.3
  - Angular-diameter distance as function of redshift, Eq. (27.78) and Ex. 27.8
  - Anisotropy of the CMB; Doppler peaks, and their use to measured the spatial geometry of the universe and thence Ω, Sec. 27.5.7, Fig. 27.6
  - Ages of the universe constrains equation of state of dark energy, Sec. 27.5.8
  - Luminosity distance; magnitude-redshift relation, Sec. 27.5.9, Ex. 27.9
- Big-bang singularity, Planck-Wheeler length and quantum gravity, Sec. 27.6
- Inflation, Sec. 27.7

Carroll, S.M., 2004. *Spacetime and Geometry: An Introduction to General Relativity*, San Francisco: Addison-Wesley.

Dicke, Robert H., Peebles, P. James E., Roll, Peter G., and Wilkinson, David T., 1965. "Cosmic-black-body radiation," *Astrophysical Journal*, **142**, 414–419.

Dodelson, S., 2003. *Modern Cosmology*, Academic Press.

Einstein, Albert, 1917. "Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie," *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, **1917 volume**, 142–152.

Einstein, Albert, 1931. "Zum kosmologischen Problem der allgemeinen Relativitätstheorie," *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften, phys.-math. Kl.*, **1931 volume**, 235–237.

Friedmann, Alexander Alexandrovichi, 1922. "Über die Krümmung des Raumes," *Zeitschrift für Physik*, **10**, 377–386.

Guth, Alan H., 1981. "Inflationary universe: A possible solution to the horizon and flatness problems," *Physical Review*, **23**, 347–356.

Hartle, J. B., 2003. *Gravity: An Introduction to Einstein's General Relativity*, San Francisco: Addison-Wesley.

Hawking, Stephen W. and Ellis, George F. R., 1968. "The cosmic black body radiation and the existence of singularities in our universe," *Astrophysical Journal*, **152**25–36.

Hubble, Edwin Powell, 1929. "A relation between distance and radial velocity among extragalactic nebulae," *Proceedings of the National Academy of Sciences*, **15**, 169–173.

Kolb, Edward W., Turner, Michael S., 1994. *The Early Universe*, Reading: Addison-Wesley.

Kuo, C.L. et. al. 2002. *Astrophysical Journal*, in press. astro-ph/0212289

Landau, Lev Davidovich, and Lifshitz, Yevgeny Michailovich, 1962. *The Classical Theory of Fields*, Addison Wesley, Reading, MA.

Lange, Andrew E. et. al., 2000. "First estimations of cosmological parameters from Boomerang," *Physical Review Letters*, submitted. astro-ph/000504.

MTW: Misner, Charles W., Thorne, Kip S. and Wheeler, John A., 1973. *Gravitation*, W. H. Freeman & Co., San Francisco.

Pais, Abraham, 1982. *'Subtle is the Lord...' The Life and Science of Albert Einstein*, Oxford University Press, New York.

Pearson, T.J. et. al., 2002. *Astrophysical Journal*, submitted. astro-ph/0205288

Penrose, Roger, 1965. *Gravitational collapse and space-time singularities*, *Physical Review Letters*, **14**, 57–59.

Penzias, Arno A., and Wilson, Robert W, 1965. "A measurement of excess antenna temperature at 4080 Mc/s," *Astrophysical Journal*, **142**, 419–421.

Peebles, P. J. E., 1993. *Principles of Physical Cosmology*, Princeton: Princeton University Press.

Perlmutter, S. et. al., "Measurements of Ω and Λ from 42 high-redshift supernovae," *Astrophysical Journal*, **517**, 565–586 (1999).

Riess, A. G. et. al., 1998. "Observational evidence from supernovae for an accelerating universe and a cosmological constant," *Astronomical Journal*, **116**, 1009.

Ryan, Michael, and Shepley, Lawrencei, 1975. *Homogeneous, Relativistic Cosmology*, Princeton University Press, Princeton, NJ.

Ryden, B.S., 2002. *Introduction to Cosmology*, San Francisco: Addison-Wesley.

Robertson, Howard Percy, 1935. "Kinematics and World Structure," *Astrophysical Journal*, **82**, 248–301; and **83**, 287–201 & 257–271.

Sakharov, Andrei D., 1965. *Zhurnal Eksperimentalnoi i Teoreticheskii Fisika*, **49**, 345.

Sunyaev, Rashid A., and Zel'dovich, Yakov B., 1970. "Small-scale fluctuations of relic radiation," *Astrophysics and Space Science*, **7**, 3–19.

Turner, Michael S., 1999. in *Proceedings of Particle Physics and the Universe (Cosmo-98)*, ed. D. O. Caldwell, AIP, Woodbury, NY; astro-ph/9904051.

Walker, Arthur Geoffrey, 1936. "On Milne's theory of world-structure," *Proceedings of the London Mathematical Society*, **42**, 90–127.

Weinberg, S., 1972. *Gravitation and Cosmology*, New York: John Wiley.

Wheeler, John Archibald, 1957, "On the nature of quantum geometrodynamics," *Annals of Physics*, **2**, 604–614.

Zel'dovich, Yakov Borisovich, 1968. "The cosmological constant and the theory of elementary particles," *Soviet Physics—Uspekhi*, **11**, 381–393.

Zel'dovich, Yakov Borisovich, and Novikov, Igor Dmitrivich, 1983. *Relativistic Astrophysics Volume 2: The Structure and Evolution of the Universe*, University of Chicago Press, Chicago.

Ph 136a 25 April 2007

CHAPTER 23

Reading: Chapter 23, *From Special to General Relativity.*

Students who were not in this class first term may need to read some sections of Chapter 1, which are cross referenced in Chapter 23. Students who were not in this class second term should read the introduction to Connection Coefficients in Sec. 10.3.

Problems

NOTE: The students in this class have a wide variety of backgrounds in relativity theory, so problems that are appropriate for some students are inappropriate (too sophisticated or too elementary) for others. Choose four problems appropriate for you from the following selection.

A. Ex. 23.1 Invariance of a Null Interval AND Ex. 23.2: Causality
B. Ex. 23.4: Index manipulation rules from duality
C. Parts (a) and (b) of Ex. 23.6: Commutation and connection coefficients for circular polar bases. Also part (b) of Ex. 23.5: Transformation matrices for circular polar bases
D. Ex. 23.9: Index gymnastics — irreducible tensorial parts of the gradient of a 4-velocity field
E. Ex. 23.10: Integration — Gauss's theorem
F. Ex. 23.11: Stress-energy tensor for a perfect fluid
G. Ex. 23.14: Stress-energy tensor for a point particle
H. Ex. 23.15: Proper reference frame
I. Ex. 23.16 Uniformly accelerated observer

Ph 136a 3 May 2007

CHAPTER 24

Reading: Chapter 24, *Fundamental Concepts of General Relativity.*

Problems

Work four of the following ten problems. Pick problems that are appropriate for *you*—not too easy; not too hard; different from anything you have done in previous courses. Note: If you are really eager to learn general relativity, you may want to work more than four problems, but please indicate which ones you want to be graded.

Exercise 24.3 Geodesic equation in an arbitrary coordinate system
Exercise 24.4 Constant of geodesic motion in a spacetime with symmetry
Exercise 24.5 Action principle for geodesic motion
Exercise 24.7 Orders of magnitude of the radius of curvature
Exercise 24.8 Components of Riemann in an arbitrary basis
Exercise 24.9 Curvature of the surface of a sphere
Exercise 24.10 Geodesic deviation on a sphere
Exercise 24.12 Newtonian limit of general relativity
Exercise 24.13 Gauge transformations in linearized theory
Exercise 24.14 External field of a stationary, linearized source

CHAPTER 25

Reading: Chapter 25, *Relativistic Stars and Black Holes*.

Problems

Work four of the following nine problems. Pick problems that are appropriate for *you*—not too easy; not too hard. I especially recommend the exercises in bold face

- **Exercise 25.1 Connection Coefficients and Riemann in Schwarzschild.** Note: If you know Mathematica or Maple fairly well, you can fairly easily write your own programs to compute connection coefficients and curvature tensors. You will find programs written by others at the following web sites:
  * The site associated with James Hartle's textbook on general relativity: http://wps.aw.com/aw_hartle_gravity_1/0,6533,512494-,00.html
  * The GRTensor web site: http://grtensor.phy.queensu.ca/
- **Exercise 25.2 The Bertotti-Robinson Solution**
- Exercise 25.3 Schwarzschild Geometry in Isotropic Coordinates
- Exercise 25.4 Star of Uniform Density
- **Exercise 25.5 Gravitational Redshift**
- Exercise 25.7 Implosion of the Surface of a Zero-Pressure Star
- **Exercise 25.8 Gore at the Singularity**
- **Exercise 25.9 Wormholes**
- **Exercise 25.11 Penrose Process, Hawking Radiation, & Black-Hole Thermodynamics**
- Exercise 25.12 Slices of Simultaneity in Schwarzschild Spacetime

Ph 136a 16 May 2007

CHAPTER 26

Reading: Chapter 26, *Gravitational Waves and Experimental Tests of General Relativity.*

Problems

Work four of the following eight problems. Pick problems that are appropriate for *you*—not too easy; not too hard.

A. Exercise 26.4 Behavior of h_+ and $h_\times$ under Rotations and Boosts
B. Exercise 26.5 Energy-momentum conservation in geometric optics limit
C. Exercise 26.6 Transformation to TT gauge
D. Exercise 26.9 Energy removed by gravitational radiation reaction
E. Exercise 26.10 Propagation of waves through an expanding universe
F. Exercise 26.11 Gravitational waves emitted by a linear oscillator
G. Exercise 26.12 Gravitational waves from waving arms
H. Exercise 26.14 Light in an interferometric gravitational wave detector in TT gauge

Ph 136a 23 May 2003

CHAPTER 27

Reading: Chapter 27, *Cosmology*.

Problems

Please put your solutions in the box beside Kip's office door, room 154 West Bridge, before 1PM Wednesday 30 May.

Work any four of the following problems:

Exercise 27.2 The 3-Sphere Geometry of a Closed Universe
Exercise 27.3 Energy Conservation for a Perfect Fluid
Exercise 27.5 Einstein's Static Universe
Exercise 27.6 Cosmological Redshift
Exercise 27.7 Cosmic Microwave Radiation in an Anisotropic Cosmological Model
Exercise 27.8 Angular-Diameter Distance
Exercise 27.9 Magnitude-Redshift Relation
Exercise 27.11 Inflationary Explanation for the Isotropy of the CMB

CHAPTER 23

Reading: Chapter 23, *From Special to General Relativity.*

Students who were not in this class first term may need to read some sections of Chapter 1, which are cross referenced in Chapter 23. Students who were not in this class second term should read the introduction to Connection Coefficients in Sec. 10.3.

Problems

NOTE: The students in this class have a wide variety of backgrounds in relativity theory, so problems that are appropriate for some students are inappropriate (too sophisticated or too elementary) for others. Choose four problems appropriate for you from the following selection.

A. Ex. 23.1 Invariance of a Null Interval AND Ex. 23.2: Causality
B. Ex. 23.4: Index manipulation rules from duality
C. Parts (a) and (b) of Ex. 23.6: Commutation and connection coefficients for circular polar bases. Also part (b) of Ex. 23.5: Transformation matrices for circular polar bases
D. Ex. 23.9: Index gymnastics — irreducible tensorial parts of the gradient of a 4-velocity field
E. Ex. 23.10: Integration — Gauss's law
F. Ex. 23.11: Stress-energy tensor for a perfect fluid
G. Ex. 23.14: Stress-energy tensor for a point particle
H. Ex. 23.15: Proper reference frame
I. Ex. 23.16 Uniformly accelerated observer

CHAPTER 24

Reading: Chapter 24, *Fundamental Concepts of General Relativity.*

Problems

Work four of the following ten problems. Pick problems that are appropriate for *you*—not too easy; not too hard; different from anything you have done in previous courses. Note: If you are really eager to learn general relativity, you may want to work more than four problems, but please indicate which ones you want to be graded.

Exercise 24.3 Geodesic equation in an arbitrary coordinate system
Exercise 24.4 Constant of geodesic motion in a spacetime with symmetry
Exercise 24.5 Action principle for geodesic motion
Exercise 24.7 Orders of magnitude of the radius of curvature
Exercise 24.8 Components of Riemann in an arbitrary basis
Exercise 24.9 Curvature of the surface of a sphere
Exercise 24.10 Geodesic deviation on a sphere
Exercise 24.12 Newtonian limit of general relativity
Exercise 24.13 Gauge transformations in linearized theory
Exercise 24.14 External field of a stationary, linearized source

CHAPTER 25

Reading: Chapter 25, *Relativistic Stars and Black Holes.*

Problems

Work four of the following nine problems. Pick problems that are appropriate for *you*—not too easy; not too hard. I especially recommend the exercises in bold face

- **Exercise 25.1 Connection Coefficients and Riemann in Schwarzschild.** Note: If you know Mathematica or Maple fairly well, you can fairly easily write your own programs to compute connection coefficients and curvature tensors. You will find programs written by others at the following web sites:
  * The site associated with James Hartle's textbook on general relativity: http://wps.aw.com/aw_hartle_gravity_1/0,6533,512494-,00.html
  * The GRTensor web site: http://grtensor.phy.queensu.ca/
- **Exercise 25.2 The Bertotti-Robinson Solution**
- Exercise 25.3 Schwarzschild Geometry in Isotropic Coordinates
- Exercise 25.4 Star of Uniform Density
- **Exercise 25.5 Gravitational Redshift**
- Exercise 25.7 Implosion of the Surface of a Zero-Pressure Star
- **Exercise 25.8 Gore at the Singularity**
- **Exercise 25.9 Wormholes**
- **Exercise 25.11 Penrose Process, Hawking Radiation, & Black-Hole Thermodynamics**
- Exercise 25.12 Slices of Simultaneity in Schwarzschild Spacetime

Ph 136a 18 May 2005

CHAPTER 26

Reading: Chapter 26, *Gravitational Waves and Experimental Tests of General Relativity.*

Problems

Work four of the following eight problems. Pick problems that are appropriate for *you*—not too easy; not too hard.

A. Exercise 26.4 Behavior of h_+ and $h_\times$ under Rotations and Boosts
B. Exercise 26.5 Energy-momentum conservation in geometric optics limit
C. Exercise 26.6 Transformation to TT gauge
D. Exercise 26.9 Energy removed by gravitational radiation reaction
E. Exercise 26.10 Propagation of waves through an expanding universe
F. Exercise 26.11 Gravitational waves emitted by a linear oscillator
G. Exercise 26.12 Gravitational waves from waving arms
H. Exercise 26.14 Light in an interferometric gravitational wave detector in TT gauge

Ph 136a 25 May 2003

CHAPTER 27

Reading: Chapter 27, *Cosmology*.

Problems

There is no formal assignment for this week. Here is a suggested set of problems, for which solutions are being provided.

Exercise 27.2 The 3-Sphere Geometry of a Closed Universe
Exercise 27.3 Energy Conservation for a Perfect Fluid
Exercise 27.5 Einstein's Static Universe
Exercise 27.6 Cosmological Redshift
Exercise 27.7 Cosmic Microwave Radiation in an Anisotropic Cosmological Model
Exercise 27.8 Angular-Diameter Distance
Exercise 27.9 Magnitude-Redshift Relation
Exercise 27.11 Inflationary Explanation for the Isotropy of the CMB

Physics 136a December 2004

FINAL EXAM

You will be expected to answer 8 of the following questions in three hours, closed book. The questions might carry different weights. When you are ready to take the exam, pick up a list of question numbers and their weights from Shirley Hampton, Room 151 Bridge Annex, or from JoAnn Boyd, Room 161 West Bridge. Please write your solutions in a blue book, which can be purchased at the Caltech bookstore for a nominal cost. Completed exams must be placed in Kip Thorne's mailbox at the west end of Bridge Annex before 5:00 PM Friday, December 10. Late exams will not be accepted.

Grades in this course will be determined as follows:

(i) Students who score 60% or more on the homeworks will pass the course without having to take the final. Those students who have petitioned for letter grades will receive a grade based on the homework, plus the exam if the exam is taken.

(ii) Students who are failing on the basis of homeworks and those who wish to improve their letter grades should take this final exam. For such students, the grade will be determined by a combination of homework performance and exam performance, with the proviso that the grade will be no lower than would be received based solely on homework scores.

1. Using a spacetime diagram, exhibit the Lorentz contraction of a moving, rigid rod, the time dilation of a moving clock, and the fact that the concept of simultaneity is not frame invariant.

2. Explain the geometrical meanings of the *direction* of a volume vector $\vec{\Sigma}$ and the *sense* of $\vec{\Sigma}$. What is the relationship between the direction and the sense for the case of a timelike $\vec{\Sigma}$, a spacelike $\vec{\Sigma}$, and a null $\vec{\Sigma}$?

3. Without using any coordinate system or basis, explain the meaning of "stress-energy tensor", "electromagnetic field tensor", and "Levi-Civita tensor".

4. Show that the number density of particles in phase space is Lorentz invariant (i.e. is the same in all reference frames, independently of the frames' speed of motion).

5. Millions of asteroids move in orbits around the Sun, which has a Newtonian gravitational field $\Phi(\mathbf{x})$. Ignoring collisions between the asteroids, prove that their kinetic-theory distribution function satisfies the collisionless Boltzman equation.

6. A white dwarf star made of pure hydrogen gas has almost all its mass provided by the hydrogen's protons and almost all its pressure by the hydrogen's electrons. The star is in thermodynamic equilibrium with an electron chemical potential $\mu = \tilde{\mu} + m_e$ that varies with radius but lies in the range $0 < \tilde{\mu} \ll m_e$, and with a temperature $T \ll \tilde{\mu}/k$. Here k is Boltzmann's constant and m_e is the electron rest mass. Sketch a derivation of the equation of state $P = P(\rho)$ of the star's matter

7. For photons derive the relationship between the number density in phase space, the mean occupation number, and the specific intensity.

8. Explain what is meant by a "transport coefficient" and sketch the steps by which one can compute transport coefficients from the Boltzmann transport equation.

9. The shear of a (nonrelativistic) fluid is defined to be the trace-free part of the symmetrized gradient of its velocity, $\sigma_{ij} = v_{(i,j)} - \frac{1}{3}\delta_{ij}v_{k,k}$. Viscosity resists this shear by producing a stress $T_{ij} = -2\eta\sigma_{ij}$. Here η is the coefficient of shear viscosity. Make a rough estimate of η for photons in a shearing hydrogen gas at a temperature $T = 10^5$K and density $\rho = 10^{-16}\mathrm{g\,cm^{-3}}$.

10. What is the statistical equilibrium distribution function that describes an ensemble of systems of the following type: Each system is a collection of photons in the interior of a spherical container whose walls freely emit and absorb the photons. The container has volume V and temperature T, and it rotates with angular velocity Ω. What is the distribution function for a subensemble of this ensemble in which each subsystem is a mode of the electromagnetic field that has angular frequency ω and spherical harmonic indices l, m?

11. How would one measure the chemical potential of the nitrogen molecules in a room full of air?

12. Explain what is meant by the ergodic hypothesis in statistical mechanics, and give one concrete example of how it could be applied.

13. Explain what is meant by canonical ensemble, grand canonical ensemble, microcanonical ensemble, and Gibbs ensemble. Give an example of a system whose fluctuations away from equilibrium are characterized by the canonical ensemble.

14. Explain quantitatively the relationship between the second law of thermodynamics and the principle that, when a system held at fixed pressure and temperature can exist in two phases, it will evolve toward the phase with minimum Gibbs potential.

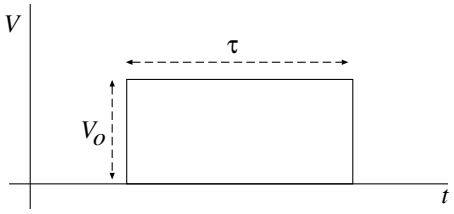
15. Explain the difference between a first-order phase transition and a second-order phase transition, and give one example of each.

16. A mixture of hydrogen and oxygen is inserted into a chamber with rigid, perfectly insulating walls. A spark ignites an explosion in which much of the hydrogen and oxygen forms water. Since no heat is added to or removed from the system through its walls, and one normally says that (heat added) = (temperature)×(change in entropy), one might think that the explosion leaves the system's entropy unchanged. Explain why this is not so, and explain how you would go about computing how much water is formed and how much entropy increase accompanies that formation.

17. Explain the meaning of correlation function and spectral density and the relationship between them. If a correlation function has a relaxation time τ_r, what does that imply about the form of its spectral density?

18. A 100 gram pendulum with a 1Hz swinging frequency is set swinging with a large amplitude in air at room temperature (300 K). The air molecules produce drag on the pendulum, causing its swinging to die out with an amplitude e-folding time $\tau_* = 1$ hour. The pendulum then is placed absolutely at rest and absolutely vertical, and its subsequent motion is measured. Describe that subsequent motion in as much detail as you can, and as quantitatively as you can.

19. A physics experiment produces a very weak voltage signal whose time dependence has the following form:



This voltage signal is contaminated by noise with spectral density

$$S_V(f) = \frac{\text{constant}}{f^2}$$

The experimenter wants to know whether the signal is actually present or not, and if present, what its amplitude is. To answer this question, the experimenter performs optimal signal processing. Sketch the form of the optimal filter $K(t)$ that the experimenter should use, and explain in simple terms why the optimal filter has this form.

20. Give one specific example of each of the following types of random processes: A Gaussian process; a non-Gaussian process; a Markoff process; a non-Markoff process; a stationary process; a non-stationary process.

21. An experimenter measures the average value $\bar{y}_T$ of some random process y over a time interval T. The experimenter repeats this process over and over again, and finds a variance $\sigma_{\bar{y}_T}$ for this quantity. What properties must the spectral density of y have in order for $\sigma_{\bar{y}_T}$ to have the following behaviors as T goes to infinity: (i) $\sigma_{\bar{y}_T} \to 0$? (ii) $\sigma_{\bar{y}_T} \to \infty$?

22. Explain what is meant by "Fokker-Planck equation". Give a specific example of a physical system whose evolution is governed by the Fokker-Planck equation, and describe qualitatively the evolution.

23. State the conditions under which geometric optics is valid and use the eikonal approximation to derive the ray equations for $\mathbf{k}, \mathbf{x}, \omega$.

24. State Fermat's principle and use it to show that a wave propagating in a stratified atmosphere of refractive index $n(z)$ propagates at an angle $\propto \text{cosec}^{-1}(n)$ to the direction of the density gradient.

25. Use matrix methods to calculate the magnification of a simple refracting telescope comprising two thin converging lenses.

26. Distinguish between Fraunhofer and Fresnel diffraction and estimate the distance beyond which we can form a Fraunhofer diffraction pattern on a screen when visual light passes through a 1 mm aperture.

27. A grating comprises a large number of parallel strips of constant width w. Alternate strips retard a normally incident plane wave by phase π. Sketch the Fraunhofer diffraction pattern. Explain qualitatively the effect of reducing the number of strips.

28. Explain what is meant by the visibility of interference fringes and show that the normalized brightness distribution of a distant, extended, incoherent source is the Fourier transform of the visibility of the interference fringes formed by a pair of Young's slits.

29. Define and show that you know what is meant by the degrees of spatial and temporal coherence.

30. A visual light spectral line of width 0.1nm from a star at ~ 10 light years is received at earth. Estimate the longitudinal and lateral coherence lengths.

31. A delta function pulse from a satellite is scattered through an rms angle $\theta \sim 1$ arcsec by a layer of turbulence in the atmosphere some 5km above ground. Estimate the duration of the received pulse, the Fresnel length and the horizontal scale of the diffraction pattern created on the ground

32. Explain what is meant by a caustic and show how one can be formed in reflection by a spherical mirror.

33. Explain what is meant by Babinet's Principle and give one example of its application.

34. It is proposed to construct a 16m optical telescope on the lunar surface. Estimate the linear resolution achievable when observing a star at a distance of 10 light years.

35. A distant point source is observed through a circular aperture at a point on the optic axis at a distance z. Explain what is meant by a Fresnel half period zone and show that the radius of the n'th zone is $(n\lambda z)^{1/2}$, where λ is the wavelength.

36. A coherently illuminated source is imaged by a single converging lens. Explain qualitatively how the image will be modified if a small circular aperture (low pass filter) is placed in the back focal plane of the lens. What happens if the aperture is replaced by a circular disk (high pass filter)?

37. Outline the physical principles underlying a phase contrast microscope, and describe the conditions under which it is useful.

38. A plane wave of frequency ω is incident upon a Fabry-Perot etalon of thickness d and refractive index n. The wavefronts make an angle θ with the surface of the etalon. Show that the phase difference between radiation that it is transmitted at both surfaces and radiation that is reflected once at each surface is given by

$$\phi = 2n\omega d \cos\theta / c$$

Hence explain qualitatively how an etalon can be used as a spectrometer.

39. Explain what is meant by phase conjugation in optics, explain its relationship to time reversal, and describe one possible application.

40. Explain what is meant by "squeezed light" and describe an experimental setup that will produce it.

41. Explain how nonlinear optics could be used to *triple* the frequency of a monochromatic light beam. What order of nonlinearity would be required (e.g., how many tensor indices) for the dielectric susceptibility? Suppose (as is normally the case) that the dielectric medium has dispersion. How would that dispersion impede the frequency tripling, and how might this impediment be surmounted?

42. In three-wave mixing, two waves (with frequencies and wave numbers ω_1, ω_2 and $\mathbf{k}_1$, $\mathbf{k}_2$) interact nonlinearly to produce a third wave ω_3, $\mathbf{k}_3$. Sketch a derivation of the relationships between ω_1, ω_2, ω_3 and between $\mathbf{k}_1$, $\mathbf{k}_2$, $\mathbf{k}_3$. Interpret these relationships in terms of photon energies and momenta. Suppose that $\omega_1 > \omega_2 > \omega_3$. Which of the waves attenuate due to the three-wave mixing and which ones grow?

43. Explain how holography works and discuss its relationship to phase conjugation.