

A SPACETIME PRIMER

T. A. Jacobson

July 30, 1998

Contents

1 Spacetime	3
1.1 Differential structure	3
1.2 Spacetime diagrams	4
1.3 Causal structure	5
1.4 Metrical structure	8
2 Mathematical formulation	11
2.1 The tangent space	11
2.2 The line element	13
2.3 Local inertial coordinates and curvature	15
2.4 Metric = causal cone + scale	17
2.5 Deep background	18
2.6 Problems	21
3 Free-fall and Geodesics	25
3.1 Curves	25
3.2 Inertial motion	26
3.3 Lightlike free-fall and geodesics	28
3.4 Conserved quantities along a geodesic	30
3.5 Field theory in curved spacetime	30
3.6 Problems	31
4 Special Relativity	35
4.1 Minkowski space	35
4.2 4-velocity and 4-acceleration	36
4.3 4-momentum	37
4.4 Voyage to the galactic center	38

Chapter 1

Spacetime

We begin by introducing the concept of spacetime, and the structures it is assumed to possess. First the discussion will use just words and pictures, to give a feeling for what is going on. Afterwards, these ideas will be given a precise mathematical formulation.

Spacetime is the collection of all *events*. An event is a “place and time”. Nothing special has to happen there and then in order for an event to be an “event”. In any case, physics today is based on quantum field theory, and quantum fields permeate all of spacetime with, if nothing else, vacuum fluctuations. So something is always “occurring” at an event. Moreover, as we shall see, the spacetime metric tensor is a dynamical field that takes values at every event.

1.1 Differential structure

It is assumed that the events form a 4-dimensional *continuum*, or *manifold*. That is, they can be put into 1-1 correspondence with 4-tuples of real numbers, called *coordinates*. Coordinates that are related by smooth invertible functions are all on the same footing. The spacetime need not be covered by any single coordinate system. Rather, it may be covered by more than one patch, as long as where the patches overlap, the coordinates are related by a smooth, invertible transformation. A maximal collection of smoothly related coordinate patches is called a *differentiable structure*.

We have so far assumed that spacetime *can be* equipped with a 4-dimensional differentiable structure. Actually, the relevant assumption is that spacetime possesses a *particular* differentiable structure. This is an important distinction, since a given set of points can be given many different differentiable structures,

even with different dimensions. For instance, the set of 4-tuples of real numbers can be put into 1-1 correspondence with the set of n -tuples for any n .

As an example of a physically defined coordinate system, consider an idealized form of the global positioning system. Adopt four satellites with precise clocks on board that orbit the earth, and continuously transmit signals coded with the time on the clock at which they were sent. Within some spacetime region, every event E will be reached by a signal from each of the four satellites, and E can be labeled by the four times at which those signals were sent. Within some open range of clock times, every 4-tuple of clock times will determine a unique event. Thus these four times provide a coordinatization of spacetime, at least in some region.¹ (The coordinate system will break down somewhere if there is enough spacetime curvature to ruin the 1-1 nature of the labeling.)

This example was designed to illustrate the fact that the four coordinates of spacetime need not be thought of as one time and three space coordinates. There *is* a sense in which $4=1+3$, but it arises from the nature of the causal structure in spacetime.

1.2 Spacetime diagrams

Pictures illustrating relationships in spacetime can be drawn in perspective or in a plane, by suppressing one or two spacelike dimensions respectively. These pictures can be very helpful in appreciating spacetime relationships.

As an example, Figure 1.1 depicts the earth with a satellite in orbit around it, and another satellite with a thruster rocket, hovering without orbiting at the altitude of the orbiting satellite. The curve indicating the history of each satellite is called the *world line* of the satellite.

As another example, Figure 1.2 illustrates the 4-time coordinate system discussed above. Suppressing one spatial dimension, only three satellites are required. The world lines of three satellites are indicated, together with an

¹To go from these four times to the latitude and longitude of a receiver on the earth involves a calculation that takes into account a model of the orbital dynamics of the satellites and the rotation of the earth, and involves important contributions from both gravitational and relative motion effects of general and special relativity. As a further aside, it is amusing to note that a similar (but redundant) coordinate system has been chosen in an attempt to communicate, to extraterrestrials, where and when we are located: On a plaque carried by the *Pioneer 10* and *11* spacecrafts, which left the solar system in the 1970's, our location was specified by indicating the then current rotation periods of fourteen pulsars. As these periods are very stable, but gradually lengthening, they provide good clocks for this purpose.

Figure 1.1: earth

event E . Emanating down from E is a cone composed of those events that are connected to E by light signals. The clock times t_1, t_2, t_3 where this cone cuts the satellite world lines are the coordinates of the event E .

Figure 1.2: 4times

1.3 Causal structure

Perhaps more fundamental than the differential structure is the *causal structure* of spacetime. The causal structure specifies for every pair of events A and B whether A can influence B , or B can influence A , or neither. These three possibilities are mutually exclusive.² The causal order is transitive, in the sense that if A can influence B , and B can influence C , then A can influence C .

Because of transitivity, it is not necessary to specify the causal relations between all pairs of events. Rather, from relations between events in localized neighborhoods covering the spacetime, relations between more widely separated events are determined by transitivity. In the continuum model of spacetime, it

²Unless, of course, there are closed timelike loops in spacetime, which is a possibility that is sometimes considered. For a recent review see, for example, K. S. Thorne, in *General Relativity and Gravitation 1992*, eds. R. J. Gleiser, C. N. Kozameh, and O. M. Moreschi (Institute of Physics Publishing, 1993), p 295.

suffices to know the causal relations between each event E and the events in an *infinitesimal* neighborhood E .

In relativistic spacetime, the causal structure in a neighborhood of each event E is determined by a 3-dimensional surface having the topology of a double cone whose vertex is E . This cone is called the *causal cone* or *light cone* at E . One half of the cone is called the *future cone*, and the other, the *past cone*. Only events lying inside or on the future cone can be influenced by E , and E can only be influenced by events inside or on the past cone. The collection of these cones for all events defines the *causal structure* in the spacetime. Actually, the causal structure itself really consists of the collection of light cones extending out only to an infinitesimal neighborhood of each event, since the global causal relations can be built up from these by repetition.

Figure 1.3: cone

Figure 1.3 depicts an event and the light cone in a neighborhood of the event. Event A is *future timelike*, A' is *past timelike*, B and B' are *spacelike*, C is *future lightlike*, and C' is *past lightlike* related to E .

It is instructive to contrast relativistic causal structure with the Newtonian one. Newtonian spacetime is also a 4-dimensional continuum. Each event E in Newtonian spacetime lies in a 3-dimensional subspace consisting of all the events that occur “at the same time” as E . That is, the spacetime is “layered” into spatial surfaces of absolute simultaneity. (See Figure 1.4.) In Newtonian physics, an event E can have a causal influence on any other event that occurs to the future of the simultaneity surface in which E lies, and can be influenced by events to the past. Events simultaneous with E are not causally related to E ; these are the spacelike related events.

Note that the collection of events timelike related to E is 4-dimensional in both relativistic and Newtonian spacetime. By contrast, the collection of spacelike related events is also 4-dimensional in the relativistic case, whereas it is 3-dimensional in the Newtonian case.

Curves or world lines are said to be timelike, spacelike, or lightlike, according

Figure 1.4: newton

as (infinitesimally) nearby events along them stand in those relations. Sometimes a timelike curve is referred to as an *observer*, since it is an idealized representation of a history of “here and now”’s.³ Note that a timelike curve at an event must always extend into the *interior* of the light cone at that event, whereas a spacelike curve must remain outside the light cone. This is illustrated in Figure 1.5.

Figure 1.5: curves

Just as the surfaces of simultaneity have an observer-independent status in Newtonian spacetime, so do the light cone surfaces in relativistic spacetime. However, whereas the Newtonian spacetime is layered or “foliated” by the simultaneity surfaces, the light cones of all the events of a relativistic spacetime are mutually intersecting. This is also true even if only the future cones are included. (See Figure 1.6).

One can, however, foliate a region of spacetime with future cones by selecting a particular timelike world line, and including only those cones whose vertex lies along that world line. This is depicted in Figure 1.7.

One can suggest how a nonrelativistic causal structure arises from a relativistic one by reference to Figure 1.7. If light travel times are very short compared with other timescales of interest, then it is as if the cones are opened up and

³George Gamow’s autobiography is called *My World Line*.

Figure 1.6: tangle

Figure 1.7: bondi

flattened out. The relativistic, observer-dependent foliation of spacetime illustrated in Figure 1.7 then goes over into the observer-independent Newtonian type foliation of Figure 1.4.

1.4 Metrical structure

Along any segment of a timelike worldline, a definite elapsed time exists. This is sometimes called the *proper time* along the worldline. It is the time that would be measured by an ideal clock with no spatial extension, moving along the worldline. (The nature and properties of this temporal structure are discussed a little bit more in section 2.5 below.) Actually, since the time intervals are additive, it suffices to specify the time intervals along the *infinitesimal* timelike displacements.

The time interval between two events depends on the world line along which it is defined. (The twin “paradox”.) For example, a clock on the orbiting satellite in Figure 1.1 advances *less* between events A and B than a clock on the hovering satellite.⁴ Newtonian physics this is not the case, since each event occurs at some absolute time, and the time interval between the events is just

⁴What is the timelike path with *greatest* elapsed time connecting A to B ? See problem ??.)

the difference between the corresponding absolute times.

There is also a *spatial* metric structure in spacetime, but it is determined by the structures already discussed. Spatial distances can be defined by “light cone radar” and timing measurements. For instance, the “distance” from A to B in Figure 1.8 can be defined as half the time CC' as measured by an observer for whom the times CA and AC' are equal. Although this definition depends

Figure 1.8: dist

on the observer world line, it becomes unique, to first order in displacements, as B approaches A . Infinitesimal distances are uniquely determined in this way, and from them the spatial lengths of finite spacelike curves can be built up by integration. In view of this construction, and the fundamental role played by the causal structure, a much better name for spacetime would be “timespace”.

Note that a significant rearrangement of structure has occurred in the transition from Newtonian to relativistic spacetime. In the Newtonian case, an absolute time function defines both the causal structure (surfaces of simultaneity) and the temporal structure, and the spatial metric is an independently specified piece of structure which adorns the surfaces of simultaneity. In the relativistic case, the spatial metric can be constructed from the causal and temporal structures, with no additional input. On the other hand, it takes much more information (ten functions as opposed to one) to specify the causal and temporal structures in the relativistic case.

Chapter 2

Mathematical Formulation of Spacetime Structure

In this chapter we give a precise formulation of the structures of spacetime sketched in the previous chapter. The key idea is that, because of the transitivity of the causal relation and the additivity of time intervals, and because one adopts a continuum model of spacetime, it suffices to introduce structure only in the “infinitesimal neighborhood” of each event. This neighborhood is conveniently and precisely described by the concept of the “tangent space” at each event.

2.1 The tangent space

Suppose x^μ ($\mu = 0, 1, 2, 3$) are generic coordinates for some patch of spacetime (with no particular metrical significance). An infinitesimal displacement at a given point (event) is specified by differentials dx^μ . In terms of a different set of coordinates x'^μ , the same displacement is specified by other differentials dx'^μ . The relation between the differentials is given, to first order in dx , by the chain rule:

$$dx'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} dx^\nu. \quad (2.1)$$

(Here the *Einstein summation convention* has been employed, according to which repeated indices are summed over their four numerical values.) To first order, it doesn't matter whether the partial derivatives are evaluated at the beginning or end of the displacement, since the difference between these would also be of first order, and hence would make a *second* order contribution to (2.1).) Thus, for a given infinitesimal displacement, the differentials in one

coordinate system are linearly related to those in another by the Jacobian matrix $\partial x'^\mu / \partial x^\nu$ of the transformation. It is to be understood here and below that terms of higher order in the displacements are neglected. Thus the relations are strictly valid if used in the limit where the displacement goes to zero.

The *linearity* of the relationship 2.1 means that it makes sense to talk about the *addition* of two infinitesimal displacements, independently of the choice of coordinates used to describe them. If two such displacements are labeled by $(dx^\mu)_1$ and $(dx^\mu)_2$ in one coordinate system, and by $(dx'^\mu)_1$ and $(dx'^\mu)_2$ in another, then we have

$$(dx^\mu)_1 + (dx^\mu)_2 = \frac{\partial x'^\mu}{\partial x^\nu} \left((dx'^\mu)_1 + (dx'^\mu)_2 \right).$$

That is, the differential $(dx^\mu)_1 + (dx^\mu)_2$ is related to $(dx'^\mu)_1 + (dx'^\mu)_2$ in exactly the way (2.1) required for them to label the *same* displacement. Thus the addition of infinitesimal displacements is well defined. Similarly, scalar multiplication of displacements is well defined. Thus, the infinitesimal displacements at each point of spacetime constitute a vector space. This vector space is four dimensional since, for example, dx^0 , dx^1 , dx^2 and dx^3 form a basis.

It should be emphasized that while the infinitesimal displacements at a point form a vector space, the spacetime itself is *not* a vector space. For suppose two points are labeled by $(x^\mu)_1$ and $(x^\mu)_2$ in one coordinate system, and by $(x'^\mu)_1$ and $(x'^\mu)_2$ in another. Then in general the point labeled by $(x^\mu)_1 + (x^\mu)_2$ is *different* from the point labeled by $(x'^\mu)_1 + (x'^\mu)_2$, because the transformation relating the coordinates x^μ and x'^μ is in general not linear. Unless there is a preferred set of coordinates in terms of which the addition of points can be defined, the addition of points remains meaningless. The same goes for *finite* displacements.

The concept of infinitesimal displacements at a point can be expressed without the use of infinitesimal quantities as follows. Suppose $x^\mu(\lambda)$ describes a curve, parametrized by a real number λ . Then the four numbers $dx^\mu/d\lambda|_{\lambda=0}$ transform under a coordinate change by the same linear transformation law as do the differentials (2.1),

$$\frac{dx'^\mu}{d\lambda} = \frac{\partial x'^\mu}{\partial x^\nu} \frac{dx^\nu}{d\lambda}. \quad (2.2)$$

Thus the collection of such curve derivatives also forms a 4-dimensional vector space, called the *tangent space* at $x^\mu(0)$. The members of the tangent space are called *tangent vectors*.

Multiplying the tangent vector $dx^\mu/d\lambda$ by the infinitesimal parameter incre-

ment $d\lambda$ yields an infinitesimal displacement,

$$dx^\mu = \frac{dx^\mu}{d\lambda} d\lambda. \quad (2.3)$$

In this sense, the tangent space at a point can be thought of as an infinitely magnified copy of the space of infinitesimal displacements from that point. It should be emphasized however that the tangent vectors do *not* lie “in” the manifold. Rather, they live in the tangent space, which may perhaps be usefully pictured as “hovering over” the corresponding point in the manifold.

2.2 The line element

The causal and metrical structures are both characterized by the *line element* ds^2 . The line element is a quadratic form that assigns to every infinitesimal displacement dx^μ a number,

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu. \quad (2.4)$$

Equivalently, we can think of the quadratic form as assigning to each tangent vector v^μ a number, $v^2 \equiv g_{\mu\nu} v^\mu v^\nu$. v^2 is called the squared *norm* of v^μ . (Sometimes we are sloppy and call v^2 the norm.)

Displacements with $ds^2 = 0$ (or vectors v^μ with $g_{\mu\nu} v^\mu v^\nu = 0$) are called *lightlike* or *null*. This is how the line element specifies the causal structure. The null vectors comprise the *light cone* or *null cone*, which is assumed to be a 3-dimensional double-cone that falls apart into two disconnected pieces if the origin (zero vector) is removed. The null vectors will comprise a cone of this nature provided the quadratic form is nondegenerate and has signature $+2$, as will be explained below.

The light cone separates the tangent space into the *timelike* vectors that lie in the interior of the lightcone, and the *spacelike* vectors that lie in its exterior. For a timelike displacement, $\sqrt{-ds^2}$ gives the time elapsed along that displacement, also called the “proper time”. For a spacelike one, $\sqrt{ds^2}$ gives the spatial length. This definition of spatial length agrees with that defined by the “radar timing” described above in section 1.4. (This is easily demonstrated by a computation in the tangent space. See problem 5.) In a limiting sense, along a lightlike displacement the elapsed time and the spatial length both vanish. The physical interpretation of this statement is not clear however, since a physical clock cannot travel along a lightlike worldline.

The line element is also sometimes called the *spacetime interval* or the *metric*, although the latter term more commonly refers to the array $g_{\mu\nu}$. The metric is assumed to be symmetric ($g_{\mu\nu} = g_{\nu\mu}$), since only the symmetric part would enter ds^2 anyway. It therefore amounts to 10 independent numbers at each spacetime point. In general, these numbers depend on the spacetime point, so it is really 10 functions. These functions are usually assumed to vary smoothly.

The line element has an invariant physical significance, so it must not change if the coordinates are changed. Thus, to compensate the change (2.1) of the differentials, the functions $g_{\mu\nu}$ must change as well. In particular, if the coordinates are changed to x'^μ , we have

$$g'_{\mu\nu} dx'^\mu dx'^\nu = g_{\mu\nu} dx^\mu dx^\nu = g_{\mu\nu} \frac{\partial x^\mu}{\partial x'^\alpha} \frac{\partial x^\nu}{\partial x'^\beta} dx'^\alpha dx'^\beta. \quad (2.5)$$

Since (2.5) holds for all displacements dx'^α one evidently (Problem 6) must have¹

$$g'_{\alpha\beta} = \frac{\partial x^\mu}{\partial x'^\alpha} \frac{\partial x^\nu}{\partial x'^\beta} g_{\mu\nu}. \quad (2.6)$$

Thus, under a change of coordinates, the metric components transform linearly via contraction of each index with the inverse of the Jacobian of the transformation $x^\mu \rightarrow x'^\mu$. The linear relation (2.6) between $g'_{\alpha\beta}$ and $g_{\mu\nu}$ is an example of a *tensor transformation law*, generalizing the vector transformation law (2.2), and the metric $g_{\mu\nu}$ is an example of a *tensor*.

As mentioned above, a spacetime line element must have the property that $ds^2 = 0$ determines a cone. It turns out that this condition on the metric tensor $g_{\mu\nu}$ is equivalent to the requirement that, in the neighborhood of each spacetime point, one can find coordinates (t, x, y, z) such that *at that point* (but not in general anywhere else) the line element takes the form

$$ds^2 = -dt^2 + dx^2 + dy^2 + dz^2, \quad (2.7)$$

i.e., the metric components take the values $\eta_{\mu\nu} := \text{diag}(-1, +1, +1, +1)$. A metric having this property is called a *Lorentzian metric*, and $\eta_{\mu\nu}$ is called the *Minkowski metric*. We use here “geometrical units,” in which the speed of light is equal to unity, and time and length are measured in the same units.

The Lorentzian condition on the metric is also equivalent to the requirement that in each tangent space, the quadratic form defined by $g(v, v) \equiv g_{\mu\nu} v^\mu v^\nu$ be

¹In order to obtain (2.6), the dummy indices μ, ν are traded for α, β in the leftmost member of (2.5). Such index substitutions are a common procedure in computations with tensors.

nondegenerate and have *signature* equal to 2. That is, for any “orthonormal” basis $\{e_i\}$ satisfying $g(e_i, e_j) = \pm\delta_{ij}$, the number of positive norm vectors minus the number of negative norm ones is 2. (That is, there is exactly one negative norm vector in an orthonormal basis.) That there always exists an orthonormal basis, and that the signature is independent of the choice of this basis, is proved in problem 3.

A useful diagnostic for testing whether a metric has signature 2 in four dimensions is to compute its determinant. Although the determinant is not basis independent, its *sign* is basis independent. (Problem 4.) The sign of the determinant is positive for signatures 0 and ± 4 , and negative for signatures ± 2 . To distinguish signature 2 from -2 one can further check, for example, whether there exist at least two orthogonal positive norm vectors.

Important properties of a Lorentzian metric in the tangent space are developed in the problems at the end of this chapter.

2.3 Local inertial coordinates and curvature

It may seem that by coordinate transformations one can make the metric have any form at all, however this is clearly not the case. There are 10 independent functions in the metric, but only 4 free functions in coordinate transformations. Thus, while there is indeed a lot of freedom to alter the components of the metric tensor by coordinate transformations, there are 6 functions of invariant information coded into the metric tensor.

As stated above in section 2.2, one can always find coordinates around any given point x_0 such that a Lorentzian metric takes the Minkowski form (2.7) at x_0 , $g_{\mu\nu}(x_0) = \eta_{\mu\nu}$. If one begins with $g_{\mu\nu}$ given in arbitrary coordinates x^μ , the Minkowski condition $g'_{\mu\nu} = \eta_{\nu\mu}$ can be viewed as 10 equations on the 16 partial derivatives in the Jacobian matrix $\partial x^\mu / \partial x'^\alpha$ appearing in the transformation law (2.6). As long as the signature of $g'_{\mu\nu}$ is $+2$, the Minkowski form can be achieved at x_0 by a coordinate transformation with 6 degrees of freedom in the first partials to spare (and total freedom in the higher partials).

One thus expects that there is a 6 parameter family of Jacobian matrices that leave any given $g_{\mu\nu}$ invariant under the tensor transformation law (2.6). This 6 parameter family of linear transformations in the tangent space that preserve the metric is called the *local Lorentz group*. In terms of local coordinates in which the metric takes the Minkowski form (2.7), the Lorentz group consists of combinations of *rotations* amongst the spatial coordinates (x, y, z) , and *boosts*, which mix t with the other coordinates and correspond to transformations to a

relatively moving reference frame.

Let us now systematically try to find coordinates in the neighborhood of x_0 that give the line element the Minkowski form. This will show us where the obstruction occurs. Suppose that in the coordinate system x^μ we have $g_{\mu\nu}(x_0) = \eta_{\mu\nu}$. Any coordinate transformation $x \rightarrow x'$ whose Jacobian is a Lorentz transformation at x_0 will preserve this Minkowski form. Now, using some of the remaining freedom in the choice of coordinates, one can always arrange for all of the first partial derivatives of $g_{\mu\nu}$ to vanish at x_0 . Indeed, the condition $g'_{\mu\nu,\gamma} = 0$ is $10 \times 4 = 40$ equations, where the comma notation “ $,\gamma$ ” denotes partial differentiation with respect to x'^γ . Using the transformation law (2.6) this condition becomes 40 equations on the quantities $\partial^2 x^\mu / \partial x'^\gamma \partial x'^\alpha$, which are also effectively 40 in number due to the commutivity of partial derivatives. These 40 equations are linear and can always be solved uniquely for the second partials. Thus, the vanishing of $g'_{\mu\nu,\gamma}$ completely fixes the second partials of the coordinate transformation, once the first partials are fixed.

A coordinate system in which $g_{\mu\nu}(x_0) = \eta_{\mu\nu}$ and $g_{\mu\nu,\gamma}(x_0) = 0$ is called a *local Minkowski coordinate system* at x_0 , or a *local inertial coordinate system*. The existence of such coordinates expresses the fact that any Lorentzian manifold looks, “up to one derivative of the metric”, like Minkowski space in the neighborhood of each point.

Can we further specify the coordinates so that the second partial derivatives $g_{\mu\nu,\alpha\beta}$ also vanish at x_0 ? These partials amount to $10 \times 10 = 100$ independent quantities. On the other hand, the new functions that will appear in the transformation law for the second partials are $\partial^3 x^\mu / \partial x'^\alpha \partial x'^\beta \partial x'^\gamma$, and these comprise only $4 \times 20 = 80$ independent numbers. (The number of independent components of a totally symmetric, three index object $T_{\alpha\beta\gamma}$ is $n(n+1)(n+2)/3!$ if each index ranges over n values. See Problem 7.) This means that in general one *cannot* arrange for all the second partials $g_{\mu\nu,\alpha\beta}$ to vanish at a given point. In the generic case, at least 20 of these second partials must remain non-vanishing.

Suppose that around every point, a coordinate system (which depends on the point) exists in which *all* of the first and second partials of $g_{\mu\nu}$ vanish at that point. Then, although it is not at all obvious, it turns out that there exists a *single* coordinate system in which $g_{\mu\nu} = \eta_{\mu\nu}$ *everywhere*. More precisely, this is true at least in a coordinate patch of finite size. In this case the spacetime is said to be *flat*. If a single such coordinate patch covers all of spacetime, this is the spacetime of *special relativity*, called *Minkowski space*.

If, on the other hand, there are points at which the first and second partials of $g_{\mu\nu}$ *cannot* be simultaneously set to zero by a choice of coordinates, then the

metric is said to be *curved*. The fundamental idea of general relativity is that curvature of the metric corresponds to gravitational tidal forces.

As a simple example, consider the line element

$$ds^2 = -dt^2 + a^2(t)(dx^2 + dy^2 + dz^2), \quad (2.8)$$

which describes a spatially flat and homogeneous cosmology, in which the flat spatial metric has a time-dependent scale factor $a^2(t)$. Since $a(t)$ is only one function, and there are four free functions worth of coordinate transformations available, one might suspect that the line element (2.8) is a flat line element in disguise. However, this is not the case. If $a(t)$ depends on t , then in fact one can *not* find a change of coordinates that will put (2.8) into the Minkowski form (2.7) everywhere. This is the case for our universe which, to some approximation, can be described at large scales by a line element of the form (2.8), with a increasing approximately as $t^{2/3}$, where t is the time measured by a clock at rest with respect to the microwave background radiation.

Another example, involving the *Riemannian* signature metric on the unit sphere, is given in Problem 10.

2.4 Relation between causal and metrical structures

It is profoundly beautiful how the line element combines the causal and metrical structures into one. By contrast, in Newtonian physics, the causal and temporal metric structures are specified by the absolute time function t or, equivalently, by the differential dt . But this leaves completely unspecified the *spatial* metric. Thus, in addition to dt , Newton must specify a spatial metric $dl^2 = h_{ij}dx^i dx^j$, where x^i , ($i = 1, 2, 3$) are spatial coordinates.

It may seem that more than one function on spacetime is required to specify the temporal metric in relativity, since the ten independent components of the metric are needed to assign a time interval to all possible timelike displacements. In fact, once the causal structure has been specified, only one additional function is required to pin down the metric. To see why this is true, first note that two metrics $g_{\mu\nu}$ and $\Omega^2 g_{\mu\nu}$ related by an overall positive factor Ω^2 define the same light cones. (The two metrics are said to be *conformally related* by the *conformal factor* Ω^2 .) Conversely, if two metrics define the same light cones they are necessarily conformally related. To verify this, it suffices to analyze the situation in the tangent space at a point as follows.

Let us adopt the dot product notation $v \cdot w := g_{\mu\nu}v^\mu w^\nu$. The dot product $v \cdot w$ is also called the *inner product* of v with w , and $v \cdot v$ is called the *squared norm* or sometimes (sloppily) just the *norm* of v . All inner products $v \cdot w$ can be expressed in terms of norms via $v \cdot w = \frac{1}{2}[(v+w) \cdot (v+w) - v \cdot v - w \cdot w]$, so it suffices to determine all norms. Suppose we are given the light cone, *i.e.*, all the vectors n for which $n \cdot n = 0$. Fixing any timelike vector t , we will determine all other norms in terms of $t \cdot t$.

Let v be any vector. The plane formed by v and t is spanned by a basis of two null vectors l and n that add up to t , $t = l + n$. (See Figure 2.1.) Thus we

Figure 2.1: vtplane

can always express v as $v = \alpha l + \beta n$ for some numbers α and β . Then we have

$$v \cdot v = (\alpha l + \beta n) \cdot (\alpha l + \beta n) \quad (2.9)$$

$$= 2\alpha\beta l \cdot n \quad (2.10)$$

$$= \alpha\beta (l + n) \cdot (l + n) \quad (2.11)$$

$$= \alpha\beta t \cdot t, \quad (2.12)$$

where linearity of the inner product and $l \cdot l = 0 = n \cdot n$ have been used.

We have shown that all inner products are determined by the light cone plus the norm of one timelike vector, $t \cdot t$. This number just determines the overall scale of the metric. That is, we have the equation

$$\text{metric} = \text{causal cone} + \text{scale}.$$

Thus, although we call $g_{\mu\nu}$ the “metric”, it is in large part ($\frac{9}{10}$) just the causal structure!

2.5 Deep background

Our attribution of differential, causal and metric properties to spacetime is based ultimately on the possibility of making certain kinds of measurements. The phys-

ical processes by which these structures are determined have some fundamental limits of resolution imposed by their quantum nature, if nothing else.

It is an idealization when we extrapolate these notions to infinitesimal regions of spacetime. Even the assumption that spacetime is a continuum with a differentiable structure is an idealization whose validity is surely limited. A c -number coordinate defined by a physical process is only meaningful in some classical approximation. The true coordinates, one would think, must be q -numbers if they are fundamentally meaningful at all. Nevertheless, it is this classical, continuum idealization that we make in setting up the foundation for doing physics.

Having accepted the idealization, it is still interesting to attempt to characterize its assumptions in as fundamental a manner as is possible. One such attempt appears in a classic paper by Ehlers, Pirani and Schild² (EPS), which develops a system of axioms for spacetime structure in terms of topological and differential axioms about the properties of freely falling massive and massless point particles.

One deep question is why the causal cone is given by a quadric in the tangent space. After all, one can easily imagine a partial ordering relation that arises from an infinitesimal conical structure which is *not* a quadric. In the EPS paper, the quadratic nature of the light cone is *derived* from their axioms. This is not very satisfying however, since one of the axioms is not particularly physically natural.³

Aside from any axioms, there is a special property of quadrics that might underlie the fact that the causal structure is given by one. Namely, quadrics have the largest possible symmetry group of any conical subset of the tangent space.⁴ This is the Lorentz group, together with the conformal rescalings, a group with 7 continuous parameters.

From time to time people try to generalize the notion of the spacetime metric to allow for non-quadratic line elements. These go under the rubric “Finsler metrics”.⁵ It seems that in order to generalize known physical theories

²J. Ehlers, F.A.E. Pirani, and A. Schild, “The Geometry of Free Fall and Light Propagation,” in *General Relativity; Papers in Honor of J.L. Synge*, Oxford, Clarendon Press, 1972.

³The axiom in question can be described with reference to Fig. 1.8. Fixing the timelike curve through A and an arbitrary smooth parameter λ along the curve, the axiom states that the function $f(B) = \lambda(C)\lambda(C')$ is a twice differentiable function on spacetime. (See Axiom L_1 of the EPS paper.)

⁴I should have a reference for this but I don’t know of one. Perhaps Herman Weyl proved it. Perhaps it is not even true (see Problem 11).

⁵See, e.g., *Finsler geometry, relativity and gauge theories*, G.S. Asanov (Reidel, 1985).

to non-quadratic Finsler metrics one must introduce further structure (e.g. a spacetime volume element) and the result is not nearly as simple or “natural” as it is when a quadratic metric is the sole structure. Nevertheless, of course, it might be that the simple, quadratic metric is only an approximation that can be improved by Finslerian corrections.

Another deep question is what is the origin of the differential structure of spacetime? As remarked in section 1.1, as a point set, the same set of events could be given many different differential structures, even of different dimension. So the differential structure is real physical input in the theory. Where does it come from? In the EPS paper it is put in in the axioms, in a way that refers to the behavior of particle world lines and light rays.

Further insight into the origin of differential structure is provided by a rather remarkable fact: It turns out that not only does the causal structure determine the metrical structure up to a function, but it determines the differential structure of spacetime as well! Stated more precisely, it has been shown⁶ that if two manifolds with Lorentzian metrics (M, g) and (M', g') are causally isomorphic as causal sets, then they are necessarily diffeomorphic as manifolds (via a diffeomorphism that is a conformal isometry.) There is thus a stunning economy of structure in relativity, since the causal structure determines all spacetime structure except the conformal factor.

In view of the above observations, it would seem extremely natural to build up the theory of spacetime structure beginning not with a differentiable manifold, but with just a set of events, together with a causal partial ordering relation. There is a catch however. An arbitrary partial order on a set of events will not in general be the causal order induced by a Lorentzian metric on a manifold, even if the set of events is uncountably infinite. Nevertheless, the study of discrete (“locally finite”) *causal sets* as a possible foundation for a quantum theory of spacetime and gravity is being actively pursued.⁷

The existence of an intrinsic time interval associated to any timelike displacement is another deep mystery. The fact is that, in Nature, there are systems that can serve as *clocks*. It seems to be the case that fundamental systems all march

⁶It follows from a pair of theorems proved in S.W. Hawking, A.R. King, and P.J. McCarthy, “A new topology for curved space-time which incorporates the causal, differential, and conformal structures,” *J. Math Phys.* **17**, 174 (1976), and D. Malament, “The class of continuous curves determines the topology of spacetime”, *J. Math Phys.* **18**, 1399 (1977).

⁷See, for example, G. 't Hooft, “Quantum Gravity: A Fundamental Problem and Some Radical Ideas,” in *Recent Developments in Gravitation, Cargèse 1978*, edited by M. Levy and S. Deser (Plenum, New York, 1979); L. Bombelli et al, “Spacetime as a Causal Set,” *Phys. Lett.* **59**, 521 (1987); Comment and reply, **60**, 655-56; R.D. Sorkin, “Spacetime and Causal Sets,” in Proceedings of SILARG VII Conference, Mexico City, 1990.

to the beat of the same drummer, in the following sense: there is a large class of physical systems that mark time in a commensurate fashion. For instance, an atomic clock, a lump of decaying Carbon-14, and a rapidly spinning neutron star all “sitting next to each other” will indicate the same time interval between two given events along their common world line (once perturbing effects and the finite extent of the clocks are taken into account). It is truly remarkable that such a large collection of commensurate clocks exists in nature, and also that there seems to be only one such mutually commensurate collection. Perhaps the existence of a unique set of commensurate clocks should be traced to the existence of a common volume element in spacetime, as described below.

We have remarked already that the metric can be determined by the causal structure together with the norm of any one timelike vector. Instead of selecting a timelike vector however, a more symmetrical piece of information that can serve just as well to set the scale of the metric is the *spacetime volume element*. Mathematically, this is given by $\sqrt{-\det g} dx^0 dx^1 dx^2 dx^3$ in a given coordinate system. Since it seems somewhat less direct to measure the spacetime volume of a region than it does to read a clock, one may be disinclined to think of the volume element as fundamental. On the other hand, the volume element plays a crucial role in writing down the *action functionals* which, it may be said, are the cornerstone of contemporary physical theory. In the context of the discrete causal sets mentioned above⁷, a discrete notion of volume is defined simply by *counting* the finite number of events in a given region. Thus the extra piece of information needed to go from causal structure to metric is inherently present in a discrete causal set. In this sense, *all* of the elements of spacetime structure are embodied in the notion of a discrete causal set.

2.6 Problems

1. Show graphically and confirm algebraically that
 - (a) the sum of two *future pointing* timelike or non-parallel null vectors is a future pointing, timelike vector.
 - (b) the sum of two timelike or null vectors can also be spacelike or null;
 - (c) the sum of two null vectors can be null only if they are parallel;
 - (d) the sum of two spacelike vectors can be timelike, spacelike, or null;

(If you wish you may choose coordinates so that any given timelike, null, or spacelike vector is in one of the standard forms $(a,0,0,0)$, $(a,a,0,0)$,

or $(0, a, 0, 0)$ respectively, for some number a . These forms can always be achieved by a Lorentz transformation. This entails no loss of generality, since the properties in question are Lorentz-invariant.)

2. Show that

- (a) the sum of any two orthogonal spacelike vectors is spacelike;
- (b) a timelike vector and a null vector cannot be orthogonal;
- (c) a spacelike vector and a null vector *can* be orthogonal;
- (d) two null vectors cannot be orthogonal, unless they are parallel.

(As in the previous problem, you may assume a standard form for any one vector.)

3. Prove that the signature of the metric is a true invariant, i.e. it is independent of coordinates (or basis in the tangent space). This is a problem in linear algebra. One way to solve it is to generalize the problem somewhat as follows. Let V be an n -dimensional vector space, and let g be a quadratic form on V , i.e., a symmetric, bilinear map from $V \times V$ to the real numbers.

- (a) Show that one can always find an *orthonormal basis* e_1, \dots, e_n of V , i.e. a basis such that $g(e_i, e_j) = \pm \delta_{ij}$. (Hint: Use induction.)
- (b) The *signature* of g is defined as the number of positive norm basis vectors in an orthonormal basis minus the number of negative norm ones. Show that the signature is independent of the choice of orthonormal basis.

4. Show that while the *magnitude* of the determinant of the metric depends on the coordinate system, the *sign* does not.

5. Show that $\sqrt{ds^2}$ for infinitesimal spacelike intervals is the same as the distance defined by “radar timing” in section 1.4. This seems to be most clearly formulated as a property of the metric in the tangent space: if a spacelike vector s is related to a timelike vector t and two null vectors n and n' by $n = t + s$ and $n' = t - s$ as shown in Figure 2.2, then the squared length $s^2 = g(s, s)$ assigned to s by the metric is equal to the “radar-distance” $-t^2$.

6. Show that if $T_{\alpha\beta} V^\alpha V^\beta = 0$ for all V^α then the symmetric part $T_{(\alpha\beta)} \equiv (T_{\alpha\beta} + T_{\beta\alpha})/2$ of $T_{\alpha\beta}$ must vanish.

Figure 2.2: radar

7. Show that number of independent components of a totally symmetric k index object $T_{\alpha_1 \dots \alpha_k}$ is $n(n+1) \dots (n+k-1)/k!$ if each index ranges over n values. (This implies in particular that the number of independent components of $\partial^3 x'^{\mu} / \partial x^{\alpha} \partial x^{\beta} \partial x^{\gamma}$ is 4×20 in four spacetime dimensions.)
8. Show that the cosmological line element (2.8) gives a flat spacetime if and only if a is a constant. (Note: The two dimensional submanifold at fixed y and z is also flat if a is proportional to t .)
9. The two-dimensional line element $ds^2 = -dt^2 + t^2 dx^2$ is actually flat. Show this by finding a coordinate transformation to new coordinates τ and σ in terms of which one has $ds^2 = -d\tau^2 + d\sigma^2$. Draw lines of constant t and x on a rectangular τ - σ spacetime diagram. What region of the τ - σ Minkowski space is covered by the t - x coordinate patch?
10. In standard spherical coordinates (θ, ϕ) on the unit sphere, the line element takes the form $ds^2 = d\theta^2 + \sin^2 \theta d\phi^2$.
 - (a) Show that (θ, ϕ) provide a local Euclidean coordinate system ($g_{ij} = \delta_{ij}$ and $g_{ij,k} = 0$) at every point on the equator ($\theta = \pi/2$), but not anywhere else on the sphere.
 - (b) Show that, even at the equator, the second partial derivatives of the metric components in (θ, ϕ) coordinates do not all vanish.
 - (c) Argue that no change of coordinates can transform the line element into the Pythagorean line element $ds^2 = dx^2 + dy^2$ in any finite patch of the sphere.
11. Prove or disprove the statement that any non-quadric cone would have a smaller symmetry group in the tangent space than the 7 parameter (Lorentz transformations plus scalings) of a quadratic cone.

Chapter 3

Free-fall and Geodesics

According to general relativity, if a particle is not acted upon by any (non-gravitational) forces, it is said to be in *free-fall*, or *inertial motion*. Different gravitational fields are described by different spacetime metrics, and the possible inertial motions of a particle are determined by the metric. If true gravitational effects are present, the metric has curvature, which manifests itself via the relative accelerations of freely falling objects. Such gravitationally induced relative acceleration is said to be caused by “tidal forces”, even though, from the perspective of general relativity, there are no “forces” acting. In this chapter we will characterize the inertial motions of idealized test particles that follow timelike curves and idealized light rays that follow lightlike curves.

3.1 Curves

A curve in spacetime is a smooth function $x^\mu(\lambda)$. Smoothness implies that the curve has a well-defined tangent vector $dx^\mu/d\lambda$ at each point along the curve. A curve is timelike, spacelike, or lightlike according as its tangent vector is everywhere timelike, spacelike, or lightlike respectively. A lightlike curve is also called a *null* curve.

The elapsed proper time along a timelike curve is given by $\int \sqrt{-ds^2}$, which can be expressed as $\int (-g_{\mu\nu} \dot{x}^\mu \dot{x}^\nu)^{1/2} d\lambda$, where \dot{x}^μ denotes the tangent vector $dx^\mu/d\lambda$ to the curve. A similar expression without the $-$ sign gives the proper length of a spacelike curve. Note that, as required, these length integrals are independent of the parametrization of the curve (Problem 1). A lightlike curve has vanishing proper time/length along it.

If a timelike curve is parametrized by proper time, the norm of the tangent

vector is everywhere equal to -1 :

$$g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu = (g_{\mu\nu}dx^\mu dx^\nu)/d\tau^2 = -d\tau^2/d\tau^2 = -1. \quad (3.1)$$

(Similarly, a spacelike curve parametrized by its own length has a tangent vector of unit norm.) Thus it is often convenient to parametrize a timelike curve by the proper time along it.

3.2 Inertial motion

The inertial or free-fall world lines $x^\mu(\tau)$ parametrized by proper time τ can be characterized by the following property:

An inertial world line is one for which the coordinate acceleration $d^2x^\mu/d\tau^2$ at each point p vanishes when evaluated in a local inertial coordinate system at p .

The metric determines these free-fall motions, since it is the metric that selects out the local inertial coordinates from among all possible coordinates.

It is important to understand that this condition of vanishing acceleration does not depend on which local inertial coordinate system is used at p . To see why, note that under a coordinate change $x^\mu \rightarrow x'^\mu$, the velocity transforms as in eqn. (2.2), hence the acceleration transforms as follows:

$$\frac{d^2x'^\mu}{d\tau^2} = \frac{\partial x'^\mu}{\partial x^\alpha} \frac{d^2x^\alpha}{d\tau^2} + \frac{\partial^2 x'^\mu}{\partial x^\alpha \partial x^\beta} \frac{dx^\alpha}{d\tau} \frac{dx^\beta}{d\tau} \quad (3.2)$$

A transformation from one inertial coordinate system at p to another must have vanishing second partial derivatives at p in order to preserve the condition that $g_{\mu\nu,\rho}(p) = 0$. (See section 2.3.) Thus the coordinate acceleration transforms linearly as a 4-vector at p under a change from one inertial coordinate system to another. In particular, the condition that it vanish at p is independent of the choice of local inertial coordinates at p . On the other hand, in an arbitrary coordinate system, the coordinate acceleration will certainly not vanish.

The characterization of inertial world lines given above is almost totally impractical since, in a general curved spacetime, it necessarily refers to a different local inertial coordinate system at every point. It would be much better to be able to identify inertial motion in an arbitrary coordinate system. One way to find such a characterization is to begin with a coordinate invariant description as follows.

Recall the twin effect of special relativity. The proper time between two events is maximized by the inertial (i.e., straight in inertial coordinates) world line that connects them. Since an inertial world line in a general curved space-time looks (to second order) like a straight line in a local inertial coordinate system in the neighborhood of each point, it should maximize the proper time between infinitesimally separated points along it. Therefore the total proper time connecting the fixed endpoints of the curve should be stationary under a large class of infinitesimal variations of the curve. In fact, as will now be shown, the total proper time is stationary under *all* infinitesimal variations.¹ Furthermore, since this stationarity condition is manifestly independent of coordinates, it will yield a characterization of inertial motion that is applicable in any coordinate system.

The proper time along a world line can be written as

$$S = \int \sqrt{-L} d\lambda, \quad (3.3)$$

with L is defined by

$$L \equiv g_{\mu\nu} \dot{x}^\mu \dot{x}^\nu, \quad (3.4)$$

where the notation is as above. The condition that $x^\mu(\lambda)$ be a stationary point of the proper time (3.3) yields the Euler-Lagrange equations for the Lagrangian $\sqrt{-L}$:

$$\frac{d}{d\lambda} \frac{\partial \sqrt{-L}}{\partial \dot{x}^\alpha} - \frac{\partial \sqrt{-L}}{\partial x^\alpha} = 0. \quad (3.5)$$

As long as $L \neq 0$, (3.5) is equivalent to

$$\left(\frac{d}{d\lambda} + \frac{1}{2} L^{-1} \dot{L} \right) \frac{\partial L}{\partial \dot{x}^\alpha} - \frac{\partial L}{\partial x^\alpha} = 0. \quad (3.6)$$

In order to simplify (3.6) let us now specify that the originally arbitrary parameter λ is in fact τ , the proper time along the curve. Then, since with that parameter $L = -1$ (cf. (3.1)), we have $\dot{L} = 0$, and the Euler-Lagrange equation for the Lagrangian $\sqrt{-L}$ becomes identical to that for L ,

$$\frac{d}{d\lambda} \frac{\partial L}{\partial \dot{x}^\alpha} - \frac{\partial L}{\partial x^\alpha} = 0. \quad (3.7)$$

¹Although the proper time is stationary, it is in general *not* a local maximum if the endpoints are sufficiently separated. By analogy, on a sphere, a segment of a great circle going more than halfway around the sphere is a geodesic, but it is not the shortest curve between its endpoints.

Using the definition (3.4) of L , (3.7) becomes

$$\frac{d}{d\lambda}(g_{\alpha\nu}\dot{x}^\nu) - \frac{1}{2}g_{\mu\nu,\alpha}\dot{x}^\mu\dot{x}^\nu = 0. \quad (3.8)$$

The stationarity condition $\delta S = 0$ leading to (3.8) is coordinate independent, so (3.8) must hold in *any* coordinate system. That is, if it holds in one coordinate system, it will necessarily hold in any other coordinate system. (This can also be verified directly by expressing (3.8) in new coordinates using the transformation rules for \dot{x}^μ (2.2) and $g_{\mu\nu}$ (2.6). See Problem 2.) If we choose a coordinate system that is locally inertial at $x^\mu(\tau_0)$ then, at $\tau = \tau_0$, (3.8) becomes simply $\ddot{x}^\mu(\tau_0) = 0$. Thus (3.8) is in fact *equivalent* to the statement that $x^\mu(\tau)$ is an inertial world line as defined at the beginning of this section. The important thing is that (3.8) holds in any coordinate system, so that local inertial coordinates need not be invoked in order to characterize the inertial motion.

Note that equation (3.8) is a set of four coupled ordinary second order differential equations on the four functions $x^\mu(\tau)$. Thus the initial spacetime position and 4-velocity of an inertial test particle uniquely determine its subsequent motion.

3.3 Lightlike free-fall and geodesics

Lightlike inertial motion cannot be characterized with reference to proper time parametrization since the proper time along a lightlike curve vanishes. However this does not prevent us from characterizing such motion in essentially as simple a manner as in the timelike case. To this end, it is useful to generalize the language slightly and introduce the concept of a *geodesic*. In all generality, a geodesic is a curve $x^\mu(\lambda)$ with the property that the coordinate acceleration $d^2x^\mu/d\lambda^2$ at any point p is parallel to the velocity $dx^\mu/d\lambda$ at p when expressed in a coordinate system that is locally inertial at p . (It is assumed in this definition that the parametrization is non-singular, in the sense that $dx^\mu/d\lambda$ is everywhere non-zero.) This definition of geodesics is independent of which locally inertial coordinate system is used at p , for the same reason as explained above in the timelike case. It is also independent of the parametrization of the curve, as can easily be seen directly by examining the effect of reparametrizing a curve.

Under a change of parameter $\lambda \rightarrow \sigma$, the velocity and acceleration become

$$\frac{dx^\mu}{d\sigma} = \frac{d\lambda}{d\sigma} \frac{dx^\mu}{d\lambda} \quad (3.9)$$

$$\frac{d^2x^\mu}{d\sigma^2} = \left(\frac{d\lambda}{d\sigma}\right)^2 \frac{d^2x^\mu}{d\lambda^2} + \left(\frac{d^2\lambda}{d\sigma^2}\right) \frac{dx^\mu}{d\lambda}. \quad (3.10)$$

If $d^2x^\mu/d\lambda^2$ and $dx^\mu/d\lambda$ are parallel at a point, then $d^2x^\mu/d\sigma^2$ and $dx^\mu/d\sigma$ are evidently also parallel at that point, so the property of being a geodesic is independent of the parametrization.

One can always reparametrize a geodesic so that the acceleration *vanishes*, rather than just being parallel to the velocity. A geodesic parameter for which the acceleration vanishes is called an *affine parameter*. If λ and σ are both affine parameters, then the $d^2\lambda/d\sigma^2$ term in (3.10) must vanish, so they must be linearly related as $\lambda = a\sigma + b$ for some constants a and b . That is, the affine parameter along a geodesic is determined up to an overall scale and an additive constant. For timelike geodesics the proper time is an affine parameter, as is the proper length for spacelike geodesics.

Alternatively, a geodesic can be defined as a curve satisfying the “geodesic equation” (3.8), with the “overdot” indicating derivative with respect to the parameter of the curve (not necessarily the proper time). In a local inertial coordinate system at p , the geodesic equation reduces to the statement that the coordinate acceleration vanishes at p . Thus eqn. (3.8) should more explicitly be called the geodesic equation for affinely parametrized geodesics. Under an arbitrary reparametrization (3.8) will no longer hold and the coordinate acceleration will no longer vanish. Nevertheless the acceleration will necessarily remain parallel to the velocity, as was shown above.

Note that the timelike, null, or spacelike character of a geodesic is necessarily preserved along the curve. To see this, evaluate the scalar $\frac{d}{d\lambda}(g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu)$ along an affinely parametrized geodesic. In a local inertial coordinate system at a point p the derivative of $g_{\mu\nu}$ will vanish and, since the parameter is affine, the derivative of \dot{x}^μ will vanish. Thus the whole expression vanishes. Since it is a scalar, it will vanish in any coordinate system, so the squared norm of the tangent vector must be constant along an affinely parametrized geodesic.

The concept of affine parameter for a lightlike geodesic is somewhat elusive, for a couple of reasons. For one thing, a general lightlike curve that is not a geodesic has no preferred parametrization. Affine parametrization is meaningful in the lightlike case *only* for curves that are geodesics, whereas any timelike (or spacelike) curve has a proper time (or length) parametrization which is affine if the curve happens to be a geodesic. For another thing, all overall scalings for the affine parameter of a lightlike geodesic are on an equal footing, whereas in the timelike (or spacelike) cases the proper time (or length) serves as a naturally preferred affine parameter. It is perhaps helpful to note that one can think of the affine parameter along a lightlike geodesic as measuring the fraction of proper time along an infinitesimally nearby timelike geodesic. Even though the proper time is going to zero, one can fix initial and final points and then this fraction is

finite. The arbitrariness of the scale of the affine parameter then corresponds to the arbitrariness of the choice of initial and final points used in this construction.

3.4 Conserved quantities along a geodesic

If the metric is independent the coordinate $x^{\hat{\alpha}}$ in some coordinate system $\{x^\mu\}$, then the geodesic equation (3.8) immediately yields a conservation law,

$$\frac{d}{d\lambda}(g_{\hat{\alpha}\nu}\dot{x}^\nu) = 0, \quad (3.11)$$

where λ is any affine parameter for the geodesic. This is just a special case of the familiar fact that if a Lagrangian L is independent of a particular coordinate $x^{\hat{\alpha}}$, then the Euler-Lagrange equations (3.7) imply that the conjugate momentum, $\pi_{\hat{\alpha}} := \partial L / \partial \dot{x}^{\hat{\alpha}}$, is a conserved quantity.

Associated with the symmetry of the spacetime under translations of $x^{\hat{\alpha}}$ (while holding fixed the remaining coordinates $\{x^\mu\}$) there is a vector field ξ^μ , called a *Killing vector* for the metric. ξ^μ is defined by specifying that, in the coordinate system $\{x^\mu\}$, all components of ξ^μ vanish except for the $\hat{\alpha}$ -component which is unity. That is, in the coordinate system $\{x^\mu\}$,

$$\xi^\mu := \delta_{\hat{\alpha}}^\mu. \quad (3.12)$$

In terms of ξ^μ , the corresponding conserved quantity can be written as

$$g_{\mu\nu}\xi^\mu\dot{x}^\nu. \quad (3.13)$$

That is, the conserved quantity is the inner product of the Killing vector with the geodesic tangent vector.

3.5 Field theory in curved spacetime

In this chapter we have seen how a “gravitational field” affects the motion of test particles and light rays. This will suffice for most of the elementary considerations that are encountered in the initial study of general relativity. However, if one wishes to describe the propagation of fields, such as the electromagnetic field or even quantum fields, in a curved spacetime, it becomes necessary to formulate the relevant field equations in a generic, curved spacetime. The guiding principle here is the same as that which motivated the notion of a geodesic: in an infinitesimal neighborhood of each event the field equations should agree

with those in flat Minkowski space. More precisely, when examined in local inertial coordinates at a point, the field equations should agree. In fact this is perhaps too strong a requirement, since it is a matter of observation to determine whether this correspondence is precise, or only approximate. It is possible that there are local curvature “corrections” to the field equations that are not detectable in flat, or nearly flat, spacetime.

3.6 Problems

1. Show that the proper time integral $\int (-g_{\mu\nu} \dot{x}^\mu \dot{x}^\nu)^{1/2} d\lambda$ is independent of the parametrization of the curve.
2. Since the stationarity condition $\delta S = 0$ leading to (3.8) is coordinate independent, (3.8) will hold in any coordinate system if it holds in one coordinate system. Verify this directly by expressing (3.8) in new coordinates using the transformation rules for \dot{x}^μ (2.2) and $g_{\mu\nu}$ (2.6).
3. Use the variational principle $\delta \int g_{ij} \dot{x}^i \dot{x}^j = 0$ to find the equation satisfied by (affinely parametrized) geodesics on the unit 2-sphere. Show that the solutions to this equation are precisely the great circles. (You may use spherical symmetry to simplify your task.) Using the fact that spherical coordinates are locally Euclidean on the equator (Problem 10), give an independent argument showing that the equator is a geodesic.
4. In Chapter 1 an example was mentioned involving two satellites, one orbiting the earth (in free-fall) at fixed radius, and the other hovering without orbiting (along an accelerated world line) at the same fixed radius and constant angular position (Fig. 1.1). Suppose both satellites are in the equatorial plane $\theta = \pi/2$, with the hovering one fixed at $\phi = 0$. Using the line element for a static, spherically symmetric empty spacetime, show that the proper time between two successive encounters (events A and B in the figure) is longer along the world line of the hovering satellite. Is the proper time a local maximum along the orbiting world line? Along the hovering world line? Describe the world line along which the proper time is an absolute maximum between events A and B .
5. Show that conformally related metrics $g_{\mu\nu}$ and $\Omega^2 g_{\mu\nu}$ (with $\Omega(x)$ any nowhere vanishing function) determine the same *null* geodesics, but with a different definition of affine parametrization. Show that the timelike and spacelike geodesics are *not* the same for the two metrics.

6. A spacetime with line element $ds^2 = \Omega^2(-dt^2 + \sum_i dx^i dx^i)$, $i = 1, 2, 3$ is called *conformally flat*. (Ω is any nowhere vanishing function.) Using the coordinates in which the metric takes the above form,
- Find the geodesic equation for affinely parametrized geodesics in a conformally flat spacetime.
 - For timelike geodesics, find the equation for the spatial components of the acceleration.
 - Find the low velocity limit ($dx^i/d\tau \ll 1$) of the spatial acceleration, assuming Ω is independent of t .

7. (a) Show that the affinely parametrized geodesic equation (3.8) is equivalent to the equation

$$\ddot{x}^\beta + \Gamma^\beta_{\mu\nu} \dot{x}^\mu \dot{x}^\nu = 0, \quad (3.14)$$

where the *Christoffel* symbol $\Gamma^\beta_{\mu\nu}$ is defined by

$$\Gamma^\beta_{\mu\nu} := \frac{1}{2} g^{\beta\gamma} (g_{\gamma\mu,\nu} + g_{\gamma\nu,\mu} - g_{\mu\nu,\gamma}), \quad (3.15)$$

with $g^{\beta\gamma}$ defined as the *inverse metric*,

$$g^{\beta\gamma} g_{\gamma\sigma} = \delta^\beta_\sigma. \quad (3.16)$$

Note that in a local inertial coordinate system at a point p one has $\Gamma^\beta_{\mu\nu}|_p = 0$.

- Show that although \ddot{x}^β does not transform as a vector under a change of coordinates, and neither does $\Gamma^\beta_{\mu\nu} \dot{x}^\mu \dot{x}^\nu$, the sum $\ddot{x}^\beta + \Gamma^\beta_{\mu\nu} \dot{x}^\mu \dot{x}^\nu$ is a vector. This vector is called the *covariant acceleration vector* of the curve $x^\beta(\lambda)$.
8. *Rotation symmetry in the Euclidean plane* about the origin gives rise to a Killing vector field ξ^i defined up to an overall constant rescaling.
- Sketch ξ^i on the plane.
 - Give the components of ξ^i in both polar and Cartesian coordinates.
 - Calculate the norm of ξ^i as a function of position. Explain geometrically why the norm is not a constant.
 - Evaluate in both polar and Cartesian coordinates the quantity $g_{ij} \xi^i \dot{x}^j$ that is conserved along affinely parametrized geodesics in the plane, and show geometrically that it is indeed conserved.

9. *Killing's equation in general coordinates*: Find a covariant equation satisfied by any Killing vector ξ^λ by using the fact that $\frac{d}{d\lambda}(g_{\alpha\nu}\xi^\alpha\dot{x}^\nu) = 0$ along *any* affinely parametrized geodesic $x^\mu(\lambda)$. Show that in a coordinate system for which the components of ξ^λ are $\delta_{\hat{\alpha}}^\lambda$ this equation reduces to the simple statement that $g_{\mu\nu,\hat{\alpha}} = 0$, i.e., the metric components are independent of $x^{\hat{\alpha}}$.
10. *Synchronous or Gaussian Normal Coordinates*: For any spacetime metric, one can always find coordinates (t, x^i) such that the line element takes the form

$$ds^2 = -dt^2 + g_{ij}dx^i dx^j \quad (3.17)$$

($i, j = 1, 2, 3$), although the coordinates will in general be singular beyond some region. To construct such a coordinate system, start with an arbitrary 3-dimensional spacelike surface Σ_0 , labeled with coordinates x^i . At each point of Σ_0 fire the geodesic orthogonal to Σ_0 and use proper time along these geodesics as the fourth coordinate. By construction on Σ_0 we have $g_{00} = -1$ and $g_{0i} = 0$ so, on Σ_0 , the line element takes the above form. Show that it has this form *everywhere* (until the geodesics cross) by showing that $\partial g_{0\mu}/\partial t = 0$ as a consequence of the geodesic equation.

11. *Free-fall coordinates*: Show that, given a geodesic γ , it is always possible to choose a coordinate system that is locally inertial at every point along γ . This is in general *not* possible for an arbitrary nongeodesic curve. (For a discussion of this coordinate system see *Gravitation*, by C.W. Misner, K.S. Thorne, and J.A. Wheeler (Freeman), section 13.6.)
12. *Kaluza-Klein theory*: Imagine a 5-dimensional spacetime with line element

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu + (dx^5 + A_\mu dx^\mu)^2, \quad (3.18)$$

where $\mu, \nu = 0, 1, 2, 3$ and $g_{\mu\nu}$ and A_μ are independent of the coordinate x^5 . Consider the equation for a geodesic $(x^\alpha(\tau), x^5(\tau))$ in this metric. Because the metric components are independent of x^5 , the momentum $p_5 := \dot{x}^5 + A_\mu \dot{x}^\mu$ is conserved. Show that $x^\alpha(\tau)$ satisfies the geodesic equation with an additional term of the form $(e/m)F_{\mu\nu}\dot{x}^\nu$, where $e/m \equiv p_5$ is the (conserved) momentum in the x^5 direction, and $F_{\mu\nu} = A_{\mu,\nu} - A_{\nu,\mu}$ is the usual electromagnetic field strength tensor corresponding to a 4-potential A_μ .

Chapter 4

Special Relativity

4.1 Minkowski space

The spacetime of special relativity is a flat spacetime that can be covered by a single coordinate system (t, x, y, z) in terms of which the line element takes the Minkowski form

$$ds^2 = -dt^2 + dx^2 + dy^2 + dz^2. \quad (4.1)$$

This spacetime is called *Minkowski space* or *Minkowski spacetime* and the metric is called the *Minkowski metric*. There is a 10-parameter family of coordinate systems in which ds^2 takes the Minkowski form. They are linearly related to each other via some combination of translations (4), rotations (3) and boosts (3). Coordinates for which the line element takes the Minkowski form (4.1) are called *inertial* or *Minkowski coordinates*. In Minkowski coordinates, the metric components have the values

$$g_{\mu\nu} = \eta_{\mu\nu} = \text{diag}(-1, 1, 1, 1) \quad (4.2)$$

The proper time along an infinitesimal displacement is given in terms of inertial coordinates by

$$d\tau = \sqrt{-ds^2} = dt \sqrt{1 - (dx/dt)^2 - (dy/dt)^2 - (dz/dt)^2}. \quad (4.3)$$

Define γ by $\gamma = (1 - v^2)^{-1/2}$, with $v^2 = v^i v^i$ and $v^i = dx^i/dt$, $i = x, y, z$. Then one can write

$$d\tau = \gamma^{-1} dt. \quad (4.4)$$

Note that $d\tau \leq dt$, since $\gamma \geq 1$, and $d\tau = dt$ only when $v^i = 0$. This leads to the *twin effect*: The proper time along a timelike curve joining two events A and B at time coordinates t_1 and t_2 is given by

$$\Delta\tau = \int_{t_A}^{t_B} \gamma^{-1} dt \leq t_B - t_A.$$

If the inertial coordinates are chosen so that the coordinates of the two events differ only in t , the curve with longest proper time joining then two events will be the one with zero velocity, i.e., the straight line.

Figure 4.1: Twin Effect

Along any curve other than the straight one, the elapsed proper time is *less* than $t_B - t_A$. In fact, $\Delta\tau$ can be made arbitrarily small by traveling arbitrarily “close” to the light cone, as suggested by path ACB in Fig. 4.1.¹

4.2 4-velocity and 4-acceleration

Along a timelike curve it is possible and convenient to use the proper time as a parameter. With proper time parametrization, the tangent vector to the curve $\dot{x}^\mu = dx^\mu/d\tau$, also called the *4-velocity* or just *velocity*, is always a unit vector:

$$\eta_{\mu\nu} \dot{x}^\mu \dot{x}^\nu = \frac{\eta_{\mu\nu} dx^\mu dx^\nu}{d\tau^2} = \frac{ds^2}{d\tau^2} = -1. \quad (4.5)$$

¹Of course an observer O' along ACB would say it is the observer O along AB , and not himself, that is “close” to the light cone. The situation is not symmetric however, since O' accelerates at C whereas O is unaccelerated everywhere.

(In fact this holds in an arbitrary curved spacetime as well, as was already discussed in section 3.1) Using eqn. (4.4) the 4-velocity can be expressed in terms of the coordinate velocity v^i as

$$\dot{x}^\mu = \frac{dt}{d\tau} \frac{dx^\mu}{dt} = (\gamma, \gamma v^i). \quad (4.6)$$

With proper time parametrization, (4.5) implies that the 4-acceleration \ddot{x}^μ is always orthogonal to the 4-velocity:

$$0 = \frac{d}{d\tau} (\eta_{\mu\nu} \dot{x}^\mu \dot{x}^\nu) = 2\eta_{\mu\nu} \dot{x}^\mu \ddot{x}^\nu \quad (4.7)$$

At each point P on the timelike worldline of a particle there is an inertial coordinate system in which the particle is instantaneously at rest. In such a *co-moving* coordinate system the time axis is tangent to the worldline at P (or parallel to it), the 3-velocity v^i of the particle vanishes at P , and the 4-velocity is just $\dot{x}^\mu|_P = (1, 0, 0, 0)$. According to the orthogonality relation (4.7), the co-moving 4-acceleration thus takes the form

$$\ddot{x}^\mu|_P = (0, a^i),$$

where $a^i|_P = \ddot{x}^i|_P$. The squared norm of the 4-acceleration is thus given by

$$\eta_{\mu\nu} \ddot{x}^\mu \ddot{x}^\nu = a^i a^i, \quad (4.8)$$

where $a^i = \ddot{x}^i$ is the “proper acceleration,” i.e. the 3-acceleration as measured in the instantaneous rest frame.

4.3 4-momentum

Energy and momentum conservation are unified as conservation of the total 4-momentum vector in special relativity. The vector sum of the 4-momenta of a system of particles is conserved in collisions, absorption and emission processes.

The *4-momentum* of a particle of *rest-mass* $m \neq 0$ is defined as

$$p^\mu := m \dot{x}^\mu, \quad (4.9)$$

which is a timelike vector. From the unit normalization of \dot{x}^μ (4.5) it follows that

$$\eta_{\mu\nu} p^\mu p^\nu = -m^2. \quad (4.10)$$

This invariant equation provides a more general definition of the rest mass that generalizes to massless particles as well as to quantum theory, where there is no particle trajectory and equation (4.9) is not applicable.

The *energy* E and *momentum* p^i in a particular coordinate system are defined by $p^\mu = (E, p^i)$. From (4.9) and (4.6) we have therefore

$$(E, p^i) = (\gamma m, \gamma m v^i) = (E, E v^i), \quad (4.11)$$

and the normalization equation (4.10) becomes

$$E^2 = p^i p^i + m^2. \quad (4.12)$$

Expanding $\gamma = (1 - v^2)^{-\frac{1}{2}} = 1 + \frac{1}{2}v^2 + \frac{3}{8}v^4 + \dots$ yields

$$E = \gamma m = m + \frac{1}{2}mv^2 + \frac{3}{8}mv^4 + 0(v^6). \quad (4.13)$$

The first term is the *rest energy*, the second is the *non-relativistic kinetic energy*, and the remainder is the relativistic “corrections.” Relativistically, the “kinetic energy” is just $E - m = (\gamma - 1)m$. Note that if a massive particle were to move on a lightlike worldline, its 4-momentum would diverge since γ would diverge.

For a massless particle $m\dot{x}^\mu$ vanishes unless the particle worldline is lightlike, in which case $d\tau = 0$ along the worldline and $m\dot{x}^\mu$ has the undefined value $0 \cdot \infty$. Nevertheless, the normalization equation (4.10) has a fine limit as $m^2 \rightarrow 0$, indicating that a massless particle has a lightlike 4-momentum vector. That is, $\eta_{\mu\nu}p^\mu p^\nu = 0$, or $E^2 = p^i p^i$. The 4-momentum of a massless point particle following a geodesic world line can be written as $p^\mu = dx^\mu/d\lambda$, where λ is an affine parameter scaled so as to yield the correct magnitude for p^μ . Quantum mechanically, for instance, a photon is a massless “excitation” with an energy momentum 4-vector $p^\mu = \hbar k^\mu$, where the wave-vector $k^\mu = (\omega, k^i)$ is null.

4.4 Voyage to the galactic center

In this section we consider the twin effect in a quantitative example. This will serve to illustrate how special relativistic kinematics and conservation laws can be applied to accelerated motion.

Suppose a spaceship travels from the earth to the center of the galaxy, at constant proper acceleration $g = 9.8\text{m/s}^2$ to the halfway point, then at proper deceleration g until the center is reached, with the same procedure on the trip home. How much time elapses for the voyagers during the round trip?

Figure 4.2:

First of all, let's neglect gravitational effects, and the motion of the earth relative to the center of the galaxy. In a spacetime diagram, the voyage is depicted in Figure 3. The distance to the center of the galaxy is $d = 30,000$ ly, meaning that 60,000 years passes on the earth between events A and A'' , where AA' and $A'A''$ are light-like lines.

In an inertial coordinate system at rest with respect to the earth, the world line of the spaceship is given by a curve $x^\mu(\tau) = (t(\tau), x(\tau), 0, 0)$, if we line up the x-axis with the direction of travel. We choose the origins of coordinates and proper time at the departure from earth, so that $(t(0), x(0)) = (0, 0)$.

If τ is the proper time along this curve, we have from (4.5) and (4.8) the following two scalar equations:

$$\eta_{\mu\nu} \dot{x}^\mu \dot{x}^\nu = -1 = -\dot{t}^2 + \dot{x}^2 \quad (4.14)$$

$$\eta_{\mu\nu} \ddot{x}^\mu \ddot{x}^\nu = g^2 = -\ddot{t}^2 + \ddot{x}^2 \quad (4.15)$$

Equation (4.14) implies that $(\dot{t}, \dot{x}) = (\text{ch}\alpha, \text{sh}\alpha)$ for some function $\alpha(\tau)$. Substituting this into (4.15) then yields $\dot{\alpha}^2 = g^2$. On the first leg of the journey, the relevant solution is $\alpha = g\tau$, and

$$(\dot{t}, \dot{x}) = (\text{ch}g\tau, \text{sh}g\tau) \quad (4.16)$$

$$(t, x) = g^{-1}(\text{sh}g\tau, \text{ch}g\tau - 1). \quad (4.17)$$

This gives the coordinates of the spaceship's accelerating worldline as a function of its proper time.

Halfway through the outgoing trip we have

$$x = \frac{1}{2}d = g^{-1}(\text{ch}g\tau - 1),$$

or

$$\text{ch}g\tau = \frac{1}{2}gd + 1 \quad (4.18)$$

If the acceleration lasts long enough so that the speed is close to that of light, then $g\tau \gg 1$ (and therefore also $gd \gg 1$), and the solution to (4.18) is well approximated by

$$\tau = g^{-1} \ln gd. \quad (4.19)$$

The dependence on d is therefore quite weak. As long as gd is not too huge, the proper time for the trip is of the order of g^{-1} , the time to accelerate to close to the speed of light relative to the initial rest frame.

Now let's put in the numbers. We have $g = 9.8\text{m/s}^2 = 1.03c/y$, so it is a cute accident that the surface gravity of the earth is just about $1y^{-1}$ in our units. Thus for $d = 30,000$ light-years we have $gd = 30,000$, so for the first half of the outgoing trip $\tau = \ln 30,000 = 10.3$ years. From the symmetry of the four segments of the trip we conclude that, during the round trip, the elapsed time on the spaceship is $\tau \simeq 41.2$ years. Meanwhile, the elapsed time on the earth is $t = g^{-1}\text{sh}g\tau = 60,004$ years, to the approximations we have made.

At the midway point, the ship has a γ -factor of $\gamma = \dot{t} = 15,001$ relative to the earth, and a speed of roughly $(1 - 2 \times 10^{-9})c$.² During most of the trip (all but the first, last, and middle two years), the spaceship is moving at close to the speed of light relative to the earth. This is why the total elapsed time on the earth is only 4 years more than twice the light travel time to the center.

Finally, suppose that instead of traveling to the center of our galaxy the voyagers wished to make a round trip to a destination, say, 3 billion light years away, 10^5 times further than the center of the galaxy. According to equation (4.19), this would lengthen the proper time required for the round trip by only $4 \times \ln 10^5 y \simeq 46$ years.

What are the fuel requirements for such a voyage? The most efficient possible rocket (from the point of view of minimizing the mass of the fuel) is one

²By way of comparison, a proton in the Tevatron accelerated up to an energy of 1 TeV has a γ -factor of $\gamma = E/m = 1\text{TeV}/939\text{MeV} \simeq 10^3$.

that ejects exhaust at the speed of light.³ For instance, one can imagine a matter-anti-matter annihilation rocket that ejects γ -rays out the back with perfect collimation. For such a rocket it is particularly easy to determine the mass of the required fuel.

The energy and momentum of the exhaust are equal to each other since the exhaust is massless. Thus 4-momentum conservation on the first half of the outgoing trip implies that the change in energy of the spaceship is equal to minus the change in its momentum. Thus rest frame of the earth one has $E + p = \text{constant} = m_i$, where m_i is the initial rest mass of the spaceship plus fuel. At the halfway point of the outgoing leg, $2\gamma \simeq gd \gg 1$, so $E + p = \gamma m_f(1 + v) \simeq 2\gamma m_f$. Therefore $m_0/m \simeq 2\gamma \simeq gd$.

The ratio m_i/m_f can also be computed in a slightly more “invariant” fashion, as will now be shown for the sake of illustration. 4-momentum conservation of ship-exhaust system is expressed by the 4-vector equation $p_i - p_f = k$, where p_i and p_f are the initial and final 4-momenta of the ship plus fuel, and k is the null 4-momentum of the exhaust. “Squaring” both sides of this relation (i.e., taking the squared norm) yields $-m_i^2 - m_f^2 - 2p_i \cdot p_f = 0$. We can evaluate the invariant $p_i \cdot p_f$ in the initial rest frame, where $p_i = (m_i, 0)$ and $p_f = (\gamma m_f, \gamma m_f v)$, yielding $p_i \cdot p_f = -\gamma m_i m_f$. We thus have $m_i^2 + m_f^2 = 2\gamma m_i m_f$. Since $2\gamma \simeq gd \gg 1$ is so large we therefore have $m_i/m_f \simeq gd$.

A similar computation applies to each of the other three parts of the round trip. Thus, for a complete round trip, the initial mass of the ship plus fuel must be $(gd)^4$ times greater than the final mass of the returning ship. For the voyage to the center of the galaxy this yields a mass ratio of $(30,000)^4 = 8.1 \times 10^{17}$. For a ship of mass 10^7 kg, the fuel must have a mass of 8×10^{24} kg, or just about the mass of the earth. For a voyage to a destination 3 billion light years away, the mass ratio would be greater by a factor of $(10^5)^4 = 10^{20}$. The initial mass would therefore need to be about 10^{45} kg, or about 10^{15} solar masses. It would clearly be a good idea to design a “ramjet” engine to eliminate the need to carry all the fuel with the ship.

³Exercise: Formulate this statement precisely and prove or disprove it.