To appear in: *Handlbook of the Philosophy of Physics*,
eds. J. Butterfield and J. Earman, Elsevier

Classical Relativity Theory*

Version 2.4

David B. Malament
Department of Logic and Philosophy of Science
3151 Social Science Plaza
University of California, Irvine
Irvine, CA 92697-5100
dmalamen@uci.edu

Abstract

The essay that follows is divided into two parts. In the first (section 2), I give a brief account of the structure of classical relativity theory. In the second (section 3), I discuss three special topics: (i) the status of the relative simultaneity relation in the context of Minkowski spacetime; (ii) the "geometrized" version of Newtonian gravitation theory (also known as Newton-Cartan theory); and (iii) the possibility of recovering the global geometric structure of spacetime from its "causal structure".

Keywords: relativity theory; spacetime structure; simultaneity; Newtonian gravitation theory

*I am grateful to Jeremy Butterfield, Erik Curiel, and John Earman for comments on earlier drafts.

Contents

1 Introduction

The essay that follows is divided into two parts. In the first, I give a brief account of the structure of classical relativity theory.[1] In the second, I discuss three special topics.

My account in the first part (section 2) is limited in several respects. I do not discuss the historical development of classical relativity theory, nor the evidence we have for it. I do not treat "special relativity" as a theory in its own right that is superseded by "general relativity". And I do not describe known exact solutions to Einstein's equation. (This list could be continued at great length.[2]) Instead, I limit myself to a few fundamental ideas, and present them as clearly and precisely as I can. The account presupposes a good understanding of basic differential geometry, and at least passing acquaintance with relativity theory itself.[3]

In section 3, I first consider the status of the relative simultaneity relation in the context of Minkowski spacetime. At issue is whether the standard relation, the one picked out by Einstein's "definition" of simultaneity, is conventional in character, or is rather in some significant sense forced on us. Then I describe the "geometrized" version of Newtonian gravitation theory (also known as Newton-Cartan theory). It is included here because it helps to clarify what is and is not distinctive about classical relativity theory. Finally, I consider to what extent the global geometric structure of spacetime can be recovered from its "causal structure".[4]

[1] I speak of "classical" relativity theory because considerations involving quantum mechanics will play no role. In particular, there will be no discussion of quantum field theory in curved spacetime, or of attempts to formulate a quantum theory of gravitation. (For the latter, see Rovelli (this volume, chapter 12).)

[2] Two important topics that I do not consider figure centrally in other contributions to this volume, namely the initial value formulation of relativity theory (Earman, chapter 15), and the Hamiltonian formulation of relativity theory (Belot, chapter 2).

[3] A review of the needed differential geometry (and "abstract-index notation" that I use) can be found, for example, in Wald [1984] and Malament [unpublished]. (Some topics are also reviewed in sections 3.1 and 3.2 of Butterfield (this volume, chapter 1).) In preparing part 1, I have drawn heavily on a number of sources. At the top of the list are Geroch [unpublished], Hawking and Ellis [1972], O'Neill [1983], Sachs and Wu [1977a; 1977b], and Wald [1984].

[4] Further discussion of the foundations of classical relativity theory, from a slightly different point of view, can be found in Rovelli (this volume, chapter 12).

2 The Structure of Relativity Theory

2.1 Relativistic Spacetimes

Relativity theory determines a class of geometric models for the spacetime structure of our universe (and subregions thereof such as, for example, our solar system). Each represents a possible world (or world-region) compatible with the constraints of the theory. It is convenient to describe these models in stages. We start by characterizing a broad class of "relativistic spacetimes", and discussing their interpretation. Later we introduce further restrictions involving global spacetime structure and Einstein's equation.

We take a *relativistic spacetime* to be a pair (M, g_{ab}), where M is a smooth, connected, four-dimensional manifold, and g_{ab} is a smooth, semi-Riemannian metric on M of Lorentz signature $(1,3)$.[5]

We interpret M as the manifold of point "events" in the world.[6] The interpretation of g_{ab} is given by a network of interconnected physical principles. We list three in this section that are relatively simple in character because they make reference only to point particles and light rays. (These objects alone suffice to determine the metric, at least up to a constant.) In the next section, we list a fourth that concerns the behavior of (ideal) clocks. Still other principles involving generic matter fields will come up later.

We begin by reviewing a few definitions. In what follows, let (M, g_{ab}) be a fixed relativistic spacetime, and let ∇_a be the derivative operator on M determined by g_{ab}, i.e., the unique (torsion-free) derivative operator on M satisfying

---

[5] The stated signature condition is equivalent to the requirement that, at every point p in M, the tangent space M_p has a basis $\overset{1}{\xi}{}^a, ..., \overset{4}{\xi}{}^a$ such that, for all i and j in $\{1, 2, 3, 4\}$,

$$\overset{i}{\xi}{}^a \, \overset{i}{\xi}_a = \begin{cases} +1 & \text{if } i = 1 \\ -1 & \text{if } i = 2,\, 3,\, 4 \end{cases}$$

and $\overset{i}{\xi}{}^a \, \overset{j}{\xi}_a = 0$ if $i \neq j$. (Here we are using the abstract-index notation. 'a' is an abstract index, while 'i' and 'j' are normal counting indices.) It follows that given any vectors $\eta^a = \sum_{i=1}^{4} \overset{i}{k} \, \overset{i}{\xi}{}^a$, and $\rho^a = \sum_{j=1}^{4} \overset{j}{l} \, \overset{j}{\xi}{}^a$ at p,

$$\begin{aligned} \eta^a \rho_a &= \overset{1}{k}\,\overset{1}{l} - \overset{2}{k}\,\overset{2}{l} - \overset{3}{k}\,\overset{3}{l} - \overset{4}{k}\,\overset{4}{l} \\ \eta^a \eta_a &= \overset{1}{k}\,\overset{1}{k} - \overset{2}{k}\,\overset{2}{k} - \overset{3}{k}\,\overset{3}{k} - \overset{4}{k}\,\overset{4}{k}. \end{aligned}$$

[6] We use 'event' as a neutral term here and intend no special significance. Some might prefer to speak of "equivalence classes of coincident point events", or "point event locations", or something along those lines.

the compatibility condition $\nabla_a\, g_{bc} = \mathbf{0}$.

Given a point p in M, and a vector η^a in the tangent space M_p at p, we say η^a is:

$$\begin{array}{lllll} \textit{timelike} & \text{if} & \eta^a\eta_a & > & 0 \\ \textit{null (or lightlike)} & \text{if} & \eta^a\eta_a & = & 0 \\ \textit{causal} & \text{if} & \eta^a\eta_a & \geq & 0 \\ \textit{spacelike} & \text{if} & \eta^a\eta_a & < & 0. \end{array}$$

In this way, g_{ab} determines a "null-cone structure" in the tangent space at every point of M. Null vectors form the boundary of the cone. Timelike vectors form its interior. Spacelike vectors fall outside the cone. Causal vectors are those that are either timelike or null. This classification extends naturally to *curves*. We take these to be smooth maps of the form $\gamma\colon I \to M$ where $I \subseteq \mathbb{R}$ is a (possibly infinite, not necessarily open) interval.[7] γ qualifies as *timelike* (respectively *null*, *causal*, *spacelike*) if its tangent vector field $\vec{\gamma}$ is of this character at every point.

A curve $\gamma_2\colon I_2 \to M$ is called an *(orientation preserving) reparametrization* of the curve $\gamma_1\colon I_1 \to M$ if there is a smooth map $\tau\colon I_2 \to I_1$ of I_2 onto I_1, with positive derivative everywhere, such that $\gamma_2 = (\gamma_1 \circ \tau)$. The property of being timelike, null, etc. is preserved under reparametrization.[8] So there is a clear sense in which our classification also extends to *images* of curves.[9]

A curve $\gamma\colon I \to M$ is said to be a *geodesic (with respect to g_{ab})* if its tangent field ξ^a satisfies the condition: $\xi^n \nabla_n\, \xi^a = \mathbf{0}$. The property of being a geodesic is not, in general, preserved under reparametrization. So it does not transfer to curve images. But, of course, the related property of being a *geodesic up to reparametrization* does carry over. (The latter holds of a curve if it can be reparametrized so as to be a geodesic.)

Now we can state the first three interpretive principles. For all curves $\gamma : I \to M$,

C1 γ is timelike iff its image $\gamma[I]$ could be the worldline of a massive point particle (i.e., a particle with positive mass);[10]

[7] If I is not an open set, we can understand smoothness to mean that there is an open interval $\overline{I} \subseteq \mathbb{R}$, with $I \subset \overline{I}$, and a smooth map $\overline{\gamma}\colon \overline{I} \to M$, such that $\overline{\gamma}(s) = \gamma(s)$ for all $s \in I$.

[8] This follows from the fact that, in the case just described, $\vec{\gamma}_2 = \frac{d\tau}{ds}\, \vec{\gamma}_1$, with $\frac{d\tau}{ds} > 0$.

[9] The difference between curves and curve images, i.e., between maps $\gamma\colon I \to M$ and sets $\gamma[I]$, matters. We take worldlines to be instances of the latter, i.e., construe them as point sets rather than parametrized point sets.

[10] We will later discuss the concept of mass in relativity theory. For the moment, we take it to be just a primitive attribute of particles.

C2 γ can be reparametrized so as to be a null geodesic iff $\gamma[I]$ could be the trajectory of a light ray;[11]

P1 γ can be reparametrized so as to be a timelike geodesic iff $\gamma[I]$ could be the worldline of a *free*[12] massive point particle.

In each case, a statement about geometric structure (on the left) is correlated with a statement about the behavior of particles or light rays (on the right).

Several comments and qualifications are called for. First, we are here working within the framework of relativity as traditionally understood, and ignoring speculations about the possibility of particles ("tachyons") that travel faster than light. (Their worldlines would come out as images of spacelike curves.) Second, we have built in the requirement that "curves" be smooth. So, depending on how one models collisions of point particles, one might want to restrict attention here to particles that do not experience collisions.

Third, the assertions require qualification because the status of "point particles" in relativity theory is a delicate matter. At issue is whether one treats a particle's own mass-energy as a source for the surrounding metric field g_{ab} – in addition to other sources that may happen to be present. (Here we anticipate our discussion of Einstein's equation.) If one does, then the curvature associated with g_{ab} may blow up as one approaches the particle's worldline. And in this case one *cannot* represent the worldline as the image of a curve in M, at least not without giving up the requirement that g_{ab} be a smooth field on M. For this reason, a more careful formulation of the principles would restrict attention to "test particles", i.e., ones whose own mass-energy is negligible and may be ignored for the purposes at hand.

Fourth, the modal character of the assertions (i.e., the reference to possibility) is essential. It is simply not true, to take the case of C1, that all timelike curve images *are*, in fact, the worldlines of massive particles. The claim is that, as

[11] For certain purposes, even within classical relativity theory, it is useful to think of light as constituted by streams of "photons", and take the right side condition here to be "$\gamma[I]$ could be the worldline of a photon". The latter formulation makes C2 look more like C1 and P1, and draws attention to the fact that the distinction between massive particles and mass 0 particles (such as photons) has direct significance in terms of relativistic spacetime structure.

[12] "Free particles" here must be understood as ones that do not experience any forces (except "gravity"). It is one of the fundamental principles of relativity theory that gravity arises as a manifestation of spacetime curvature, not as an external force that deflects particles from their natural, straight (geodesic) trajectories. We will discuss this matter further in section 2.4.

least so far as the laws of relativity theory are concerned, they *could* be. Of course, judgments concerning what could be the case depend on what conditions are held fixed in the background. The claim that a particular curve image could be the worldline of a massive point particle must be understood to mean that it could so long as there are, for example, no barriers in the way. Similarly, in C2 there is an implicit qualification. We are considering what trajectories are available to light rays when no intervening material media are present, i.e., when we are dealing with light rays *in vacua*.

Though these four concerns are important and raise interesting questions about the role of idealization and modality in the formulation of physical theory, they have little to do with relativity theory as such. Similar difficulties arise when one attempts to formulate corresponding principles within the framework of Newtonian gravitation theory.

It follows from the cited interpretive principles that the metric g_{ab} is determined (up to a constant) by the behavior of point particles and light rays. We make this claim precise in a pair of propositions about "conformal structure" and "projective structure".

Let $\bar{g}_{ab}$ be a second smooth metric of Lorentz signature on M. We say that $\bar{g}_{ab}$ is *conformally equivalent* to g_{ab} if there is a smooth map $\Omega\colon M \to \mathbb{R}$ on M such that $\bar{g}_{ab} = \Omega^2 g_{ab}$. ($\Omega$ is called a *conformal factor*. It certainly need not be constant.) Clearly, if $\bar{g}_{ab}$ and g_{ab} are conformally equivalent, then they agree in their classification of vectors and curves as timelike, null, etc.. The converse is true as well.[13] Conformally equivalent metrics on M do not agree, in general, as to which curves on M qualify as geodesics or even just as geodesics up to reparametrization. But, it turns out, they do necessarily agree as to which *null* curves are geodesics up to reparametrization.[14] And the converse is true, once

[13]If the two metrics agree as to which vectors and curves belong to any one of the three categories, then they must agree on all. And in that case, they must be conformally equivalent. See Hawking and Ellis [1972, p. 61].

[14]This follows because the property of being the image of a null geodesic can be captured in terms of the existence or non-existence of (local) timelike and null curves connecting points in M. The relevant technical lemma can be formulated as follows.

> A curve $\gamma\colon I \to M$ can be reparametrized so as to be a null geodesic iff γ is null and for all $s \in I$, there is an open set $O \subseteq M$ containing $\gamma(s)$ such that, for all $s_1, s_2 \in I$, if $s_1 \leq s \leq s_2$, and if $\gamma([s_1, s_2]) \subseteq O$, then there is no timelike curve from $\gamma(s_1)$ to $\gamma(s_2)$ within O.

(Here $\gamma([s_1, s_2])$ is the image of γ as restricted to the interval $[s_1, s_2]$.) For a proof, see Hawking and Ellis [1972, p. 103].

again.[15]

Putting the pieces together, we have the following proposition. Clauses (1) and (2) correspond to C1 and C2 respectively.

Proposition 2.1.1. *Let $\bar{g}_{ab}$ be a second smooth metric of Lorentz signature on M. Then the following conditions are equivalent.*

(1) $\bar{g}_{ab}$ and g_{ab} agree as to which curves on M are timelike.

(2) $\bar{g}_{ab}$ and g_{ab} agree as to which curves on M can be reparameterized so as to be null geodesics.

(3) $\bar{g}_{ab}$ and g_{ab} are conformally equivalent.

In this sense, *the spacetime metric g_{ab} is determined up to a conformal factor, independently, by the set of possible worldlines of massive point particles, and by the set of possible trajectories of light rays.*

Next we turn to projective structure. Let $\overline{\nabla}_a$ be a second derivative operator on M. We say that $\overline{\nabla}_a$ and ∇_a are *projectively equivalent* if they agree as to which curves are geodesics up to reparametrization (i.e., if, for all curves γ, γ can be reparametrized so as to be a geodesic with respect to $\overline{\nabla}_a$ iff it can be so reparametrized with respect to ∇_a). And if $\bar{g}_{ab}$ is a second metric on M of Lorentz signature, we say that it is *projectively equivalent* to g_{ab} if its associated derivative operator $\overline{\nabla}_a$ is projectively equivalent to ∇_a.

It is a basic result, due to Hermann Weyl [1921], that if $\bar{g}_{ab}$ and g_{ab} are conformally *and* projectively equivalent, then the conformal factor that relates them must be constant. It is convenient for our purposes, with interpretive principle P1 in mind, to cast it in a slightly altered form that makes reference only to timelike geodesics (rather than arbitrary geodesics).

Proposition 2.1.2. *Let $\bar{g}_{ab}$ be a second smooth metric on M with $\bar{g}_{ab} = \Omega^2 g_{ab}$. If $\bar{g}_{ab}$ and g_{ab} agree as to which timelike curves can be reparametrized so as to be geodesics, then Ω is constant.*

The spacetime metric g_{ab}, we saw, is determined up to a conformal factor, independently, by the set of possible worldlines of massive point particles, and by the set of possible trajectories of light rays. The proposition now makes clear the sense in which it is fully determined (up to a constant) by those sets together with the set of possible worldlines of free massive particles.[16]

[15]For if the metrics agree as to which curves are null geodesics up to reparametrization, they must agree as to which vectors at arbitrary points are null, and this, we know, implies that the metrics are conformally equivalent.

[16]As Weyl put it [1950, p. 103],

Our characterization of relativistic spacetimes is extremely loose. Many further conditions might be imposed. For the moment, we consider just one.

(M, g_{ab}) is said to be *temporally orientable* if there exists a continuous timelike vector field τ^a on M. Suppose the condition is satisfied. Then two such fields τ^a and $\hat{\tau}^a$ on M are said to be *co-oriented* if $\tau^a \hat{\tau}_a > 0$ everywhere, i.e., if τ^a and $\hat{\tau}^a$ fall in the same lobe of the null-cone at every point of M. Co-orientation is an equivalence relation (on the set of continuous timelike vector fields on M) with two equivalence classes. A *temporal orientation* of (M, g_{ab}) is a choice of one of those two equivalence classes to count as the "future" one. Thus, a non-zero causal vector ξ^a at a point of M is said to be *future directed* or *past directed* with respect to the temporal orientation $\mathcal{T}$ depending on whether $\tau^a \xi_a > 0$ or $\tau^a \xi_a < 0$ at the point, where τ^a is any continuous timelike vector field in $\mathcal{T}$. Derivatively, a causal curve $\gamma\colon I \to M$ is said to be *future directed* (resp. *past directed*) with respect to $\mathcal{T}$ if its tangent vectors at every point are.

In what follows, we assume that our background spacetime (M, g_{ab}) is temporally orientable, and that a particular temporal orientation has been specified. Also, given events p and q in M, we write $p \ll q$ (resp. $p < q$) if there is a future-directed timelike (resp. causal) curve that starts at p and ends at q.[17]

2.2 Proper Time

So far we have discussed relativistic spacetime structure without reference to either "time" or "space". We come to them in this section and the next.

Let $\gamma : [s_1, s_2] \to M$ be a future-directed timelike curve in M with tangent field ξ^a. We associate with it an elapsed *proper time* (relative to g_{ab}) given by

$$|\gamma| = \int_{s_1}^{s_2} (g_{ab}\, \xi^a\, \xi^b)^{\frac{1}{2}}\; ds.$$

> ... it can be shown that the metrical structure of the world is already fully determined by its inertial and causal structure, that therefore mensuration need not depend on clocks and rigid bodies but that light signals and mass points moving under the influence of inertia alone will suffice.

(For more on Weyl's "causal-inertial" method of determining the spacetime metric, see Coleman and Korté [2001, section 4.9].)

[17] It follows immediately that if $p \ll q$, then $p < q$. The converse does not hold, in general. But the only way the second condition can be true, without the first being true as well, is if the only future-directed causal curves from p to q are null geodesics (or reparametrizations of null geodesics). See Hawking and Ellis [1972, p. 112].

This elapsed proper time is invariant under reparametrization of γ, and is just what we would otherwise describe as the length of (the image of) γ. The following is another basic principle of relativity theory.

P2 Clocks record the passage of elapsed proper time along their worldlines.

Again, a number of qualifications and comments are called for. Our formulation of C1, C2, and P1 was rough. The present formulation is that much more so. We have taken for granted that we know what "clocks" are. We have assumed that they have worldlines (rather than worldtubes). And we have overlooked the fact that ordinary clocks (e.g., the alarm clock on the nightstand) do not do well at all when subjected to extreme acceleration, tidal forces, and so forth. (Try smashing the alarm clock against the wall.) Again, these concerns are important and raise interesting questions about the role of idealization in the formulation of physical theory. (One might construe an "ideal clock" as a point-sized test object that perfectly records the passage of proper time along its worldline, and then take P2 to assert that real clocks are, under appropriate conditions, to varying degrees of accuracy, approximately ideal.) But as with our concerns about the status of point particles, they do not have much to do with relativity theory as such. Similar ones arise when one attempts to formulate corresponding principles about clock behavior within the framework of Newtonian theory.

Now suppose that one has determined the conformal structure of spacetime, say, by using light rays. Then one can use clocks, rather than free particles, to determine the conformal factor. One has the following simple result, which should be compared with proposition 2.1.2.[18]

Proposition 2.2.1. *Let $\bar{g}_{ab}$ be a second smooth metric on M with $\bar{g}_{ab} = \Omega^2 g_{ab}$. Further suppose that the two metrics assign the same lengths to all timelike curves, i.e., $|\gamma|_{\bar{g}_{ab}} = |\gamma|_{g_{ab}}$ for all timelike curves $\gamma : I \to M$. Then $\Omega = 1$ everywhere. (Here $|\gamma|_{g_{ab}}$ is the length of γ relative to g_{ab}.)*

P2 gives the whole story of relativistic clock behavior (modulo the concerns noted above). In particular, it implies the path dependence of clock readings. If two clocks start at an event p, and travel along different trajectories to an event

[18] Here we not only determine the metric up to a constant, but determine the constant as well. The difference is that here, in effect, we have built in a choice of units for spacetime distance. We could obtain a more exact counterpart to proposition 2.1.2 if we worked, not with intervals of elapsed proper time, but rather with ratios of such intervals.

q, then, in general, they will record different elapsed times for the trip. (E.g., one will record an elapsed time of 3,806 seconds, the other 649 seconds.) This is true no matter how similar the clocks are. (We may stipulate that they came off the same assembly line.) This is the case because, as P2 asserts, the elapsed time recorded by each of the clocks is just the length of the timelike curve it traverses in getting from p to q and, in general, those lengths will be different.

Suppose we consider all future-directed timelike curves from p to q. It is natural to ask if there are any that minimize or maximize the recorded elapsed time between the events. The answer to the first question is 'no'. Indeed, one has the following proposition.

Proposition 2.2.2. *Let p and q be events in M such that $p \ll q$. Then, for all $\epsilon > 0$, there exists a future-directed timelike curve γ from p to q with $|\gamma| < \epsilon$. (But there is no such curve with length 0, since all timelike curves have non-zero length.)*

Though some work is required to give the proposition an honest proof (see O'Neill [1983, pp. 294-5]), it should seem intuitively plausible. If there is a timelike curve connecting p and q, there also exists a jointed, zig-zag null curve that connects them. It has length 0. But we can approximate the jointed null

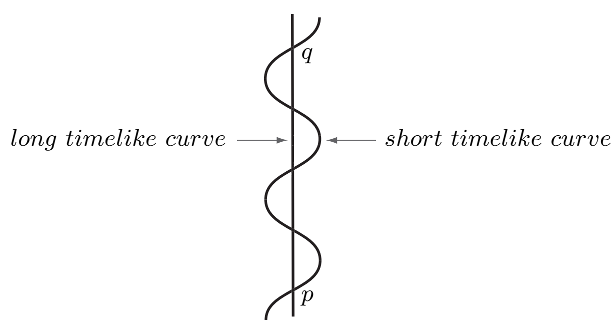

Figure 2.2.1: A long timelike curve from p to q and a very short one that swings back-and-forth, and approximates a broken null curve.

curve arbitrarily closely with smooth timelike curves that swing back and forth. So (by the continuity of the length function), we should expect that, for all $\epsilon > 0$, there is an approximating timelike curve that has length less than ϵ. (See figure 2.2.1.)

The answer to the second question (Can one *maximize* recorded elapsed time between p and q?) is 'yes' if one restricts attention to local regions of spacetime.

In the case of positive definite metrics, i.e., ones with signature of form $(n, 0)$, we know, geodesics are *locally shortest* curves. The corresponding result for Lorentz metrics is that timelike geodesics are *locally longest* curves.

Proposition 2.2.3. *Let $\gamma : I \to M$ be a future-directed timelike curve. Then γ can be reparametrized so as to be a geodesic iff for all $s \in I$, there exists an open set O containing $\gamma(s)$ such that, for all $s_1, s_2 \in I$ with $s_1 \leq s \leq s_2$, if the image of $\overline{\gamma} = \gamma_{|[s_1,s_2]}$ is contained in O, then $\overline{\gamma}$ (and its reparametrizations) are longer than all other timelike curves in O from $\gamma(s_1)$ to $\gamma(s_2)$. (Here $\gamma_{|[s_1,s_2]}$ is the restriction of γ to the interval $[s_1, s_2]$.)*

The proof of the proposition is very much the same as in the positive definite case. (See Hawking and Ellis [1972, p. 105].) Thus of all clocks passing locally from p to q, that one will record the greatest elapsed time that "falls freely" from p to q. To get a clock to read a smaller elapsed time than the maximal value one will have to accelerate the clock. Now acceleration requires fuel, and fuel is not free. So proposition 2.2.3 has the consequence that (locally) "saving time costs money". And proposition 2.2.2 may be taken to imply that (locally) "with enough money one can save as much time as one wants".

The restriction here to local regions of spacetime is essential. The connection described between clock behavior and acceleration does not, in general, hold on a global scale. In some relativistic spacetimes, one can find future-directed timelike geodesics connecting two events that have different lengths, and so clocks following the curves will record different elapsed times between the events even though *both* are in a state of free fall. Furthermore – this follows from the preceding claim by continuity considerations alone – it can be the case that of two clocks passing between the events, the one that undergoes acceleration during the trip records a *greater* elapsed time than the one that remains in a state of free fall.

The connection we have been considering between clock behavior and acceleration was once thought to be paradoxical. (I am thinking of the "clock (or twin) paradox".) Suppose two clocks, A and B, pass from one event to another in a suitably small region of spacetime. Further suppose A does so in a state of free fall, but B undergoes acceleration at some point along the way. Then, we know, A will record a greater elapsed time for the trip than B. This was thought paradoxical because it was believed that "relativity theory denies the possibility of distinguishing "absolutely" between free fall motion and accelerated motion". (If we are equally well entitled to think that it is clock B that is in a state of free fall, and A that undergoes acceleration, then, by parity of reasoning, it should be B that records the greater elapsed time.) The resolution of the paradox, if one can call it that, is that relativity theory makes no such denial. The situations of A and B here are *not* symmetric. The distinction between accelerated motion and free fall makes every bit as much sense in relativity theory as it does in Newtonian physics.

In what follows, unless indication is given to the contrary, a "timelike curve" should be understood to be a future-directed timelike curve, parametrized by elapsed proper time, i.e., by arc length. In that case, the tangent field ξ^a of the curve has unit length ($\xi^a \xi_a = 1$). And if a particle happens to have the image of the curve as its worldline, then, at any point, ξ^a is called the particle's *four-velocity* there.

2.3 Space/Time Decomposition at a Point and Particle Dynamics

Let γ be a timelike curve representing the particle O with four-velocity field ξ^a. Let p be a point on the image of γ, and let λ^a be a vector at p. There is a natural decomposition of λ^a into components parallel to, and orthogonal to, ξ^a:

$$\lambda^a = \underbrace{(\lambda^b \xi_b)\xi^a}_{\textit{parallel to } \xi^a} + \underbrace{(\lambda^a - (\lambda^b \xi_b)\xi^a)}_{\textit{orthogonal to } \xi^a}. \tag{1}$$

These are standardly interpreted, respectively, as the "temporal" and "spatial" components of λ^a (relative to ξ^a). In particular, the three-dimensional subspace of M_p consisting of vectors orthogonal to ξ^a is interpreted as the "infinitesimal" simultaneity slice of O at p.[19] If we introduce the tangent and orthogonal projection operators

$$k_{ab} = \xi_a \, \xi_b \tag{2}$$

$$h_{ab} = g_{ab} - \xi_a \, \xi_b \tag{3}$$

then the decomposition can be expressed in the form

$$\lambda^a = k^a{}_b \, \lambda^b + h^a{}_b \, \lambda^b. \tag{4}$$

We can think of k_{ab} and h_{ab} as the relative temporal and spatial metrics determined by ξ^a. They are symmetric and satisfy

$$k^a{}_b \, k^b{}_c = k^a{}_c \tag{5}$$

$$h^a{}_b \, h^b{}_c = h^a{}_c. \tag{6}$$

[19] Here we simply take for granted the standard identification of "relative simultaneity" with orthogonality. We will return to consider its justification in section 3.1.

Many standard textbook assertions concerning the kinematics and dynamics of point particles can be recovered using these decomposition formulas. For example, suppose that the worldline of a second particle $\overline{O}$ also passes through p and that its four-velocity at p is $\overline{\xi}^a$. (Since ξ^a and $\overline{\xi}^a$ are both future-directed, they are co-oriented, i.e., $(\xi^a\,\overline{\xi}_a) > 0$.) We compute the speed of $\overline{O}$ as determined by O. To do so, we take the spatial magnitude of $\overline{\xi}^a$ relative to O and divide by its temporal magnitude relative to O:

$$v = \mathit{speed\ of}\ \overline{O}\ \mathit{relative\ to}\ O = \frac{\|h^a{}_b\,\overline{\xi}^b\|}{\|k^a{}_b\,\overline{\xi}^b\|}. \tag{7}$$

(Given any vector μ^a, we understand $\|\mu^a\|$ to be $(\mu^a\mu_a)^{\frac{1}{2}}$ if μ^a is causal, and $(-\mu^a\mu_a)^{\frac{1}{2}}$ if it is spacelike.) From (2), (3), (5), and (6), we have

$$\|k^a{}_b\,\overline{\xi}^b\| = (k^a{}_b\,\overline{\xi}^b\,k_{ac}\,\overline{\xi}^c)^{\frac{1}{2}} = (k_{bc}\,\overline{\xi}^b\,\overline{\xi}^c)^{\frac{1}{2}} = (\overline{\xi}^b\,\xi_b) \tag{8}$$

and

$$\|h^a{}_b\,\overline{\xi}^b\| = (-h^a{}_b\,\overline{\xi}^b\,h_{ac}\,\overline{\xi}^c)^{\frac{1}{2}} = (-h_{bc}\,\overline{\xi}^b\,\overline{\xi}^c)^{\frac{1}{2}} = ((\overline{\xi}^b\,\xi_b)^2 - 1)^{\frac{1}{2}}. \tag{9}$$

So

$$v = \frac{((\overline{\xi}^b\,\xi_b)^2 - 1)^{\frac{1}{2}}}{(\overline{\xi}^b\,\xi_b)} < 1. \tag{10}$$

Thus, as measured by O, no massive particle can ever attain the maximal speed 1. (A similar calculation would show that, as determined by O, light always travels with speed 1.) For future reference, we note that (10) implies:

$$\overline{\xi}^b\,\xi_b = \frac{1}{\sqrt{1 - v^2}}. \tag{11}$$

It is a basic fact of relativistic life that there is associated with every point particle, at every event on its worldline, a *four-momentum* (or *energy-momentum*) vector P^a. In the case of a massive particle with four-velocity $\overline{\xi}^a$, P^a is proportional to $\overline{\xi}^a$, and the (positive) proportionality factor is just what we would otherwise call the *mass* (or *rest mass*) m of the particle. So we have $P^a = m\,\overline{\xi}^a$. In the case of a "photon" (or other mass 0 particle), no such characterization is available because its worldline is the image of a null (rather than timelike) curve. But we can still understand its four-momentum vector at the event in question to be a future-directed null vector that is tangent to its worldline there. If we think of the four-momentum vector P^a as fundamental, then we can, in both cases, recover the mass of the particle as the length of P^a: $m = (P^a P_a)^{\frac{1}{2}}$. (It is strictly positive in the first case, and 0 in the second.)

Now suppose a massive particle O has four-velocity ξ^a at an event, and another particle, either a massive particle or a photon, has four-momentum P^a there. We can recover the usual expressions for the energy and three-momentum of the second particle relative to O if we decompose P^a in terms of ξ^a. By (4) and (2), we have

$$P^a = (P^b \xi_b)\, \xi^a \;+\; h^a{}_b P^b. \tag{12}$$

The *energy* relative to O is the coefficient in the first term: $E = P^b \xi_b$. In the case of a massive particle where $P^a = m\,\overline{\xi}^a$, this yields, by (11),

$$E = m\,(\overline{\xi}^b\,\xi_b) = \frac{m}{\sqrt{1\,-\,v^2}}. \tag{13}$$

(If we had not chosen units in which $c = 1$, the numerator in the final expression would have been mc^2 and the denominator $\sqrt{1 - \frac{v^2}{c^2}}$.) The *three-momentum* relative to O is the second term in the decomposition, i.e., the component of P^a orthogonal to ξ^a: $h^a{}_b\,P^b$. In the case of a massive particle, by (9) and (11), it has magnitude

$$p = \|h^a{}_b\, m\,\overline{\xi}^b\| = m\,((\overline{\xi}^b\,\xi_b)^2 - 1)^{\frac{1}{2}} = \frac{m\,v}{\sqrt{1\,-\,v^2}}. \tag{14}$$

Interpretive principle P1 asserts that free particles traverse the images of timelike geodesics. It can be thought of as the relativistic version of Newton's first law of motion. Now we consider acceleration and the relativistic version of the second law. Let $\gamma : I \to M$ be a timelike curve whose image is the worldline of a massive particle O, and let ξ^a be the four-velocity field of O. Then the *four-acceleration* (or just *acceleration*) field of O is $\xi^n \nabla_n\, \xi^a$, i.e., the directional derivative of ξ^a in the direction ξ^a. The four-acceleration vector is orthogonal to ξ^a. (This is clear, since $\xi^a(\xi^n \nabla_n\, \xi_a) \;=\; \frac{1}{2}\,\xi^n \nabla_n\,(\xi^a\,\xi_a) \;=\; \frac{1}{2}\,\xi^n \nabla_n\,(1) \;=\; 0$.) The magnitude $\|\xi^n \nabla_n\, \xi^a\|$ of the four-acceleration vector at a point is just what we would otherwise describe as the Gaussian curvature of γ there. It is a measure of the degree to which γ curves away from a straight path. (And γ is a geodesic precisely if its curvature vanishes everywhere.)

The notion of spacetime acceleration requires attention. Consider an example. Suppose you decide to end it all and jump off the Empire State Building. What would your acceleration history be like during your final moments? One is accustomed in such cases to think in terms of acceleration relative to the earth. So one would say that you undergo acceleration between the time of your jump and your calamitous arrival. But on the present account, that description has things backwards. Between jump and arrival you are *not* accelerating. You are in a state of free fall and moving (approximately) along a spacetime geodesic. But before the jump, and after the arrival, you *are* accelerating. The floor of the observation desk, and then later the sidewalk, push you away from a geodesic path. The all-important idea here is that we are incorporating the "gravitational field" into the geometric structure of spacetime, and particles traverse geodesics if and only if they are acted upon by no forces "except gravity".

The acceleration of any massive particle, i.e., its deviation from a geodesic trajectory, is determined by the forces acting on it (other than "gravity"). If the particle has mass $m > 0$, and the vector field F^a on $\gamma[I]$ represents the vector sum of the various (non-gravitational) forces acting on the particle, then the particle's four-acceleration $\xi^n \nabla_n \xi^a$ satisfies:

$$F^a = m\, \xi^n \nabla_n \xi^a. \tag{15}$$

This is our version of Newton's second law of motion.

Consider an example. Electromagnetic fields are represented by smooth, anti-symmetric fields F_{ab}. (Here "anti-symmetry" is the condition that $F_{ba} = -F_{ab}$.) If a particle with mass $m > 0$, charge q, and four-velocity field ξ^a is present, the force exerted by the field on the particle at a point is given by $q\, F^a_{\ b}\, \xi^b$. If we use this expression for the left side of (15), we arrive at the Lorentz law of motion for charged particles in the presence of an electromagnetic field:

$$q\, F^a_{\ b}\, \xi^b \;=\; m\, \xi^b\, \nabla_b\, \xi^a.[20] \tag{16}$$

2.4 Matter Fields

In classical relativity theory, one generally takes for granted that all that there is, and all that happens, can be described in terms of various *matter fields*, e.g., material fluids and electromagnetic fields.[21] Each such field is represented by one or more smooth tensor (or spinor) fields on the spacetime manifold M. Each is assumed to satisfy field equations involving the fields that represent it and the spacetime metric g_{ab}.

[20] Notice that the equation makes geometric sense. The acceleration vector on the right is orthogonal to ξ^a. But so is the force vector on the left since $\xi_a (F^a_{\ b}\, \xi^b) = \xi^a\, \xi^b F_{ab} = \frac{1}{2}\, \xi^a \xi^b\, (F_{ab} + F_{ba})$, and by the anti-symmetry of F_{ab}, $(F_{ab} + F_{ba}) = \mathbf{0}$.

[21] This being the case, the question arises how (or whether) one can adequately recover talk about "point particles" in terms of the matter fields. We will say just a bit about the question in this section.

For present purposes, the most important basic assumption about the matter fields is the following.

> Associated with each matter field $\mathcal{F}$ is a symmetric smooth tensor field T_{ab} characterized by the property that, for all points p in M, and all future-directed, unit timelike vectors ξ^a at p, $T^a{}_b\,\xi^b$ is the four-momentum density of $\mathcal{F}$ at p as determined relative to ξ^a.

T_{ab} is called the *energy-momentum* field associated with $\mathcal{F}$. The four-momentum density vector $T^a{}_b\,\xi^b$ at p can be further decomposed into its temporal and spatial components relative to ξ^a, just as the four-momentum of a massive particle was decomposed in the preceding section. The coefficient of ξ^a in the first component, $T_{ab}\,\xi^a\xi^b$, is the *energy density* of $\mathcal{F}$ at p as determined relative to ξ^a. The second component, $T_{nb}\,(g^{an} - \xi^a\,\xi^n)\,\xi^b$, is the *three-momentum density* of $\mathcal{F}$ at p as determined relative to ξ^a.

Other assumptions about matter fields can be captured as constraints on the energy-momentum tensor fields with which they are associated. Examples are the following. (Suppose T_{ab} is associated with matter field $\mathcal{F}$.)

Weak Energy Condition: Given any future-directed unit timelike vector ξ^a at any point in M, $T_{ab}\,\xi^a\xi^b \geq 0$.

Dominant Energy Condition: Given any future-directed unit timelike vector ξ^a at any point in M, $T_{ab}\,\xi^a\xi^b \geq 0$ *and* $T^a_b\,\xi^b$ is timelike or null.

Conservation Condition: $\nabla_a\,T^{ab} = \mathbf{0}$ at all points in M.

The first asserts that the energy density of $\mathcal{F}$, as determined by any observer at any point, is non-negative. The second adds the requirement that the four-momentum density of $\mathcal{F}$, as determined by any observer at any point, is a future-directed causal (i.e., timelike or null) vector. It captures the condition that there is an upper bound to the speed with which energy-momentum can propagate (as determined by any observer). It captures something of the flavor of principle C1 in section 2.1, but avoids reference to "point particles".[22]

The conservation condition, finally, asserts that the energy-momentum carried by $\mathcal{F}$ is locally conserved. If two or more matter fields are present in the

[22] This is the standard formulation of the dominant energy condition. The fit with C1 would be even closer if we strengthened the condition slightly so as to be appropriate, specifically, for massive matter fields: at any point p in M, if $T^a_b \neq \mathbf{0}$ there, then $T^a_b\,\xi^b$ is *timelike* for all future-directed unit timelike vectors ξ^a at p.

same region of spacetime, it need not be the case that each one individually satisfies the condition. Interaction may occur. But it *is* a fundamental assumption that the composite energy-momentum field formed by taking the sum of the individual ones satisfies it. Energy-momentum can be transferred from one matter field to another, but it cannot be created or destroyed.

The dominant energy and conservation conditions have a number of joint consequences that support the interpretations just given. We mention two. The first requires a preliminary definition.

Let (M, g_{ab}) be a fixed relativistic spacetime, and let S be an achronal subset of M (i.e., a subset in which there do *not* exist points p and q such that $p \ll q$). The *domain of dependence* $D(S)$ of S is the set of all points p in M with this property: given any smooth causal curve without (past or future) endpoint,[23] if (its image) passes through p, then it necessarily intersects S. For all standard matter fields, at least, one can prove a theorem to the effect that "what happens on S fully determines what happens throughout $D(S)$". (See Earman (this volume, chapter 15).) Here we consider just a special case.

Proposition 2.4.1. *Let S be an achronal subset of M. Further let T_{ab} be a smooth symmetric field on M that satisfies both the dominant energy and conservation conditions. Finally, assume $T_{ab} = \mathbf{0}$ on S. Then $T_{ab} = \mathbf{0}$ on all of $D(S)$.*

The intended interpretation of the proposition is clear. If energy-momentum cannot propagate (locally) outside the null-cone, and if it is conserved, and if it vanishes on S, then it must vanish throughout $D(S)$. After all, how could it "get to" any point in $D(S)$? Note that our formulation of the proposition does not presuppose any particular physical interpretation of the symmetric field T_{ab}. All that is required is that it satisfy the two stated conditions. (For a proof, see Hawking and Ellis [1972, p. 94].)

The next proposition (Geroch and Jang [1975]) shows that, in a sense, if one assumes the dominant energy condition and the conservation condition, then one can *prove* that free massive point particles traverse the images of timelike geodesics. (Recall principle P1 in section 2.3.) The trick is to find a way to talk about "point particles" in the language of extended matter fields.

[23] Let $\gamma : I \to M$ be a smooth curve. We say that a point p in M is a *future-endpoint* of γ if, for all open sets O containing p, there exists an s_0 in I such that for all $s \in I$, if $s \geq s_0$, then $\gamma(s) \in O$, i.e., the image of γ eventually enters and remains in O. (*Past-endpoints* are defined similarly.)

Proposition 2.4.2. *Let $\gamma : I \to M$ be smooth curve. Suppose that given any open subset O of M containing $\gamma[I]$, there exists a smooth symmetric field T_{ab} on M such that:*

(1) T_{ab} satisfies the dominant energy condition;

(2) T_{ab} satisfies the conservation condition;

(3) $T_{ab} = \mathbf{0}$ outside of O;

(4) $T_{ab} \neq \mathbf{0}$ at some point in O.

Then γ is timelike, and can be reparametrized so as to be a geodesic.

The proposition might be paraphrased this way. If a smooth curve in spacetime is such that arbitrarily small free bodies could contain the image of the curve in their worldtubes, then the curve must be a timelike geodesic (up to reparametrization). In effect, we are trading in "point particles" in favor of nested convergent sequences of smaller and smaller extended particles. (Bodies here are understood to be "free" if their internal energy-momentum is conserved. If a body is acted upon by a field, it is only the composite energy-momentum of the body and field together that is conserved.)

Note that our formulation of the proposition takes for granted that we can keep the background spacetime structure (M, g_{ab}) fixed while altering the fields T_{ab} that live on M. This is justifiable only to the extent that, in each case, T_{ab} is understood to represent a test body whose effect on the background spacetime structure is negligible.[24] Note also that we do not have to assume at the outset that the curve γ is timelike. That follows from the other assumptions.

We have here a precise proposition in the language of matter fields that, at least to some degree, captures principle P1 (concerning the behavior of free massive point particles). Similarly, it is possible to capture C2 (concerning the behavior of light) with a proposition about the behavior of solutions to Maxwell's equations in a limiting regime ("the geometrical limit") where wavelengths are small. It asserts, in effect, that when one passes to this limit, packets of electromagnetic waves are constrained to move along (images of) null geodesics. (See Wald [1984, p. 71].)

Now we consider an example. *Perfect fluids* are represented by three objects: a four-velocity field η^a, an energy density field ρ, and an isotropic pressure field

[24]Stronger theorems have been proved (see Ehlers and Geroch [2004]) in which it is not required that the perturbative effect of the extended body disappear entirely at each stage of the limiting process, but only that, in a certain sense, it disappear in the limit.

p (the latter two as determined by a "co-moving" observer at rest in the fluid). In the special case where the pressure p vanishes, one speaks of a *dust field*. Particular instances of perfect fluids are characterized by "equations of state" that specify p as a function of ρ. (Specifically excluded here are such complicating factors as anisotropic pressure, shear stress, and viscosity.) Though ρ is generally assumed to be non-negative (see below), some perfect fluids (e.g., to a good approximation, water) can exert negative pressure. The energy-momentum tensor field associated with a perfect fluid is:

$$T_{ab} = \rho\, \eta_a\, \eta_b - p\,(g_{ab} \;-\; \eta_a\, \eta_b). \tag{17}$$

Notice that the energy-momentum density vector of the fluid at any point, as determined by a co-moving observer (i.e., as determined relative to η^a), is $T^a_{\ b}\, \eta^b = \rho\, \eta^a$. So we can understand ρ, equivalently, as the energy density of the fluid relative to η^a, i.e., $T_{ab}\, \eta^a\, \eta^b$, or as the (rest) mass density of the fluid, i.e., the length of $\rho\, \eta^a$. (Of course, the situation here corresponds to that of a point particle with mass m and four-velocity η^a, as considered in section 2.3.)

In the case of a perfect fluid, the weak energy condition (WEC), dominant energy condition (DEC), and conservation condition (CC) come out as follows.

$$\begin{array}{lcll} \text{WEC} & \Longleftrightarrow & \rho \geq 0 \quad \text{and} & \qquad\; p \,\geq\, -\rho \\ \text{DEC} & \Longleftrightarrow & \rho \geq 0 \quad \text{and} & \rho \,\geq\, p \,\geq\, -\rho \end{array}$$

$$\text{CC} \quad \Longleftrightarrow \quad \begin{cases} (\rho + p)\, \eta^b\, \nabla_b\, \eta^a - (g^{ab} - \eta^a\, \eta^b)\, \nabla_b\, p & = \;\; \mathbf{0} \\ \eta^b\, \nabla_b\, \rho \,+\, (\rho + p)\, (\nabla_b\, \eta^b) & = \;\; 0 \end{cases}$$

Consider the two equations jointly equivalent to the conservation condition. The first is the equation of motion for a perfect fluid. We can think of it as a relativistic version of Euler's equation. The second is an equation of continuity (or conservation) in the sense familiar from classical fluid mechanics. It is easiest to think about the special case of a dust field ($p = 0$). In that case, the equation of motion reduces to the geodesic equation: $\eta^b\, \nabla_b\, \eta^a = \mathbf{0}$. That makes sense. In the absence of pressure, particles in the fluid are free particles. And the conservation equation reduces to: $\eta^b\, \nabla_b\, \rho \,+\, \rho\, (\nabla_b\, \eta^b) = 0$. The first term gives the instantaneous rate of change of the fluid's energy density, as determined by a co-moving observer. The term $\nabla_b\, \eta^b$ gives the instantaneous rate of change of its volume, per unit volume, as determined by that observer. In a more familiar notation, the equation might be written $\dfrac{d\rho}{ds} + \dfrac{\rho}{V}\dfrac{dV}{ds} = 0$ or, equivalently, $\dfrac{d(\rho V)}{ds} = 0$. (Here we use s for elapsed proper time.) It asserts that (in the absence of pressure, as determined by a co-moving observer) the energy contained in an (infinitesimal) fluid blob remains constant, even as its volume changes.

In the general case, the situation is more complex because the pressure in the fluid contributes to its energy (as determined relative to particular observers), and hence to what might be called its "effective mass density". (If you compress a fluid blob, it gets heavier.) In this case, the WEC comes out as the requirement that $(\rho + p) \geq 0$ in addition to $\rho \geq 0$. If we take $h^{ab} = (g^{ab} - \eta^a \eta^b)$, the equation of motion can be expressed as:

$$(\rho + p)\, \eta^b\, \nabla_b\, \eta^a \;=\; h^{ab}\, \nabla_b\, p.$$

This is an instance of the "second law of motion" (15) as applied to an (infinitesimal) blob of fluid. On the left we have: "effective mass density $\times$ acceleration". On the right, we have the force acting on the blob. We can think of it as minus[25] the gradient of the pressure (as determined by a co-moving observer). Again, this makes sense. If the pressure on the left side of the blob is greater than that on the right, it will move to the right. The presence of the non-vanishing term $(p\, \nabla_b \eta^b)$ in the conservation equation is now required because the energy of the blob is *not* constant when its volume changes as a result of the pressure. The equation governs the contribution made to its energy by pressure.

2.5 Einstein's Equation

Once again, let $(M,\, g_{ab})$ be our background relativistic spacetime with a specified temporal orientation.

It is one of the fundamental ideas of relativity theory that spacetime structure is not a fixed backdrop against which the processes of physics unfold, but instead participates in that unfolding. It posits a dynamical interaction between the spacetime metric in any region and the matter fields there. The interaction is governed by *Einstein's field equation*:

$$R_{ab} - \frac{1}{2}\, R\, g_{ab} - \lambda\, g_{ab} \;=\; 8\,\pi\, T_{ab}, \tag{18}$$

or, equivalently,

$$R_{ab} \;=\; 8\,\pi\, (T_{ab} - \frac{1}{2}\, T\, g_{ab}) \;-\; \lambda\, g_{ab}. \tag{19}$$

[25] The minus sign comes in because of our sign conventions.

Here λ is the *cosmological constant*, $R_{ab}\,(= R^n{}_{abn})$ is the *Ricci tensor field*, $R\,(= R^a{}_a)$ is the Riemann scalar curvature field, and T is the contracted field $T^a{}_a$.[26] We start with four remarks about (18), and then consider an alternative formulation that provides a geometric interpretation of sorts.

(1) It is sometimes taken to be a version of "Mach's principle" that "the spacetime metric is uniquely determined by the distribution of matter". And it is sometimes proposed that the principle can be captured in the requirement that "if one first specifies the energy-momentum distribution T_{ab} on the spacetime manifold M, then there is exactly one (or at most one) Lorentzian metric g_{ab} on M that, together with T_{ab}, satisfies (18)". But there is a serious problem with the proposal. In general, one *cannot* specify the energy-momentum distribution in the absence of a spacetime metric. Indeed, in typical cases the metric enters explicitly in the expression for T_{ab}. (Recall the expression (17) for a perfect fluid.) Thus, in looking for solutions to (18), one must, in general, solve simultaneously for the metric *and* matter field distribution.

(2) Given any smooth metric g_{ab} on M, there certainly exists a smooth symmetric field T_{ab} on M that, together with g_{ab}, is a solution to (18). It suffices to define T_{ab} by the left side of the equation. But the field T_{ab} so introduced will not, in general, be the energy-momentum field associated with any known matter field. It will not even satisfy the weak energy condition discussed in section 2.4. With the latter constraint on T_{ab} in place, Einstein's equation is an entirely non-trivial restriction on spacetime structure.

Discussions of spacetime structure in classical relativity theory proceed on three levels according to the stringency of the constraints imposed on T_{ab}. At the first level, one considers only "exact solutions", i.e., solutions where T_{ab} is, in fact, the aggregate energy-momentum field associated with one or more known matter fields. So, for example, one might undertake to find all perfect fluid solutions exhibiting particular symmetries. At the second level, one considers the larger class of what might be called "generic solutions", i.e., solutions where T_{ab} satisfies one or more generic constraints (of which the weak and dominant energy conditions are examples). It is at this level, for example, that the singularity theorems of Penrose and Hawking (Hawking and Ellis [1972]) are proved. Finally, at the third level, one drops all restrictions on T_{ab}, and Einstein's equation plays no role. Many results about global structure are proved at this level, e.g., the assertion that there exist closed timelike curves in any

[26] We use "geometrical units" in which the gravitational constant G, as well as the speed of light c, is 1.

relativistic spacetime (M, g_{ab}) where M is compact.

(3) The role played by the cosmological constant in Einstein's equation remains a matter of controversy. Einstein initially added the term $(-\lambda g_{ab})$ in 1917 to allow for the possibility of a static cosmological model (which, at the time, was believed necessary to properly represent the actual universe).[27] But there were clear problems with doing so. In particular, one does not recover Poisson's equation (the field equation of Newtonian gravitation theory) as a limiting form of Einstein's equation unless $\lambda = 0$. (See point (4) below.) Einstein was quick to revert to the original form of the equation after Hubble's redshift observations gave convincing evidence that the universe is, in fact, expanding. (That the theory suggested the possibility of cosmic expansion before those observations must count as one of its great successes.) Since then the constant has often been reintroduced to help resolve discrepancies between theoretical prediction and observation, and then abandoned when the (apparent) discrepancies were resolved. The controversy continues. Recent observations indicating an accelerating rate of cosmic expansion have led many cosmologists to believe that our universe is characterized by a positive value for λ. (See Earman [2001] for an overview.)

Claims about the value of the cosmological constant are sometimes cast as claims about the "energy-momentum content of the vacuum". This involves bringing the term $(-\lambda g_{ab})$ from the left side of equation (18) to the right, and re-interpreting it as an energy-momentum field, i.e., taking Einstein's equation in the form

$$R_{ab} - \frac{1}{2} R\, g_{ab} \;=\; 8\,\pi\,(T_{ab} + T_{ab}^{VAC}), \tag{20}$$

where $T_{ab}^{VAC} = \dfrac{\lambda}{8\pi}\, g_{ab}$. Here T_{ab} is still understood to represent the aggregate energy-momentum of all normal matter fields. But T_{ab}^{VAC} is now understood to represent the residual energy-momentum associated with empty space. Given any unit timelike vector ξ^a at a point, $(T_{ab}^{VAC}\, \xi^a\, \xi^b)$ is $\dfrac{\lambda}{8\pi}$. So, on this re-interpretation, λ comes out (up to the factor 8π) as the energy-density of the vacuum as determined by *any* observer, at any point in spacetime.

It should be noted that there is a certain ambiguity involved in referring to λ as the cosmological *constant* (and a corresponding ambiguity as to what counts as a solution to Einstein's equation). We can take (M, g_{ab}, T_{ab}) to qualify if it satisfies the equation for some value (or other) of λ. Or, more stringently, one can take it to qualify if it satisfies the equation for some value of λ that is fixed,

[27] He did so for other reasons as well (see Earman [2001]), but I will pass over them here.

once and for all, i.e., the same for all models (M, g_{ab}, T_{ab}). In effect, we have here two versions of "relativity theory". (See Earman [2003] for discussion of what is at stake in choosing between the two.)

(4) It is instructive to consider the relation of Einstein's equation to Poisson's equation, the field equation of Newtonian gravitation theory:

$$\nabla^2 \phi = 4\,\pi\,\rho. \tag{21}$$

Here ϕ is the Newtonian gravitational potential, and ρ is the Newtonian mass density function. In the "geometrized" formulation of the theory that we will consider in section 3.2, one trades in the potential ϕ in favor of a curved derivative operator, and Poisson's equation comes out as

$$R_{ab} = 4\,\pi\rho\,t_{ab}, \tag{22}$$

where R_{ab} is the Ricci tensor field associated with the new curved derivative operator, and t_{ab} is the temporal metric.

The geometrized formulation of Newtonian gravitation was discovered after general relativity (in the 1920s). But now, after the fact, we can put ourselves in the position of a hypothetical investigator who is considering possible candidates for a relativistic field equation, and knows about the geometrized formulation of Newtonian theory. What could be more natural than the attempt to adopt or adapt (22)? In the empty space case ($\rho = 0$), this strategy suggests the equation $R_{ab} = \mathbf{0}$, which is, of course, Einstein's equation (19) for $T_{ab} = \mathbf{0}$ and $\lambda = 0$. This seems to me, by far, the best route to the latter equation. Start with the Newtonian empty space equation ($R_{ab} = \mathbf{0}$) and then simply leave it intact!

No such simple extrapolation is possible in the general case ($\rho \neq 0$). Indeed, I know of no heuristic argument for the full version of Einstein's equation (with or without cosmological constant) that is nearly so convincing. But one can try something like the following. The closest counterparts to (22) would seem to be ones of the form: $R_{ab} = 4\pi K_{ab}$, where K_{ab} is a symmetric tensorial function of T_{ab} and g_{ab}. The possibilities for K_{ab} include T_{ab}, $g_{ab}\,T$, $T_a{}^m T_{mb}$, $g_{ab}\,(T^{mn}T_{mn})$, ..., and linear combinations of these terms. All but the first two involve terms that are second order or higher in T_{ab}. So, for example, in the special case of a dust field with energy density ρ and four-velocity η^a, they will contain occurrences of ρ^n with $n \geq 2$. (E.g., $g_{ab}\,(T^{mn}T_{mn})$ comes out as $\rho^2 g_{ab}$.) But, presumably, only terms first order in ρ should appear if the equation is to have a proper Newtonian limit. This suggests that we look for a field equation of the form

$$R_{ab} = 4\,\pi\,[k\,T_{ab} + l\,g_{ab}T] \tag{23}$$

or, equivalently,[28]

$$R_{ab} - \frac{l}{(k+4l)}\, R\, g_{ab} = 4\,\pi\, k\, T_{ab}, \tag{24}$$

for some real numbers k and l. Let $G_{ab}(k,l)$ be the field on the left side of the equation. It follows from the conservation condition that the field on the right side is divergence free, i.e., $\nabla_a\,(4\,\pi\,k\,T^{ab}) = \mathbf{0}$. So the conservation condition and (24) can hold jointly only if

$$\nabla_a\, G^{ab}(k,l) = \mathbf{0}.$$

But by the "Bianchi identity"(Wald [1984, pp. 39-40]),

$$\nabla_a\,(R^{ab} - \frac{1}{2}\, R\, g^{ab}) = \mathbf{0}. \tag{25}$$

The latter two conditions imply

$$\left[\frac{l}{(k+4l)} - \frac{1}{2}\right]\, \nabla_a(R g^{ab}) = \mathbf{0}.$$

Now $\nabla_a(R g^{ab}) = \mathbf{0}$ is an unreasonable constraint.[29] So the initial scalar term must be 0. Thus, we are left with the conclusion that the conservation condition and (24) can hold jointly only if $k + 2l = 0$, in which case (23) reduces to

$$R_{ab} = 4\,\pi\,k\left[T_{ab} - \frac{1}{2}\, g_{ab}\, T\right]. \tag{26}$$

It remains to argue that k must be 2 if (26) is to have a proper Newtonian limit. To do so, we consider, once again, the special case of a dust field with energy density ρ and four-velocity η^a. Then, $T_{ab} = \rho\,\eta_a\,\eta_b$, and $T = \rho$. If we insert these values in (26) and contract with $\eta^a\,\eta^b$, we arrive at

$$R_{ab}\,\eta^a\,\eta^b = 2\,\pi\,k\,\rho. \tag{27}$$

Now the counterpart to a four-velocity field in Newtonian theory is a vector field of unit temporal length, i.e., a field η^a where $t_{ab}\,\eta^a\,\eta^b = 1$. If we contract the geometrized version of Poisson's equation (22) with $\eta^a\,\eta^b$, we arrive at:

[28] Contraction on 'a' and 'b' in (23) yields: $R = 4\,\pi\,(k{+}4l)\,T$. Solving for T, and substituting for T in (23) yields (24).

[29] It implies that R is constant and, hence, if (23) holds, that T is constant (since (23) implies $R = 4\,\pi\,(k+4l)\,T$). But this, in turn, is an unreasonable constraint on the energy-momentum distribution T_{ab}. E.g., in the case of a dust field with $T_{ab} = \rho\,\eta^a\eta^b$, $T = \rho$, and so the constraint implies that ρ is constant. This is unreasonable since it rules out any possibility of cosmic expansion. (Recall the discussion toward the close of section 2.4.)

$R_{ab}\,\eta^a\,\eta^b \;=\; 4\,\pi\,\rho$. Comparing this expression for $R_{ab}\,\eta^a\,\eta^b$ with that in (27), we are led to the conclusion that $k = 2$, in which case (26) is just Einstein's equation (19) with $\lambda = 0$.

Summarizing now, we have suggested that if one starts with the geometrized version of Poisson's equation (22) and looks for a relativistic counterpart, one is plausibly led to Einstein's equation with $\lambda = 0$. It is worth noting that if we had started instead with a variant of (22) incorporating a "Newtonian cosmological constant"

$$R_{ab} + \lambda\, t_{ab} = 4\,\pi\rho\, t_{ab}, \tag{28}$$

we would have been led instead to Einstein's equation (19) without restriction on λ. We can think of (28) as the geometrized version of

$$\nabla^2\phi + \lambda = 4\,\pi\,\rho. \tag{29}$$

Let's now put aside the question of how one might try to motivate Einstein's equation. However one arrives at it, the equation – let's now take it in the form (18) – can be understood to assert a dynamical connection between a certain tensorial measure of spacetime curvature (on the left side) and the energy-momentum tensor field (on the right side). It turns out that one can reformulate the connection in a way that makes reference only to scalar quantities, as determined relative to arbitrary observers. The reformulation provides a certain insight into the geometric significance of the equation.[30]

Let S be any smooth spacelike hypersurface in M.[31] The background metric g_{ab} induces a (three-dimensional) metric ${}^3g_{ab}$ on S. In turn, this metric determines on S a derivative operator, an associated Riemann curvature tensor field ${}^3R^a{}_{bcd}$, and a scalar curvature field ${}^3R = ({}^3R^a{}_{bca})({}^3g^{bc})$. Our reformulation of Einstein's equation will direct attention to the values of ${}^3R$ at a point for a particular family of spacelike hypersurfaces passing through it.[32]

Let p be any point in M and let ξ^a be any future-directed unit timelike vector at p. Consider the set of all geodesics through p that are orthogonal to

[30] Another approach to its geometrical significance proceeds via the equation of geodesic deviation. See, for example, Sachs and Wu [1977b, p. 114].

[31] We can take this to mean that S is a smooth, imbedded, three-dimensional submanifold of M with the property that any curve $\gamma : I \rightarrow M$ with image in S is spacelike.

[32] In the case of a surface in three-dimensional Euclidean space, the associated Riemann scalar curvature ${}^2R$ is (up to a constant) just ordinary Gaussian surface curvature. We can think of ${}^3R$ in the present context as a higher dimensional analogue that gives averaged values of Gaussian surface curvature. This can be made precise. See, for example, Laugwitz [1965, p. 127].

ξ^a there. The (images of these) curves, at least when restricted to a sufficiently small open set containing p, sweep out a smooth spacelike hypersurface S.[33] (See figure 2.5.1.) We will call it a *geodesic hypersurface*. (We cannot speak

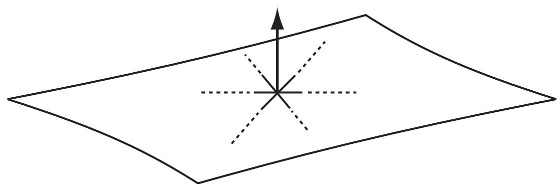

Figure 2.5.1: A "geodesic hypersurface" through a point is constructed by projecting geodesics in all directions orthogonal to a given timelike vector there.

of *the* geodesic hypersurface through p orthogonal to ξ^a because we have left open how far the generating geodesics are extended. But given any two, their restrictions to a suitably small open set containing p coincide.)

Geodesic hypersurfaces are of interest in their own right, the present context aside, because they are natural candidates for a notion of "local simultaneity slice" (relative to a timelike vector at a point). What matters here, though, is that, by the first Gauss-Codazzi equation (Wald [1984, p. 258]), we have

$$^3R = R - 2\,R_{ab}\xi^a\xi^b \tag{30}$$

at p.[34] Here we have expressed the (three-dimensional) Riemann scalar curvature of S at p in terms of the (four-dimensional) Riemann scalar curvature of M at p and the Ricci tensor there. And so, if Einstein's equation (18) holds, we have

$$^3R = -16\pi\,(T_{ab}\,\xi^a\xi^b) + 2\lambda. \tag{31}$$

at p.

[33] More precisely, let S_p be the spacelike hyperplane in M_p orthogonal to ξ^a. Then for any sufficiently small open set O in M_p containing p, the image of $(S_p \cap O)$ under the exponential map $exp : O \to M$ is a smooth spacelike hypersurface. We can take it to be S. (See, for example, Hawking and Ellis [1972, p. 33].)

[34] Let ξ^a – we use the same notation – be the extension of the original vector at p to a smooth future-directed unit timelike vector field on S that is everywhere orthogonal to S. Then the first Gauss-Codazzi equation asserts that at *all* points of S

$$^3R = R - 2\,R_{ab}\xi^a\xi^b + \pi_{ab}\,h^{ab} + \pi_{ab}\,\pi^{ab},$$

where h_{ab} is the spatial projection field $(g_{ab} - \xi_a\xi_b)$ on S, and π_{ab} is the *extrinsic curvature* field $\frac{1}{2}\,\pounds_\xi h_{ab}$ on S. But our construction guarantees that π_{ab} vanish at p.

One can also easily work backwards to recover Einstein's equation at p from the assumption that (31) holds for *all* unit timelike vectors ξ^a at p (and all geodesic hypersurfaces through p orthogonal to ξ^a). Thus, we have the following equivalence.

Proposition 2.5.1. *Let T_{ab} be a smooth symmetric field on M, and let p be a point in M. Then Einstein's equation $R_{ab} - \frac{1}{2}\, R\, g_{ab} + \lambda\, g_{ab} \;=\; 8\,\pi\, T_{ab}$ holds at p iff for all future-directed unit timelike vectors ξ^a at p, and all geodesic hypersurfaces through p orthogonal to ξ^a, the scalar curvature ${}^3R$ of S satisfies ${}^3R = [-16\pi\, (T_{ab}\, \xi^a \xi^b) + 2\lambda]$ at p.*

The result is particularly instructive in the case where $\lambda = 0$. Then (31) directly equates an intuitive scalar measure of spatial curvature (as determined relative to ξ^a) with energy density (as determined relative to ξ^a).

2.6 Congruences of Timelike Curves and "Public Space"

In this section, we consider congruences of timelike curves. We think of them as representing swarms of particles (or fluids). First, we review the standard formalism for describing the local rotation and expansion of such congruences. Then, we consider different notions of "space" and "spatial geometry" as determined relative to them.

Once again, let $(M,\, g_{ab})$ be our background relativistic spacetime (endowed with a temporal orientation). Let ξ^a be a smooth, future-directed, unit timelike vector field on M (or some open subset thereof). We understand it to be the four-velocity field of our particle swarm.

Let h_{ab} be the spatial projection field determined by ξ^a. Then the *rotation* field ω_{ab} and the *expansion* field θ_{ab} are defined as follows:

$$\omega_{ab} \;=\; h_{[a}{}^{m}\, h_{b]}{}^{n}\, \nabla_m\, \xi_n \tag{32}$$

$$\theta_{ab} \;=\; h_{(a}{}^{m}\, h_{b)}{}^{n}\, \nabla_m\, \xi_n. \tag{33}$$

They are smooth fields, orthogonal to ξ^a in both indices, and satisfy

$$\nabla_a\, \xi_b = \omega_{ab} + \theta_{ab} + \xi_a\, (\xi^n \nabla_n\, \xi_b). \tag{34}$$

We can give the two fields ω_{ab} and θ_{ab} a geometric interpretation. Let η^a be a vector field on the worldline of a particle O that is "carried along by the flow of ξ^a", i.e., $\pounds_\xi\, \eta^a = \mathbf{0}$, and is orthogonal to ξ^a at a point p. (Here $\pounds_\xi\, \eta^a$ is the Lie derivative of η^a with respect to ξ^a.) We can think of η^a at p as a spatial "connecting vector" that spans the distance between O and a neighboring particle N that is "infinitesimally close". The instantaneous velocity of N relative to O at p is given by $\xi^n \nabla_n \eta^a$. But $\xi^n \nabla_n \eta^a = \eta^n \nabla_n \xi^a$ (since $\pounds_\xi \eta^a = \mathbf{0}$). So, by (34), and the orthogonality of ξ^a with η^a at p, we have

$$\xi^n \nabla_n \eta^a = ({\omega_n}^a + {\theta_n}^a)\, \eta^n. \tag{35}$$

at the point. Here we have simply decomposed the relative velocity vector into two components. The first, $({\omega_n}^a \eta^n)$, is orthogonal to η^a (since ω_{ab} is anti-symmetric). It gives the instantaneous *rotational velocity* of N with respect to O at p.

In support of this interpretation, consider the instantaneous rate of change of the squared length $(-\eta^a \eta_a)$ of η^a at p. It follows from (35) that

$$\xi^n \nabla_n (-\eta^a \eta_a) = -\,2\, \theta_{na}\, \eta^n \eta^a. \tag{36}$$

Thus the computed rate of change depends solely on θ_{ab}. Suppose $\theta_{ab} = \mathbf{0}$. Then the instantaneous velocity of N with respect to O at p has vanishing radial component. If $\omega_{ab} \neq \mathbf{0}$, N still exhibits a non-zero velocity with respect to O. But it can only be a rotational velocity. The two conditions ($\theta_{ab} = \mathbf{0}$ and $\omega_{ab} \neq \mathbf{0}$) jointly characterize "rigid rotation".

The condition $\omega_{ab} = \mathbf{0}$, by itself, characterizes *irrotational* flow. One gains considerable insight into the condition by considering a second, equivalent formulation. Let us say that the field ξ^a is *hypersurface orthogonal* if there exist smooth, real valued maps f and g (with the same domains of definition as ξ^a) such that, at all points, $\xi_a = f \nabla_a g$. Note that if the condition is satisfied, then the hypersurfaces of constant g value are everywhere orthogonal to ξ^a.[35] Let us further say that ξ^a is *locally hypersurface orthogonal* if the restriction of ξ^a to every sufficiently small open set is hypersurface orthogonal.

Proposition 2.6.1. *Let ξ^a be a smooth, future-directed unit timelike vector field defined on M (or some open subset of M). Then the following conditions are equivalent.*

(1) $\omega_{ab} = \mathbf{0}$ everywhere.

(2) ξ^a is locally hypersurface orthogonal.

[35] For if η^a is a vector tangent to one of these hypersurfaces, $\eta^n \nabla_n g = 0$. So $\eta^n \xi_n = \eta^n (f \nabla_n g) = 0$.

The implication from (2) to (1) is immediate.[36] But the converse is non-trivial. It is a special case of Frobenius's theorem (Wald [1984, p. 436]). The qualification 'locally' can be dropped in (2) if the domain of ξ^a is, for example, simply connected.

There is a nice picture that goes with the proposition. Think about an ordinary rope. In its natural twisted state, the rope cannot be sliced by an infinite family of slices in such a way that each slice is orthogonal to all fibers. But if the rope is first untwisted, such a slicing is possible. Thus orthogonal sliceability is equivalent to fiber untwistedness. The proposition extends this intuitive equivalence to the four-dimensional "spacetime ropes" (i.e., congruences of worldlines) encountered in relativity theory. It asserts that a congruence is irrotational (i.e., exhibits no twistedness) iff it is, at least locally, hypersurface orthogonal

Suppose that our vector field ξ^a *is* irrotational and, to keep things simple, suppose that its domain of definition is simply connected. Then the hypersurfaces to which it is orthogonal are natural candidates for constituting "space" at a given "time" relative to ξ^a or, equivalently, relative to its associated set of integral curves. This is a notion of *public space* to be contrasted with *private space*, which is determined relative to individual timelike vectors or timelike curves.[37] Perhaps the best candidates for the latter are the "geodesic hypersurfaces" we considered, in passing, in section 2.5. (Given a point p and a timelike vector ξ^a there, we took a "geodesic hypersurface through p orthogonal to ξ^a" to be a spacelike hypersurface generated by geodesics through p orthogonal to ξ^a.)

The distinction between public and private space is illustrated in Figure 2.6.1. There we consider a congruence of future-directed timelike half-geodesics in Minkowski spacetime starting at some particular point p. One line O in the congruence is picked out along with a point q on it. Private space relative to O at q is a spacelike hypersurface $S_{private}$ that is flat, i.e., the metric induced on $S_{private}$ has a Riemann curvature tensor field ${}^3R^a{}_{bcd}$ that vanishes everywhere.

[36] Assume that $\xi_a = f \nabla_a g$. Then

$$\begin{aligned}
\omega_{ab} &= h_{[a}{}^m h_{b]}{}^n \nabla_m \xi_n = h_{[a}{}^m h_{b]}{}^n \nabla_m (f \nabla_n g) \\
&= f h_{[a}{}^m h_{b]}{}^n \nabla_m \nabla_n g + h_{[a}{}^m h_{b]}{}^n (\nabla_m f)(\nabla_n g) \\
&= f h_a{}^m h_b{}^n \nabla_{[m} \nabla_{n]} g + h_a{}^m h_b{}^n (\nabla_{[m} f)(\nabla_{n]} g).
\end{aligned}$$

But $\nabla_{[m} \nabla_{n]} g = \mathbf{0}$ since ∇_a is torsion-free, and the second term in the final line vanishes as well since $h_b{}^n \nabla_n g = f^{-1} h_b{}^n \xi_n = \mathbf{0}$. So $\omega_{ab} = \mathbf{0}$.

[37] The distinction between "public space" and "private space" is discussed in Rindler [1981] and Page [1983]. The terminology is due to E. A. Milne.

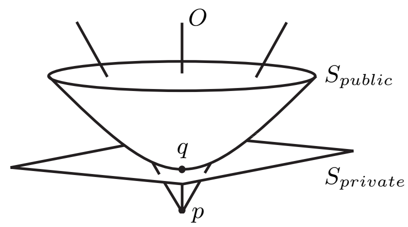

Figure 2.6.1: "Private space" $S_{private}$ at p relative to O, and "public space" S_{public} at p relative to a congruence of timelike curves of which O is a member.

In contrast, public space at q relative to the congruence is a spacelike hypersurface S_{public} of constant negative curvature. If ξ^a is the future-directed unit timelike vector field everywhere tangent to the congruence, and $h_{ab} = (g_{ab} - \xi_a\, \xi_b)$ is its associated spatial projection field, then the curvature tensor field on S_{public} associated with h_{ab} has the form ${}^3R_{abcd} = -\frac{1}{K^2}(h_{ac}\, h_{bd} - h_{ad}\, h_{bc})$, where K is the distance along O from p to q. (This is the characteristic form for a three-manifold of constant curvature $-\frac{1}{K^2}$.)

We have been considering "public space" as determined relative to an irrotational congruence of timelike curves. There is another sense in which one might want to use the term. Consider, for example, "geometry on the surface of a rigidly rotating disk" in Minkowski spacetime. (There is good evidence that Einstein's realization that this geometry is non-Euclidean played an important role in his development of relativity theory (Stachel [1980]).) One needs to ask in what sense the surface of a rotating disk *has* a geometric structure.

We can certainly model the rigidly rotating disk as a congruence of timelike curves in Minkowski spacetime. (Since the disk is two-dimensional, the congruence will be confined to a three-dimensional, timelike submanifold M' of M.) But precisely because the disk is rotating, we cannot find hypersurfaces everywhere orthogonal to the curves and understand the geometry of the disk to be the geometry induced on them – or, strictly speaking, induced on the two-dimensional manifolds determined by the intersection of the putative hypersurfaces with M' – by the background spacetime metric g_{ab}.

The alternative is to think of "space" as constituted by the "manifold of trajectories", i.e., take the individual timelike curves in the congruence to play the role of spatial points, and consider the metric induced on *this* manifold by the background spacetime metric. The construction will not work for an arbitrary congruence of timelike curves. It is essential that we are dealing here with a "stationary" system. (The metric induced on the manifold of trajectories (when the construction works) is fixed and frozen.) But it *does* work for these systems, at least. More precisely, anticipating the terminology of the following section, it works if the four-velocity field of the congruence in question is proportional to a Killing field. (The construction is presented in detail in Geroch [1971, Appendix].)

Thus we have two notions of "public space". One is available if the four-velocity field of the congruence in question is irrotational; the other if it is proportional to a Killing field. Furthermore, if the four-velocity field is irrotational *and* proportional to Killing field, as is the case when we dealing with a "static" system, then the two notions of public space are essentially equivalent.

2.7 Killing Fields and Conserved Quantities

Let κ^a be a smooth vector field on M. We say it is a *Killing field* if $\pounds_\kappa\, g_{ab} = \mathbf{0}$, i.e., if the Lie derivative with respect to κ^a of the metric vanishes.[38] This is equivalent to the requirement that the "flow maps" $\{\Gamma_s\}$ generated by κ^a are all isometries. (See Wald [1984, p. 441].) For this reason, Killing fields are sometimes called "infinitesimal generators of smooth one-parameter families of isometries" or "infinitesimal symmetries". The defining condition can also be expressed as[39]

$$\nabla_{(a}\, \kappa_{b)} \;=\; \mathbf{0}. \tag{37}$$

This is "Killing's equation".

Given any two smooth vector fields ξ^a and μ^a on M, the *bracket* or *commutator* field $[\xi, \mu]^a$ defined by $[\xi, \mu]^a = \pounds_\xi\, \mu^a$ is also smooth. The set of smooth vector fields on M forms a Lie algebra with respect to this operation, i.e., the bracket operation is linear in each slot; it is anti-symmetric ($[\xi, \mu]^a = -[\mu, \xi]^a$); and it satisfies the Jacobi identity

$$[[\xi, \mu], \nu]^a \;+\; [[\nu, \xi], \mu]^a \;+\; [[\mu, \nu], \xi]^a \;=\; \mathbf{0} \tag{38}$$

for all smooth vector fields ξ^a, μ^a, and ν^a on M. It turns out that the bracket field of two Killing fields is also a Killing field. So it follows, as well, that the set of Killing fields on M has a natural Lie algebra structure.

[38] We drop the index on κ here to avoid giving the impression that $\pounds_\kappa\, g_{ab}$ is a three index tensor field. Lie derivatives are always taken with respect to (contravariant) vector fields, so no ambiguity is introduced when the index is dropped. A similar remark applies to our bracket notation below.

[39] This follows since $\pounds_\kappa\, g_{ab} = \kappa^n \nabla_n\, g_{ab} + g_{nb} \nabla_a\, \kappa^n + g_{an} \nabla_b\, \kappa^n$, and ∇_a is compatible with g_{ab}, i.e., $\nabla_n g_{ab} = \mathbf{0}$.

The discussion of smooth symmetries in spacetime, and their associated conserved quantities, is naturally cast in the language of Killing fields. For example, we can use that language to capture precisely the following intuitive notions. A spacetime is *stationary* if it has a Killing field that is everywhere timelike. It is *axially symmetric* if it has a Killing field that is everywhere spacelike, and has integral curves that are closed. (The "axis" in this case is the set of points, possibly empty, where the Killing field vanishes.) Finally, a spacetime is *spherically symmetric* if it has three Killing fields $\overset{1}{\sigma}{}^a, \overset{2}{\sigma}{}^a, \overset{3}{\sigma}{}^a$ that (i) are everywhere spacelike, (ii) are linearly dependent at every point, i.e., $\overset{1}{\sigma}{}^{[a}\, \overset{2}{\sigma}{}^{b}\, \overset{3}{\sigma}{}^{c]} = \mathbf{0}$, and (iii) exhibit the same commutation relations as do the generators of the rotation group in three dimensions:

$$[\overset{1}{\sigma}, \overset{2}{\sigma}]^a = \overset{3}{\sigma}{}^a, \qquad [\overset{2}{\sigma}, \overset{3}{\sigma}]^a = \overset{1}{\sigma}{}^a, \qquad [\overset{3}{\sigma}, \overset{1}{\sigma}]^a = \overset{2}{\sigma}{}^a. \tag{39}$$

Now we consider, very briefly, two types of conserved quantity. One is an attribute of massive point particles, the other of extended bodies. Let κ^a be an arbitrary Killing field, and let $\gamma : I \to M$ be a timelike curve, with unit tangent field ξ^a, whose image is the worldline of a point particle with mass $m > 0$. Consider the quantity $J = (P^a \kappa_a)$, where $P^a = m\, \xi^a$ is the four-momentum of the particle. It certainly need not be constant on $\gamma[I]$. But it will be if γ is a geodesic. For in that case, $\xi^n \nabla_n\, \xi^a = \mathbf{0}$ and hence, by (37),

$$\xi^n \nabla_n J = m\, (\kappa_a\, \xi^n \nabla_n\, \xi^a + \xi^n \xi^a\, \nabla_n\, \kappa_a) \;=\; m\, \xi^n \xi^a\, \nabla_{(n}\, \kappa_{a)} = \mathbf{0}. \tag{40}$$

Thus, the value of J (construed as an attribute of massive point particles) is constant for *free* particles.

We refer to J as the conserved quantity associated with κ^a. If κ^a is timelike, and if the flow maps determined by κ^a have the character of translations[40], then J is called the *energy* of the particle (associated with κ^a).[41] If it is spacelike,

[40] In Minkowski spacetime, one has an unambiguous classification of Killing fields as generators of translations, spatial rotations, boosts (and linear combinations of them). No such classification is available in general. Killing fields are just Killing fields. But sometimes a Killing field in a curved spacetime resembles a Killing field in Minkowski spacetime in certain respects, and then the terminology may carry over naturally. For example, in the case of asymptotically flat spacetimes, one can classify Killing fields by their asymptotic behavior.

[41] If κ^a is of unit length everywhere, this usage accords well with that in section 2.3. For there ascriptions of energy to point particles were made relative to unit timelike vectors, and the value of the energy at any point was taken to be the inner product of that unit timelike vector with the particle's four-momentum vector. If κ^a is, at least, of constant length, then one can always rescale it so as to achieve agreement of usage. But, in general, Killing fields, timelike or otherwise, are not of constant length, and so the current usage must be regarded as a generalization of that earlier usage.

and if the flow maps have the character of translations, then J is called the component of *linear momentum* of the particle (associated with κ^a). Finally, if κ^a is spacelike, and if the flow maps have the character of rotations, then it is called the component of *angular momentum* of the particle (associated with κ^a).

It is useful to keep in mind a certain picture that helps one to "see" why the angular momentum of free particles (to take that example) is conserved. It involves an analogue of angular momentum in Euclidean plane geometry. Figure 2.7.1 shows a rotational Killing field κ^a in the Euclidean plane, the image of a

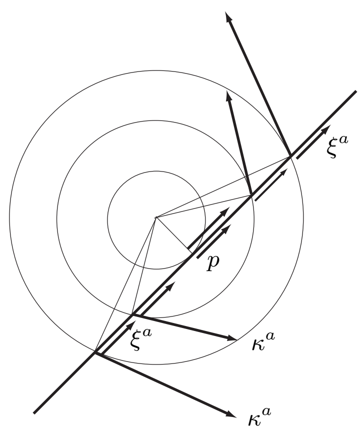

Figure 2.7.1: κ^a is a rotational Killing field. (It is everywhere orthogonal to a circle radius, and proportional to it in length.) ξ^a is a tangent vector field of constant length on the line. The inner-product between them is constant. (Equivalently, the length of the projection of κ^a onto the line is constant.)

geodesic (i.e., a line L), and the tangent field ξ^a to the geodesic. Consider the quantity $J = \xi^a \kappa_a$, i.e., the inner product of ξ^a with κ^a, along L. Exactly the same proof as before (in equation (40)) shows that J is constant along L.[42] But here we can better visualize the assertion.

Let us temporarily drop indices and write $\kappa \cdot \xi$ as one would in ordinary Euclidean vector calculus (rather than $\xi^a \kappa_a$). Let p be the point on L that is closest to the center point where κ vanishes. At that point, κ is parallel to ξ. As one moves away from p along L, in either direction, the length $|\kappa|$ of κ grows, but the angle $\angle(\kappa, \xi)$ between the vectors increases as well. It is at

[42] The mass m played no special role.

least plausible from the picture (and easy to check directly with an argument involving similar triangles) that the length of the projection of κ onto the line is constant. Equivalently, the inner product $\kappa{\cdot}\xi = cos(\angle(\kappa,\xi))\,\|\kappa\|\,\|\xi\|$ is constant.

That is how to think about the conservation of angular momentum for free particles in relativity theory. It does not matter that in the latter context we are dealing with a Lorentzian metric and allowing for curvature. The claim is still that a certain inner product of vector fields remains constant along a geodesic, and we can still think of that constancy as arising from a compensatory balance of two factors.

Let us now turn to the second type of conserved quantity, the one that is an attribute of extended bodies. Let κ^a be an arbitrary Killing field, and let T_{ab} be the energy-momentum field associated with some matter field. Assume it satisfies the conservation condition. Then $(T^{ab}\,\kappa_b)$ is divergence free:

$$\nabla_a(T^{ab}\kappa_b) \;=\; \kappa_b\,\nabla_a T^{ab} \,+\, T^{ab}\nabla_a\kappa_b \;=\; T^{ab}\nabla_{(a}\kappa_{b)} \;=\; \mathbf{0}. \qquad (41)$$

(The second equality follows from the conservation condition for T^{ab} (in section 2.4) and the symmetry of T^{ab}; the third from the fact that κ^a is a Killing field.) It is natural, then, to apply Stokes' theorem to the vector field $(T^{ab}\kappa_b)$.

Consider a bounded system with aggregate energy-momentum field T_{ab} in an otherwise empty universe. Then there exists a (possibly huge) timelike world tube such that T_{ab} vanishes outside the tube (and vanishes on its boundary). Let S_1 and S_2 be (non-intersecting) spacelike hypersurfaces that cut the tube

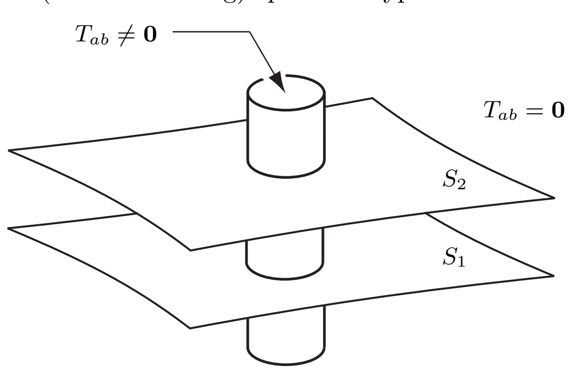

Figure 2.7.2: The integrated energy (relative to a background timelike Killing field) over the intersection of the world tube with a spacelike hypersurface is independent of the choice of hypersurface.

as in figure 2.7.2, and let N be the segment of the tube falling between them (with boundaries included). By Stokes' theorem,

$$\begin{aligned}\int_{S_2} (T^{ab}\kappa_b)\, dS_a - \int_{S_1} (T^{ab}\kappa_b)\, dS_a & \\ &= \int_{S_2 \cap\, \partial N} (T^{ab}\kappa_b)\, dS_a - \int_{S_1 \cap\, \partial N} (T^{ab}\kappa_b)\, dS_a \\ &= \int_{\partial N} (T^{ab}\kappa_b)\, dS_a = \int_N \nabla_a (T^{ab}\kappa_b)\, dV = 0.\end{aligned}$$

Thus, the integral $\int_S (T^{ab}\kappa_b)\, dS_a$ is independent of the choice of spacelike hypersurface S intersecting the world tube, and is, in this sense, a conserved quantity (construed as an attribute of the system confined to the tube). An "early" intersection yields the same value as a "late" one. Again, the character of the background Killing field κ^a determines our description of the conserved quantity in question. If κ^a is timelike, we take $\int_S (T^{ab}\kappa_b)\, dS_a$ to be the aggregate energy of the system (associated with κ^a). And so forth.

For further discussion of symmetry and conservation principles in general relativity, see Brading and Castellani (this volume, chapter 13).

3 Special Topics

3.1 Relative Simultaneity in Minkowski Spacetime

We noted in section 2.3, when discussing the decomposition of vectors at a point into their "temporal" and "spatial" components relative to a four-velocity vector there, that we were taking for granted the standard identification of relative simultaneity with orthogonality. Here we return to consider the justification of that identification.

Rather than continue to cast the discussion as one concerning the decomposition of the tangent space at a particular point, it is convenient to construe it instead as one about the structure of Minkowski spacetime, the regime of so-called "special relativity". Doing so will bring it closer to the framework in which traditional discussions of the status of the relative simultaneity relation have been conducted.

Minkowski spacetime is a relativistic spacetime (M, g_{ab}) characterized by three conditions: (i) M is the manifold $\mathbb{R}^4$; (ii) (M, g_{ab}) is flat, i.e., g_{ab} has vanishing Riemann curvature everywhere; and (iii) (M, g_{ab}) is geodesically complete, i.e., every geodesic (with respect to g_{ab}) can be extended to arbitrarily large parameter values in both directions.

By virtue of these conditions, Minkowski spacetime can be canonically identified with its tangent space at any point, and so it inherits the structure of a "metric affine space" in the following sense. Pick any point o in M, and let V be the tangent space M_o at o. Then there is a map $(p,q) \mapsto \overrightarrow{pq}$ from $M \times M$ to V with the following two properties.

(1) For all p, q and r in M, $\overrightarrow{pq} + \overrightarrow{qr} = \overrightarrow{pr}$.

(2) For all p in M, the induced map $q \mapsto \overrightarrow{pq}$ from M to V is a bijection.[43]

The triple consisting of the point set M, the vector space V, and the map $(p,q) \mapsto \overrightarrow{pq}$ forms an *affine space*. If we add to this triple the inner product on V defined by g_{ab} it becomes a (Lorentzian) *metric affine space*. (For convenience we will temporarily drop the index notation and write $\langle v, w \rangle$ instead of $g_{ab}v^a w^b$ for v and w in V.) We take all this structure for granted in what follows, i.e., we work with Minkowski spacetime and construe it as a metric affine space in the sense described. This will simplify the presentation considerably.

We also use an obvious notation for orthogonality. Given four points p, q, r, s in M, we write $\overrightarrow{pq} \perp \overrightarrow{rs}$ if $\langle \overrightarrow{pq}, \overrightarrow{rs} \rangle = 0$. And given a line[44] O in M, we write $\overrightarrow{pq} \perp O$ if $\overrightarrow{pq} \perp \overrightarrow{rs}$ for all points r and s on O.

Now consider a timelike line O in M. What pairs of points (p,q) in M should qualify as being "simultaneous relative to O"? That is the question we are considering. The standard answer is that they should do so precisely if $\overrightarrow{pq} \perp O$.

In traditional discussions of relative simultaneity, the standard answer is often cast in terms of "epsilon" values. The connection is easy to see. Let p be any point that is not on our timelike line O. Then there exist unique points r and s on O (distinct from one another) such that $\overrightarrow{rp}$ and $\overrightarrow{ps}$ are future-directed null vectors. (See figure 3.1.1.) Now let q be any point on O. (We think of it as a candidate for being judged simultaneous with p relative to O.) Then

[43] If exp is the exponential map from M_o to M, we can take $\overrightarrow{pq}$ to be the vector

$$(exp^{-1}(q) - exp^{-1}(p))$$

in M_o. All other standard properties of affine spaces follow from these two. E.g., it follows that $\overrightarrow{pq} = \mathbf{0} \iff p = q$, for all p and q in M. (Here $\mathbf{0}$ is the zero vector in V.)

[44] In the present context we can characterize a *line* in more than one way. We can take it to be the image of a maximally extended geodesic that is non-trivial, i.e., not a point. Equivalently, we can take it to be a set of points of the form $\{r : \overrightarrow{pr} = \epsilon\, \overrightarrow{pq} \text{ for some } \epsilon \text{ in } \mathbb{R}\}$ where p and q are any two (distinct) points in M.

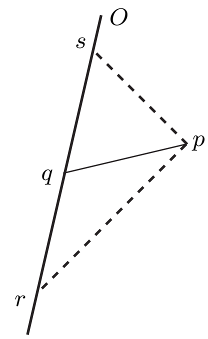

Figure 3.1.1: The $\epsilon = \frac{1}{2}$ characterization of relative simultaneity: p and q are simultaneous relative to O iff q is midway between r and s.

$\overrightarrow{rq} = \epsilon\,\overrightarrow{rs}$ for some $\epsilon \in \mathbb{R}$. A simple computation[45] shows that

$$\epsilon = \frac{1}{2} \iff \overrightarrow{pq} \perp \overrightarrow{rs}. \tag{42}$$

So the standard (orthogonality) relation of relative simultaneity in special relativity may equally well be described as the "$\epsilon = \frac{1}{2}$" relation of relative simultaneity.

Yet another equivalent formulation involves the "one-way speed of light". Suppose a light ray travels from r to p with speed c_+ relative to O, and from p to s with speed c_- relative to O. We saw in section 2.3 that *if* one adopts the standard criterion of relative simultaneity, then it follows that $c_+ = c_-$. The converse is true as well. For if $c_+ = c_-$, then, as determined relative to O, it should take as much time for light to travel from r to p as from p to s. And in that case, a point q on O should be judged simultaneous with p relative to O precisely if it is midway between r and s. So we are led, once again, to the "$\epsilon = \frac{1}{2}$" relation of relative simultaneity.

Now is adoption of the standard relation a matter of convention, or is it in some significant sense forced on us?

---

[45] First note that, since $\overrightarrow{ps}$ and $\overrightarrow{pr}$ are null,

$$0 = \langle \overrightarrow{ps}, \overrightarrow{ps} \rangle = \langle \overrightarrow{pr} + \overrightarrow{rs}, \overrightarrow{pr} + \overrightarrow{rs} \rangle = 2\langle \overrightarrow{pr}, \overrightarrow{rs} \rangle + \langle \overrightarrow{rs}, \overrightarrow{rs} \rangle.$$

It follows that

$$\langle \overrightarrow{pq}, \overrightarrow{rs} \rangle = \langle \overrightarrow{pr} + \overrightarrow{rq}, \overrightarrow{rs} \rangle = \langle \overrightarrow{pr} + \epsilon\overrightarrow{rs}, \overrightarrow{rs} \rangle = \langle \overrightarrow{pr}, \overrightarrow{rs} \rangle + \epsilon\langle \overrightarrow{rs}, \overrightarrow{rs} \rangle = (\epsilon - \frac{1}{2})\langle \overrightarrow{rs}, \overrightarrow{rs} \rangle,$$

which implies (42).

There is, of course, a large literature devoted to this question.[46] It is not my purpose to review it here, but I do want to draw attention to certain remarks of Howard Stein [1991, pp. 153-4] that seem to me particularly insightful. He makes the point that determinations of conventionality require a context.

> There are really two distinct aspects to the issue of the "conventionality" of Einstein's concept of relative simultaneity. One may assume the position of Einstein himself at the outset of his investigation – that is, of one confronted by a problem, trying to find a theory that will deal with it satisfactorily; or one may assume the position of (for instance) Minkowski – that is, of one confronted with a theory already developed, trying to find its most adequate and instructive formulation.

The problem Einstein confronted was (in part) that of trying to account for our apparent inability to detect any motion of the earth with respect to the "aether". A crucial element of his solution was the proposal that we think about simultaneity a certain way (i.e., in terms of the "$\epsilon = \frac{1}{2}$ criterion"), and resolutely follow through on the consequences of doing so. Stein emphasizes just how different that proposal looks when we consider it, not from Einstein's initial position, but rather from the vantage point of the finished theory, i.e., relativity theory conceived as an account of invariant spacetime structure.

> [For] Einstein, the question (much discussed since Reichenbach) whether the evidence really shows that that the speed of light *must* be regarded as the same in all directions and for all observers is not altogether appropriate. A person devising a theory does not have the responsibility, at the outset, of showing that the theory being developed is the only possible one given the evidence. [But] once Einstein's theory had been developed, and had proved successful in dealing with all relevant phenomena, the case was quite transformed; for we know that *within* this theory, there is only one "reasonable" concept of simultaneity (and in terms of that concept, the velocity of light is indeed as Einstein supposed); therefore an alternative will only present itself if someone succeeds in constructing, not simply a

[46] Classic statements of the conventionalist position can be found in Reichenbach [1958] and Grünbaum [1973]. Grünbaum has recently responded to criticism of his views in [forthcoming]. An overview of the debate with many references can be found in Janis [2002].

> different empirical criterion of simultaneity, but an essentially different (and yet viable) theory of electrodynamics of systems in motion. No serious alternative theory is in fact known. (emphasis in original)

My goal in the remainder of this section is to formulate three elementary uniqueness results, closely related to one another, that capture the sense in which "there is only one 'reasonable' concept of (relative) simultaneity" within the framework of Minkowski spacetime.

It will help to first consider an analogy. In some formulations of Euclidean plane geometry, the relation of congruence between angles is taken as primitive along with that of congruence between line segments (and other relations suitable for formulating axioms about affine structure). But suppose we have a formulation in which it is not, and we undertake to *define* a notion of angle-congruence in terms of the other primitives. The standard angle-congruence relation can certainly be defined this way, and there is a clear sense in which it is the only reasonable candidate. Consider any two angles in the Euclidean plane. (Let's agree that an "angle" consists of two rays, i.e., half-lines, with a common initial point.) Whatever else is the case, presumably, it is only reasonable to count them as congruent, i.e., equal in "size", if there is an isometry of the Euclidean plane that maps one angle onto the other.[47] So though we have here a notion of angle-congruence that is introduced "by definition", there is no interesting sense in which it is conventional in character.

A situation very much like this arises if we think about "one-way light speeds" in terms of Minkowskian spacetime geometry. Indeed, the claim that the speed of light *in vacuo* is the same in all directions and for all inertial observers is naturally represented as a claim about angle congruence (for a special type of angle) in Minkowski spacetime.

Let us take a "light-speed angle" to be a triple of the form (p, T, N), where p is a point in M, T is a future-pointing timelike ray with initial point p, and N is a future-pointing null ray with initial point p. (See figure 3.1.2.) Then we can represent systematic attributions of one-way light speed as maps of the form: $(p, T, N) \mapsto v(p, T, N)$. (We understand $v(p, T, N)$ to be the speed that an observer with (half) worldline T at p assigns to the light signal with (half) worldline N.) So, for example, the principle that the speed of light is the same in all directions and for all inertial observers comes out as the condition that $v(p, T, N) = v(p', T', N')$ for all light-speed angles (p, T, N) and (p', T', N').

[47] In this context, a one-to-one map of the Euclidean plane onto itself is an "isometry" if it preserves the relation of congruence between line segments.

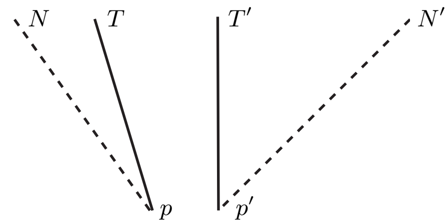

Figure 3.1.2: Congruent "light speed angles" in Minkowski spacetime.

Now it is natural to regard $v(p, T, N)$ as a measure of the "size" of the angle (p, T, N). If we do so, then, just as in the Euclidean case, we can look to the background metric to decide when two angles have the same size. That is, we can take them to be congruent iff there is an isometry of Minkowski spacetime that maps one to the other. But on this criterion, all light-speed angles are congruent (proposition 3.1.1). So we are led back to the principle that the (one-way) speed of light is the same in all directions and for all inertial observers and, hence, back to the standard relative simultaneity relation.

Proposition 3.1.1. *Let (p, T, N) and (p', T', N') be any two light speed angles in Minkowski spacetime. Then there is an isometry ϕ of Minkowski spacetime such that $\phi(p) = p'$, $\phi[T] = T'$, and $\phi[N] = N'$.*[48]

Once again, let O be a timelike line in M, and let Sim_O be the standard relation of simultaneity relative to O. (So $(p, q) \in Sim_O$ iff $\overrightarrow{pq} \perp O$, for all p and q in M.) Further, let S be an arbitrary two-place relation on M that we regard as a candidate for the relation of "simultaneity relative to O". Our second uniqueness result asserts that if S satisfies three conditions, including an invariance condition, then $S = Sim_O$.[49]

The first two conditions are straightforward.

S1 S is an equivalence relation (i.e., S is reflexive, symmetric, and transitive).

S2 For all points $p \in M$, there is a unique point $q \in O$ such that $(p, q) \in S$.

[48] The required isometry can be realized in the form $\phi = \phi_3 \circ \phi_2 \circ \phi_1$ where (i) ϕ_1 is a translation that takes p to p', (ii) ϕ_2 is a boost (based at p') that maps $\phi_1[T]$ to T', and (iii) ϕ_3 is a rotation about T' that maps $(\phi_2 \circ \phi_1)[N]$ to N'.

[49] Many propositions of this form can be found in the literature. (See Budden [1998] for a review.) Ours is intended only as an example. There are lots of possibilities here depending on exactly how one formulates the conditions that S must satisfy. The proofs are all very much the same.

If S satisfies (S1), it has an associated family of equivalence classes. We can think of them as "simultaneity slices" (as determined relative to O). Then (S2) asserts that every simultaneity slice intersects O in exactly one point.

The third condition is intended to capture the requirement that S is determined by the background geometric structure of Minkowski spacetime and by O itself. But there is one subtle point here. It makes a difference whether temporal orientation counts as part of that background geometric structure or not. Let's assume for the moment that it does not.

Let us say, quite generally, that S is *implicitly definable* in a structure of the form $(M, ...)$ if it is invariant under all symmetries of $(M, ...)$, i.e., for all such symmetries $\phi : M \to M$, and all points p and q in M,

$$(p, q) \in S \iff (\phi(p), \phi(q)) \in S. \tag{43}$$

(Here, a *symmetry* of $(M, ...)$ is understood to be a bijection $\phi : M \to M$ that preserves all the structure in "...".) This is certainly a very weak sense of definability.[50]

We are presently interested in the structure $(M, \langle\ \rangle, O)$.[51] Its symmetries are bijections $\phi : M \to M$ such that, for all p, q, r, s in M,

$$\langle \overrightarrow{\phi(p)\,\phi(q)},\ \overrightarrow{\phi(r)\,\phi(s)} \rangle = \langle \overrightarrow{pq},\ \overrightarrow{rs} \rangle \tag{44}$$

and

$$p \in O \iff \phi(p) \in O. \tag{45}$$

They are generated by maps of the following three types: (i) translations ("up and down") in the direction of O, (ii) spatial rotations that leave fixed every point in O, and (iii) temporal reflections with respect to spacelike hyperplanes orthogonal to O. Our second uniqueness result can be formulated as follows.[52]

Proposition 3.1.2. *Let S be a two-place relation on Minkowski spacetime that satisfies conditions (S1), (S2), and is implicitly definable in $(M, \langle\ \rangle, O)$. Then $S = Sim_O$.*

[50] We follow Budden [1998] in using "implicit definability in $(M, ...)$" as a convenient device for organizing various closely related uniqueness results. One moves from one to another simply by shifting the choice of "..." .

[51] Strictly speaking, since we are here thinking of Minkowski spacetime as a metric affine space, we should include elements of structure (to the right of M) that turn the point set M into an affine space, i.e., add a four-dimensional vector space V and a map from $M \times M$ to V satisfying conditions (1) and (2) listed at beginning of this section. But it is cumbersome to do so every time. Let them be understood in what follows.

[52] It is a close variant of one presented in Hogarth [1993].

As it turns out, the full strength of the stated invariance condition is not needed here. It suffices to require that S be invariant under maps of type (iii).[53]

Suppose now that we *do* want to consider temporal orientation as part of the background structure that may play a role in the determination of S. Then the requirement of implicit definability in $(M, \langle\ \rangle, O)$ should be replaced by that of implicit definability in $(M, \mathcal{T}, \langle\ \rangle, O)$, where $\mathcal{T}$ is the background temporal orientation. Maps of type (i) and (ii) qualify as symmetries of this enriched structure (too), but temporal reflections of type (iii) do not. Implicit definablity in $(M, \mathcal{T}, \langle\ \rangle, O)$ is a weaker condition than implicit definability in $(M, \langle\ \rangle, O)$, and proposition 3.1.2 fails if the latter condition is replaced by the former.

But we can still get a uniqueness result if we change the set-up slightly: namely, we think of O as merely one member of a congruence of parallel timelike lines $\mathcal{F}_O$, the *frame* of O, and think of simultaneity as determined relative to the latter. Implicit definability in $(M, \mathcal{T}, \langle\ \rangle, \mathcal{F}_O)$ is a stronger condition than implicit definability in $(M, \mathcal{T}, \langle\ \rangle, O)$ because there are symmetries of the former – (iv) translations taking O to other lines in $\mathcal{F}_O$, and (v) spatial rotations that leave fixed the points of some line in $\mathcal{F}_O$ other than O – that are not symmetries of the latter. Our variant formulation of the uniqueness result is the following.[54]

Proposition 3.1.3. *Let S be a two-place relation on Minkowski spacetime that satisfies conditions (S1), (S2), and is implicitly definable in $(M, \mathcal{T}, \langle\ \rangle, \mathcal{F}_O)$. Then $S = Sim_O$.*

The move from proposition 3.1.2 to proposition 3.1.3 involves a trade-off. We drop the requirement that S be invariant under maps of type (iii), but add the requirement that it be invariant under those of type (iv) and (v).[55]

Once again, many variations of these results can be found in the literature. For example, if one subscribes to a "causal theory of time (or spacetime)", one

[53]The key step in the proof is the following. Let p be a point in M. By (S2), there is a unique point q on O such that $\langle p, q\rangle \in S$. Let $\phi : M \to M$ be a reflection with respect to the hyperplane orthogonal to O that passes through p. Then $\phi(p) = p$, $\phi(q) \in O$, and S is invariant under ϕ. Hence $\langle p, \phi(q)\rangle = \langle \phi(p), \phi(q)\rangle \in S$. Since $\phi(q) \in O$, it follows by the uniqueness condition in (S2) that $\phi(q) = q$. But the only points left fixed by ϕ are those on the hyperplane orthogonal to O that passes through p. So p and q are both on that hyperplane, and $\overrightarrow{pq}$ is orthogonal to O, i.e., $\langle p, q\rangle \in Sim_O$.

[54]It is closely related to propositions in Spirtes [1981], Stein [1991], and Budden [1998].

[55]It is a good exercise to check that one does not need the full strength of the stated invariance condition to derive the conclusion. It suffices to require that S be invariant under maps of type (i) and (v), or, alternatively, invariant under maps of type (ii) and (iv).

will want to consider what candidate simultaneity relations are determined by the causal structure of Minkowski spacetime (in adddition to the line O). Let κ be the symmetric two-place relation of "causal connectibility" in M, i.e., the relation that holds of two points p and q if $\overrightarrow{pq}$ is a causal vector. Clearly, every symmetry of $(M, \langle\ \rangle)$ is a symmetry of (M, κ). So the requirement of implicit definability in $(M,\ \kappa,\ O)$ is at least as strong as that of implicit definability in $(M,\ \langle\ \rangle,\ O)$. It follows that we can substitute the former for the latter in proposition 3.1.2. Similarly, we can substitute the requirement of implicit definability in $(M,\ \mathcal{T},\ \kappa,\ \mathcal{F}_O)$ for that of implicit definability in $(M,\ \mathcal{T},\ \langle\ \rangle,\ \mathcal{F}_O)$ in proposition 3.1.3.

3.2 Geometrized Newtonian Gravitation Theory

The "geometrized" formulation of Newtonian gravitation theory was first introduced by Cartan [1923; 1924], and Friedrichs [1927], and later developed by Trautman [1965], Künzle [1972; 1976], Ehlers [1981], and others.

It is significant for several reasons. (1) It shows that several features of relativity theory once thought to be uniquely characteristic of it do not distinguish it from (a suitably reformulated version of) Newtonian gravitation theory. The latter too can be cast as a "generally covariant" theory in which (a) gravity emerges as a manifestation of spacetime curvature, and (b) spacetime structure is "dynamical", i.e., participates in the unfolding of physics rather than being a fixed backdrop against which it unfolds.

(2) It helps one to see where Einstein's equation "comes from", at least in the empty-space case. (Recall the discussion in section 2.5.) It also allows one to make precise, in coordinate-free, geometric language, the standard claim that Newtonian gravitation theory (or, at least, a certain generalized version of it) is the "classical limit" of general relativity. (See Künzle [1976] and Ehlers [1981].)

(3) It clarifies the gauge status of the Newtonian gravitational potential. In the geometrized formulation of Newtonian theory, one works with a single curved derivative operator $\overset{g}{\nabla}_a$. It can be decomposed (in a sense) into two pieces – a flat derivative operator ∇_a and a gravitational potential ϕ – to recover the standard formulation of the theory.[56] But in the absence of special boundary conditions, the decomposition will not be unique. Physically, there is no unique way to

[56] As understood here, the "standard" formulation is not that found in undergraduate textbooks, but rather a "generally covariant" theory of four-dimensional spacetime structure in which gravity is not geometrized.

divide into "inertial" and "gravitational" components the forces experienced by particles. Neither has any direct physical significance. Only their "sum" does. It is an attractive feature of the geometrized formulation that it trades in two gauge quantities for this sum.

(4) The clarification described in (3) also leads to a solution, or dissolution, of an old conceptual problem about Newtonian gravitation theory, namely the apparent breakdown of the theory when applied (in cosmology) to a hypothetically infinite, homogeneous mass distribution. (See Malament [1995] and Norton [1995; 1999].)

In what follows, we give a brief overview of the geometrized formulation of Newtonian gravitation theory, and say a bit more about points (1) and (3). We start by characterizing a new class of geometrical models for the spacetime structure of our universe (or subregions thereof) that is broad enough to include the models considered in both the standard and geometrized versions of Newtonian theory. We take a *classical spacetime* to be a structure $(M, t_{ab}, h^{ab}, \nabla_a)$ where (i) M is a smooth, connected, four-dimensional differentiable manifold; (ii) t_{ab} is a smooth, symmetric, covariant tensor field on M of signature (1, 0, 0, 0);[57] (iii) h^{ab} is a smooth, symmetric, contravariant tensor field on M of signature $(0, 1, 1, 1)$; (iv) ∇_a is a smooth derivative operator on M; and (v) the following two conditions are met:

$$h^{ab}\, t_{bc} \;=\; \mathbf{0} \tag{46}$$

$$\nabla_a\, t_{bc} \;=\; \mathbf{0} \;=\; \nabla_a\, h^{bc}. \tag{47}$$

We refer to them, respectively, as the "orthogonality" and "compatibility" conditions.

M is interpreted as the manifold of point events (as before); t_{ab} and h^{ab} are understood to be temporal and spatial metrics on M, respectively; and ∇_a is

[57]The stated signature condition is equivalent to the requirement that, at every point p in M, the tangent space M_p has a basis $\overset{1}{\xi}{}^a, \ldots, \overset{4}{\xi}{}^a$ such that, for all i and j in $\{1, 2, 3, 4\}$,

$$t_{ab}\, \overset{i}{\xi}{}^a \overset{i}{\xi}{}^b = \begin{cases} 1 & \text{if } i = 1 \\ 0 & \text{if } i = 2,\, 3,\, 4 \end{cases}$$

and $t_{ab}\, \overset{i}{\xi}{}^a \overset{j}{\xi}{}^b = 0$ if $i \neq j$. Similarly, the signature condition on h^{ab} stated in (iii) requires that, at every point, the co-tangent space have a basis $\overset{1}{\sigma}_a, \ldots, \overset{4}{\sigma}_a$ such that, for all i and j in $\{1, 2, 3, 4\}$,

$$h^{ab}\, \overset{i}{\sigma}_a \overset{i}{\sigma}_b = \begin{cases} 0 & \text{if } i = 1 \\ 1 & \text{if } i = 2,\, 3,\, 4 \end{cases}$$

and $h^{ab}\, \overset{i}{\sigma}_a \overset{j}{\sigma}_b = 0$ if $i \neq j$.

understood to be an affine structure on M. Collectively, these objects represent the spacetime structure presupposed by classical, Galilean relativistic dynamics. We review, briefly, how they do so.

In what follows, let $(M, t_{ab}, h^{ab}, \nabla_a)$ be a fixed classical spacetime.

Consider, first, t_{ab}. Given any vector ξ^a at a point, it assigns a "temporal length" $(t_{ab}\, \xi^a\, \xi^b)^{\frac{1}{2}} \geq 0$. The vector ξ^a is classified as *timelike* or *spacelike* depending on whether its temporal length is positive or zero. It follows from the signature of t_{ab} that the subspace of spacelike vectors at any point is three-dimensional. It also follows from the signature that at every point there exists a covariant vector t_a, unique up to sign, such that $t_{ab} = t_a t_b$. We say that the structure $(M, t_{ab}, h^{ab}, \nabla_a)$ is *temporally orientable* if there is a continuous (globally defined) vector field t_a such that this decomposition holds at every point. Each such field t_a (which, in fact, must be smooth because t_{ab} is) is a *temporal orientation*. A timelike vector ξ^a qualifies as *future-directed* relative to t_a if $t_a\, \xi^a > 0$; otherwise it is *past-directed*. Let us assume in what follows that $(M, t_{ab}, h^{ab}, \nabla_a)$ is temporally orientable and that a temporal orientation t_a has been selected.

From the compatibility condition, it follows that t_a is closed, i.e., $\nabla_{[a}\, t_{b]} = \mathbf{0}$. So, at least locally, it must be exact, i.e., of form $t_a = \nabla_a\, t$ for some smooth function t. We call any such function a *time function*. If M has a suitable global structure, e.g., if it is simply connected, then a globally defined time function $t : M \rightarrow \mathbb{R}$ must exist. In this case, spacetime can be decomposed into a one-parameter family of global ($t =$ constant) "time slices". One can speak of "space" at a given "time". A different choice of time function would result in a different zero-point for the time scale, but would induce the same time slices and the same elapsed intervals between them.

We say that a smooth curve is *timelike* (respectively *spacelike*) if its tangent field is timelike (respectively spacelike) at every point. In what follows, unless indication is given to the contrary, it should further be understood that a "timelike curve" is future-directed and parametrized by its t_{ab} - length. In this case, its tangent field ξ^a satisfies the normalization condition $t_a \xi^a = 1$. Also, in this case, if a particle happens to have the image of the curve as its worldline, then, at any point, ξ^a is called the particle's *four-velocity*, and $\xi^n \nabla_n\, \xi^a$ its *four-acceleration*, there.[58] If the particle has mass m, then its four-acceleration

[58] Here we take for granted an interpretive principle that corresponds to $C1$: (i) a curve is timelike iff its image could be the worldline of a point particle. Other principles we can formulate at this stage correspond to $P1$ and $P2$: (ii) a timelike curve can be reparametrized

field satisfies the equation of motion

$$F^a = m\, \xi^n \nabla_n\, \xi^a, \tag{48}$$

where F^a is a spacelike vector field (on the image of its worldline) that represents the net force acting on the particle. This is, once again, our version of Newton's second law of motion. Recall (15). Note that the equation makes geometric sense because four-acceleration vectors are necessarily spacelike.[59]

Now consider h^{ab}. It serves as a spatial metric, but just how it does so is a bit tricky. In Galilean relativistic mechanics, we have no notion of spatial length for timelike vectors, e.g., four-velocity vectors, since having one is tantamount to a notion of absolute rest. (We can take a particle to be at rest if its four-velocity has spatial length 0 everywhere.) But we *do* have a notion of spatial length for spacelike vectors, e.g., four-acceleration vectors. (We can, for example, use measuring rods to determine distances between simultaneous events.) h^{ab} serves to give us one without the other.

We cannot take the spatial length of a vector σ^a to be $(h_{ab}\, \sigma^a \sigma^b)^{\frac{1}{2}}$ because the latter is not well-defined. (Since h^{ab} has degenerate signature, it is not invertible, i.e., there does not *exist* a field h_{ab} satisfying $h^{ab} h_{bc} = \delta^a{}_c$.) But if σ^a is spacelike, we can use h^{ab} to assign a spatial length to it indirectly. It turns out that: (i) a vector σ^a is spacelike iff it can be expressed in the form $\sigma^a = h^{ab}\, \lambda_b$, and (ii) if it can be so expressed, the quantity $(h^{ab}\, \lambda_a\, \lambda_b)$ is independent of the choice of λ_a. Furthermore, the signature of h^{ab} guarantees that $(h^{ab}\, \lambda_a\, \lambda_b) \geq 0$. So if σ^a is spacelike, we can take its spatial length to be $(h^{ab}\, \lambda_a\, \lambda_b)^{\frac{1}{2}}$, for any choice of corresponding λ_a.

One final preliminary remark about classical spacetimes is needed. It is crucial for our purposes, as will be clear, that the compatibility condition (47) does not determine a unique derivative operator. (It is a fundamental result that the compatibility condition $\nabla_a\, g_{bc} = \mathbf{0}$ determines a unique derivative operator if g_{ab} is a semi-Riemannian metric, i.e., a smooth, symmetric field that is invertible (i.e., non-degenerate). But neither t_{ab} nor h^{ab} is invertible.)

Because h^{ab} is not invertible, we cannot raise *and* lower indices with it. But we can, at least, raise indices with it, and it is sometimes convenient to do so.

---

so as to be a geodesic (with respect to ∇_a) iff its image could be the worldline of a free particle; (iii) clocks record the passage of elapsed t_{ab} - length along their worldlines. (Here, in contrast the relativistic setting, we have only massive particles to consider and, until we geometrize Newtonian gravity, do not count a particle as "free" if it is subject to "gravitational force".)

[59] By the compatibility condition, $t_a\, \xi^n \nabla_n\, \xi^a = \xi^n \nabla_n\, (t_a\, \xi^a) = \xi^n \nabla_n\, (1) = 0$.

So, for example, if $R^a{}_{bcd}$ is the Riemann curvature tensor field associated with ∇_a, we can understand $R^{ab}{}_{cd}$ to be an abbreviation for $h^{bn}\, R^a{}_{ncd}$.

Let us now, finally, consider Newtonian gravitation theory. In the standard (non-geometrized) version, one works with a flat derivative operator ∇_a and a gravitational potential ϕ, the latter understood to be a smooth, real-valued function on M. The gravitational force on a point particle with mass m is given by $-\, m\, h^{ab}\, \nabla_b\, \phi$. (Notice that this is a spacelike vector by the orthogonality condition.) Using our convention for raising indices, we can also express the vector as: $-\, m\, \nabla^a\, \phi$. It follows that if the particle is subject to no forces except gravity, and if it has four-velocity ξ^a, it satisfies the equation of motion

$$-\nabla^a\, \phi \;=\; \xi^n\, \nabla_n\, \xi^a. \tag{49}$$

(Here we have just used $-\, m\, \nabla^a\, \phi$ for the left side of (48).) It is also assumed that ϕ satisfies Poisson's equation:

$$\nabla^a\, \nabla_a\, \phi \;=\; 4\, \pi\, \rho, \tag{50}$$

where ρ is the Newtonian mass-density function (another smooth real-valued function on M). (The expression on the left side is an abbreviation for: $h^{ab}\, \nabla_b\, \nabla_a\, \phi$.)

In the geometrized formulation of the theory, gravitation is no longer conceived as a fundamental "force" in the world, but rather as a manifestation of spacetime curvature (just as in relativity theory). Rather than thinking of point particles as being deflected from their natural straight (i.e., geodesic) trajectories, one thinks of them as traversing geodesics in curved spacetime. So we have a geometry problem. Starting with the structure $(M, t_{ab}, h^{ab}, \nabla_a)$, can we find a new derivative operator $\overset{g}{\nabla}_a$, also compatible with the metrics t_{ab} and h^{ab}, such that a timelike curve satisfies the equation of motion (49) with respect to the original derivative operator ∇_a iff it is a geodesic with respect to $\overset{g}{\nabla}_a$? The following proposition (essentially due to Trautman [1965]) asserts that there is exactly one such $\overset{g}{\nabla}_a$. It also records several facts about the Riemann curvature tensor field $\overset{g}{R}{}^a{}_{bcd}$ associated with $\overset{g}{\nabla}_a$.

In formulating the proposition, we make use of the following basic fact about derivative operators. Given any two such operators $\overset{1}{\nabla}_a$ and $\overset{2}{\nabla}_a$ on M, there is a unique smooth tensor field $C^a{}_{bc}$, symmetric in its covariant indices, such that, for all smooth fields $\alpha^{a\ldots b}{}_{c\ldots d}$ on M,

$$\overset{2}{\nabla}_n \, \alpha^{a\ldots b}{}_{c\ldots d} \;=\; \overset{1}{\nabla}_n \, \alpha^{a\ldots b}{}_{c\ldots d} + C^r{}_{nc}\, \alpha^{a\ldots b}{}_{r\ldots d} + \ldots + C^r{}_{nd}\, \alpha^{a\ldots b}{}_{c\ldots r} - C^a{}_{nr}\, \alpha^{r\ldots b}{}_{c\ldots d} - \ldots - C^b{}_{nr}\, \alpha^{a\ldots r}{}_{c\ldots d}. \tag{51}$$

In this case, we say that "the action of $\overset{2}{\nabla}_a$ relative to that of $\overset{1}{\nabla}_a$ is given by $C^a{}_{bc}$".[60] Conversely, given any one derivative operator $\overset{1}{\nabla}_a$ on M, and any smooth, symmetric field $C^a{}_{bc}$on M, (51) defines a new derivative operator $\overset{2}{\nabla}_a$ on M. (See Wald [1984, p. 33].)

Proposition 3.2.1 (Geometrization Theorem)**.** *Let $(M, t_{ab}, h^{ab}, \nabla_a)$ be a classical spacetime with ∇_a flat ($R^a{}_{bcd} = \mathbf{0}$). Further, let ϕ and ρ be smooth real valued functions on M satisfying Poisson's equation: $\nabla^a \nabla_a \phi = 4\pi\rho$. Finally, let $\overset{g}{\nabla}_a$ be the derivative operator on M whose action relative to that of ∇_a is given by $C^a{}_{bc} = -t_{bc}\nabla^a\phi$. Then all the following hold.*

(G1) $(M, t_{ab}, h^{ab}, \overset{g}{\nabla}_a)$ is a classical spacetime.

(G2) $\overset{g}{\nabla}_a$ is the unique derivative operator on M such that, for all timelike curves on M with four-velocity fields ξ^a,

$$\xi^n \, \overset{g}{\nabla}_n \, \xi^a \;=\; \mathbf{0} \iff -\nabla^a\phi \;=\; \xi^n \, \nabla_n \, \xi^a. \tag{52}$$

(G3) The curvature field $\overset{g}{R}{}^a{}_{bcd}$ associated with $\overset{g}{\nabla}_a$ satisfies:

$$\overset{g}{R}_{bc} \;=\; 4\pi\rho\, t_{bc} \tag{53}$$

$$\overset{g}{R}{}^{ab}{}_{cd} \;=\; \mathbf{0} \tag{54}$$

$$\overset{g}{R}{}^{[a}{}_{(b}{}^{c]}{}_{d)} \;=\; \mathbf{0}. \tag{55}$$

(53) is the geometrized version of Poisson's equation. The proof proceeds by more-or-less straight forward computation using (51).[61]

[60] Clearly, if the action of $\overset{2}{\nabla}_a$ relative to that of $\overset{1}{\nabla}_a$ is given by $C^a{}_{bc}$, then, conversely, the action of $\overset{1}{\nabla}_a$ relative to that of $\overset{2}{\nabla}_a$ is given by $-C^a{}_{bc}$. In the sum on the right side of (51), there is one term involving $C^a{}_{bc}$ for each index in $\alpha^{a\ldots b}{}_{c\ldots d}$. In each case, the index in question is contracted with $C^a{}_{bc}$, and the term carries a coefficient of $+1$ or -1 depending on whether the index in question is in covariant (down) or contravariant (up) position. (The components of $C^a{}_{bc}$ in a particular coordinate system are obtained by subtracting the Christoffel symbols associated with $\overset{1}{\nabla}_a$ (in that coordinate system) from those associated with $\overset{2}{\nabla}_a$.)

[61] Here is a sketch. By (51),

$$\overset{g}{\nabla}_a \, t_{bc} = \nabla_a \, t_{bc} + C^r{}_{ab}\, t_{rc} + C^r{}_{ac}\, t_{br} = \nabla_a \, t_{bc} + (-t_{ab}\nabla^r\phi)\, t_{rc} + (-t_{ac}\nabla^r\phi)\, t_{br}.$$

We can also work in the opposite direction. In geometrized Newtonian gravitation theory, one *starts* with a curved derivative operator $\overset{g}{\nabla}_a$ satisfying (53), (54), (55), and with the principle that point particles subject to no forces (except "gravity") traverse geodesics with respect to $\overset{g}{\nabla}_a$. (54) and (55) function as integrability conditions that ensure the possibility of working backwards to recover the standard formulation in terms of a gravitational potential and flat derivative operator.[62] We have the following recovery, or de-geometrization, theorem (also essentially due to Trautman [1965]).

The first term in the far right sum vanishes by the compatibility condition (47); the second and third do so by the orthogonality condition (46) since, for example, $(\nabla^r \phi)\, t_{br} = (h^{rm}\, t_{br})\, \nabla_m \phi$. So $\overset{g}{\nabla}_a$ is compatible with t_{bc}. Much the same argument shows that it is also compatible with h^{ab}. This give us (G1).

For (G2), let $\overset{g}{\nabla}_a$ (temporarily) be an arbitrary derivative operator on M whose action relative to that ∇_a is given by some field $C^a{}_{bc}$. Let p be an arbitrary point in M, and let ξ^a be the four-velocity field of an arbitrary timelike curve through p. Then, by (51),

$$\xi^n \, \overset{g}{\nabla}_n \, \xi^a \;=\; \xi^n \nabla_n \, \xi^a - C^a{}_{rn}\, \xi^r \xi^n .$$

It follows that $\overset{g}{\nabla}_n$ will satisfy (G2) iff $C^a{}_{rn} \xi^r \xi^n = -\nabla^a \phi$ or, equivalently,

$$[C^a{}_{rn} + (\nabla^a \phi)\, t_{rn}]\, \xi^r \xi^n = \mathbf{0},$$

for all future-directed unit timelike vectors ξ^a at all points p. But the space of future-directed unit timelike vectors at any p spans the tangent space M_p there, and the field in brackets is symmetric in its covariant indices. So, $\overset{g}{\nabla}_n$ will satisfy (G2) iff $C^a{}_{rn} \;=\; -(\nabla^a \phi)\, t_{rn}$ everywhere.

Finally, for (G3) we use the fact that $\overset{g}{R}{}^a{}_{bcd}$ can be expressed as a sum of terms involving $R^a{}_{bcd}$ and $C^a{}_{bc}$ (see Wald [1984, p. 184]), and then substitute for $C^a{}_{bc}$:

$$\begin{aligned} \overset{g}{R}{}^a{}_{bcd} \;&=\; R^a{}_{bcd} + 2\, \nabla_{[c}\, C^a{}_{d]b} + 2\, C^n{}_{b[c} C^a{}_{d]n} \\ &=\; R^a{}_{bcd} - 2\, t_{b[d} \nabla_{c]}\, \nabla^a \phi \;=\; -\, 2\, t_{b[d} \nabla_{c]}\, \nabla^a \phi . \end{aligned}$$

(Here $C^n{}_{b[c} C^a{}_{d]n}$ turns out to be $\mathbf{0}$ by the orthogonality condition, and $\nabla_{[c}\, C^a{}_{d]b}$ turns out to be $-t_{b[d} \nabla_{c]}\, \nabla^a \phi$ by the compatibility condition. For the final equality we use our assumption that $R^a{}_{bcd} = \mathbf{0}$.) (54) and (55) now follow from the orthogonality condition and (for (55)) from the fact that $\nabla^{[c} \nabla^{a]} \phi = \mathbf{0}$ for *any* smooth function ϕ. Contraction on 'a' and 'd' yields

$$\overset{g}{R}_{bc} = t_{bc} (\nabla_a \nabla^a \phi).$$

So (53) follows from our assumption that $\nabla^a \, \nabla_a \, \phi \;=\; 4\, \pi\, \rho$ (and the fact that $\nabla_a \, \nabla^a \phi \;=\; \nabla^a \, \nabla_a \phi$).

[62] I am deliberately passing over some subtleties here. Geometrized Newtonian gravitation theory comes in several variant formulations. (See Bain [2004] for a careful review of the differences.) The one presented here is essentially that of Trautman [1965]. In other weaker formulations (such as that in Künzle [1972]), condition (54) is dropped, and it is *not* possible to fully work back to the standard formulation (in terms of a gravitational potential and flat derivative operator) unless special global conditions on spacetime structure are satisfied.

Proposition 3.2.2 (Recovery Theorem)**.** *Let $(M, t_{ab}, h^{ab}, \overset{g}{\nabla}_a)$ be a classical spacetime that, together with a smooth, real-valued function ρ on M, satisfies conditions (53), (54), (55). Then, at least locally (and globally if M is, for example, simply connected), there exists a smooth, real-valued function ϕ on M and a flat derivative operator ∇_a such that all the following hold.*

(R1) $(M, t_{ab}, h^{ab}, \nabla_a)$ is a classical spacetime.

(R2) For all timelike curves on M with four-velocity fields ξ^a, the geometrization condition (52) is satisfied.

(R3) ∇_a satisfies Poisson's equation: $\nabla^a \nabla_a \phi = 4\pi\rho$.

The theorem is an existential assertion of this form: given $\overset{g}{\nabla}_a$ satisfying certain conditions, there exists (at least locally) a smooth function ϕ on M and a flat derivative operator ∇_a such that $\overset{g}{\nabla}_a$ arises as the "geometrization" of the pair (∇_a, ϕ). But, as claimed at the beginning of this section, we do *not* have uniqueness unless special boundary conditions are imposed on ϕ.

For suppose ∇_a is flat, and the pair (∇_a, ϕ) satisfies (R1), (R2), (R3). Let ψ be any smooth function (with the same domain as ϕ) such that $\nabla^a \nabla^b \psi$ vanishes everywhere, but $\nabla^b \psi$ does not.[63] If we set $\overline{\phi} = \phi + \psi$, and take $\overline{\nabla}_a$ to be the derivative operator relative to which the action of $\overset{g}{\nabla}_a$ is given by $\overline{C}^a{}_{bc} = -t_{bc}\nabla^a\overline{\phi}$, then $\overline{\nabla}_a$ is flat and the pair $(\overline{\nabla}_a, \overline{\phi})$ satisfies conditions (R1), (R2), (R3) as well.[64]

---

[63]We can think of $\nabla^b\psi$ as the "spatial gradient" of ψ. The stated conditions impose the requirement that $\nabla^b\psi$ be constant on all spacelike submanifolds ("time slices"), but not vanish on all of them.

[64]It follows directly from the way $\overline{\nabla}_a$ was defined that the pair $(\overline{\nabla}_a, \overline{\phi})$ satisfies conditions (R1) and (R2). (The argument is almost exactly the same as that used in an earlier note to prove (G1) and (G2) in the Geometrization Theorem.) What must be shown that is that $\overline{\nabla}_a$ is flat, and that the pair $(\overline{\nabla}_a, \overline{\phi})$ satisfies Poisson's equation: $\overline{\nabla}^a\,\overline{\nabla}_a\,\overline{\phi} = 4\pi\rho$. We do so by showing that (i) $\overline{R}^a{}_{bcd} = R^a{}_{bcd}$, (ii) $\nabla^a\nabla_a\psi = \mathbf{0}$, and (iii) $\overline{\nabla}^a\,\overline{\nabla}_a\,\alpha = \nabla^a\nabla_a\alpha$, for *all* smooth scalar fields α on M. (It follow immediately from (ii) and (iii) that $\overline{\nabla}^a\,\overline{\nabla}_a\,\overline{\phi} = \overline{\nabla}^a\,\overline{\nabla}_a\,\phi + \overline{\nabla}^a\,\overline{\nabla}_a\,\psi = \nabla^a\nabla_a\phi + \nabla^a\nabla_a\psi = 4\pi\rho$.)

We know from the uniqueness clause of (G2) in the Geometrization Theorem that the action of $\overset{g}{\nabla}_a$ with respect to ∇_a is given by the field $C^a{}_{bc} = -t_{bc}\nabla^a\phi$. It follows that the action of $\overline{\nabla}_a$ relative to that of ∇_a is given by $\widehat{C}^a{}_{bc} = -\overline{C}^a{}_{bc} + C^a{}_{bc} = -t_{bc}\nabla^a(-\overline{\phi} + \phi) = t_{bc}\nabla^a\psi$. So, arguing almost exactly as we did in the proof of (G3) in the Geometrization Theorem, we have

$$\overline{R}^a{}_{bcd} = R^a{}_{bcd} + 2\,t_{b[d}\nabla_{c]}\nabla^a\psi. \tag{56}$$

Now it follows from $\nabla^a\nabla^b\psi = \mathbf{0}$ that

$$\nabla_c\nabla^a\psi = t_c\,(\xi^n\nabla_n\nabla^a\psi), \tag{57}$$

where $t_{ab} = t_a t_b$, and ξ^n is any smooth future-directed unit timelike vector field on M.

$$\begin{tikzcd}
(\nabla_a, \phi) \arrow[dr] & \\
& \overset{g}{\nabla}_a \\
(\overline{\nabla}_a, \overline{\phi}) \arrow[ur] &
\end{tikzcd}$$

But, because $\nabla^b \psi$ is non-vanishing (somewhere or other), the pairs (∇_a, ϕ) and $(\overline{\nabla}_a, \overline{\phi})$ are distinct decompositions of $\overset{g}{\nabla}_a$. Relative to the first, a point particle (with mass m and four-velocity ξ^a) has acceleration $\xi^n \nabla_n \xi^a$ and is subject to a gravitational force $-m \nabla^a \phi$. Relative to the second, it has acceleration $\xi^n \overline{\nabla}_r \xi^a = \xi^n \nabla_n \xi^a - \nabla^a \psi$ and is subject to a gravitational force $-m \overline{\nabla}^a \overline{\phi} = -m \nabla^a \phi - m \nabla^a \psi$.

As suggested at the beginning of the section, we can take this non uniqueness of recovery result to capture in precise mathematical language the standard claim that Newtonian gravitational force is a gauge quantity. By the argument just given, if we can take the force on a point particle with mass m to be $-m \nabla^a \phi$, we can equally well take it to be $-m \nabla^a (\phi + \psi)$, where ψ is any field satisfying $\nabla^a \nabla^b \psi = \mathbf{0}$.

3.3 Recovering Global Geometric Structure from "Causal Structure"

There are many interesting and important issues concerning the global structure of relativistic spacetimes that might be considered here – the nature and significance of singularities, the cosmic censorship hypothesis, the possibility of "time travel", and others.[65] But we limit ourselves to a few remarks about one rather special topic.

In our discussion of relativistic spacetime structure, we started with geometric models (M, g_{ab}) exhibiting several levels of geometric structure, and used

---

Hence, $t_{b[d} \nabla_{c]} \nabla^a \psi = t_b\, t_{[d}\, t_{c]} (\xi^n \nabla_n \nabla^a \psi) = \mathbf{0}$. This, together with (56), gives us (i). And (ii) follows directly from (57). Finally, for (iii), notice that

$$\begin{aligned}
\overline{\nabla}^a \overline{\nabla}_a \alpha &= h^{ar} \overline{\nabla}_r \overline{\nabla}_a \alpha = h^{ar} \overline{\nabla}_r \nabla_a \alpha = h^{ar} (\nabla_r \nabla_a \alpha + \widehat{C}^n{}_{ra} \nabla_n \alpha) \\
&= \nabla^a \nabla_a \alpha + h^{ar}\, t_{ra} (\nabla^n \psi)(\nabla_n \alpha) = \nabla^a \nabla_a \alpha.
\end{aligned}$$

The final equality follows from the orthogonality condition.

[65] Earman [1995] offers a comprehensive review of many of them. (On the topic of singularities, I can also recommend Curiel [1999].)

the latter to define the (two-place) relations $\ll$ and $<$ on M.[66] The latter are naturally construed as relations of "causal connectibility (or accessibility)". The question arises whether it is possible to work backwards, i.e., *start* with the pair $(M, \ll)$ or $(M, <)$, with M now construed as a bare point set, and recover the geometric structure with which we began. The question is suggested by long standing interest on the part of some philosophers in "causal theories" of time or spacetime. It also figures centrally in a certain approach to quantum gravity developed by Rafael Sorkin and co-workers. (See, e.g., Sorkin [1995; forthcoming].)

Here is one way to make the question precise. (For convenience, we work with the relation $\ll$.)

Let (M, g_{ab}) and $(\overline{M}, \overline{g}_{ab})$ be (temporally oriented) relativistic spacetimes. We say that a bijection $\phi : M \to \overline{M}$ between their underlying point sets is a *causal isomorphism* if, for all p and q in M,

$$p \ll q \quad \Longleftrightarrow \quad \phi(p) \ll \phi(q). \tag{58}$$

Now we ask: Does a causal isomorphism have to be a homeomorphism? a diffeomorphism? a conformal isometry?[67]

Without further restrictions on (M, g_{ab}) and $(\overline{M}, \overline{g}_{ab})$, the answer is certainly 'no' to all three questions. Unless the "causal structure" (i.e., the structure determined by $\ll$) of a spacetime is reasonably well behaved, it provides no useful information at all. For example, let us say that a spacetime is *causally degenerate* if $p \ll q$ for all points p and q. *Any* bijection between two causally degenerate spacetimes qualifies as a causal isomorphism. But we can certainly find causally degenerate spacetimes whose underlying manifolds have different topologies (e.g., Gödel spacetime and a rolled-up version of Minkowski spacetime).

There is a hierarchy of "causality conditions" that is relevant here. (See, e.g., Hawking and Ellis [1972, section 6.4].) They impose, with varying degrees of stringency, the requirement that there exist no closed, or "almost closed", timelike curves. Here are three.

[66] Recall that $p \ll q$ holds if there is a future-directed timelike curve that runs from p to q; and $p < q$ holds if there is a future-directed causal curve that runs from p to q.

[67] We know in advance that a causal isomorphism need not be a (full) isometry because conformally equivalent metrics g_{ab} and $\Omega^2 \, g_{ab}$ on a manifold M determine the same relation $\ll$. The best one can ask for is that it be a conformal isometry, i.e., that it be a diffeomorphism that preserves the metric up to a conformal factor.

chronology: There do not exist closed timelike curves. (Equivalently, for all p, it is *not* the case that $p \ll p$.)

future (resp. past) distinguishability: For all points p, and all sufficiently small open sets O containing p, no future directed (resp. past directed) timelike curve that starts at p, and leaves O, ever returns to O.

strong causality: For all points p, and all sufficiently small open sets O containing p, no future directed timelike curve that starts in O, and leaves O, ever returns to O.

It is clear that strong causality implies both future distinguishability and past distinguishability, and that each of the distinguishability conditions (alone) implies chronology. Standard examples (Hawking and Ellis [1972]) establish that the converse implications do not hold, and that neither distinguishability condition implies the other.

The names "future distinguishability" and "past distinguishability" are easily explained. For any p, let $I^+(p)$ be the set $\{q : p \ll q\}$ and let $I^-(p)$ be the set $\{q : q \ll p\}$. Then future distinguishability is equivalent to the requirement that, for all p and q,

$$I^+(p) = I^+(q) \;\Rightarrow\; p = q.$$

And the counterpart requirement with I^+ replaced by I^- is equivalent to past distinguishability.

We mention all this because it turns out that one gets a positive answer to all three questions above if one restricts attention to spacetimes that are *both* future and past distinguishing.

Proposition 3.3.1. *Let* (M, g_{ab}) *and* $(\overline{M}, \overline{g}_{ab})$ *be (temporally oriented) relativistic spacetimes that are past and future distinguishing, and let* $\phi : M \to \overline{M}$ *be a causal isomorphism. Then* ϕ *is a diffeomorphism and preserves* g_{ab} *up to a conformal factor, i.e.,* $\phi_\star g_{ab}$ *is conformally equivalent to* $\overline{g}_{ab}$.

A proof is given in Malament [1977]. A counterexample given there also shows that the proposition fails if the initial restriction on causal structure is weakened to past distinguishability or to future distinguishability alone.

References

[Bain, 2004] J. Bain. Theories of Newtonian gravity and empirical distinguishability. *Studies in the History and Philosophy of Modern Physics*, 35:345–376, 2004.

[Budden, 1998] T. Budden. Geometric simultaneity and the continuity of special relativity. *Foundations of Physics Letters*, 11:343–357, 1998.

[Cartan, 1923] E. Cartan. Sur les variétés a connexion affine et la théorie de la relativité généralisée. *Annales Scientifiques de l'Ecole Normale Supérieure*, 40:325–412, 1923.

[Cartan, 1924] E. Cartan. Sur les variétés a connexion affine et la théorie de la relativité généralisée. *Annales Scientifiques de l'Ecole Normale Supérieure*, 41:1–25, 1924.

[Coleman and Korté, 2001] R. Coleman and H. Korté. Hermann Weyl: Mathematician, physicist, philosopher. In E. Scholz, editor, *Hermann Weyl's* Raum-Zeit-Materie *and a General Introduction to his Scientific Work*. Birkhäuser Verlag, 2001.

[Curiel, 1999] E. Curiel. The analysis of singular spacetimes. *Philosophy of Science*, 66:S119–S145, 1999.

[Earman, 1995] J. Earman. *Bangs, Crunches, Whimpers, and Shrieks*. Oxford University Press, 1995.

[Earman, 2001] J. Earman. Lambda: The constant that refuses to die. *Archives for History of Exact Sciences*, 55:189–220, 2001.

[Earman, 2003] J. Earman. The cosmological constant, the fate of the universe, unimodular gravity, and all that. *Studies in the History and Philosophy of Modern Physics*, 34:559–577, 2003.

[Ehlers and Geroch, 2004] J. Ehlers and R. Geroch. Equation of motion of small bodies in relativity. *Annals of Physics*, 309:232–236, 2004.

[Ehlers, 1981] J. Ehlers. Über den Newtonschen Grenzwert der Einsteinschen Gravitationstheorie. In J. Nitsch, J. Pfarr, and E. W. Stachow, editors, *Grundlagen Probleme der Modernen Physik*. Wissenschaftsverlag, 1981.

[Friedrichs, 1927] K. Friedrichs. Eine Invariante Formulierung des Newtonschen Gravitationsgesetzes und der Grenzüberganges vom Einsteinschen zum Newtonschen Gesetz. *Mathematische Annalen*, 98:566–575, 1927.

[Geroch and Jang, 1975] R. Geroch and P. S. Jang. Motion of a body in general relativity. *Journal of Mathematical Physics*, 16:65–67, 1975.

[Geroch, 1971] R. Geroch. A method for generating solutions of Einstein's equation. *Journal of Mathematical Physics*, 12:918–924, 1971.

[Geroch, unpublished] R. Geroch. General Relativity. Unpublished Lecture Notes from 1972, University of Chicago, unpublished.

[Grünbaum, 1973] A. Grünbaum. *Philosophical Problems of Space and Time*. Reidel, 2nd enlarged edition, 1973.

[Grünbaum, forthcoming] A. Grünbaum. David Malament and the conventionality of simultaneity: A reply. In *Collected Papers*, volume 1. Oxford University Press, forthcoming.

[Hawking and Ellis, 1972] S.W. Hawking and G. F. R. Ellis. *The Large Scale Structure of Space-Time*. Cambridge University Press, 1972.

[Hogarth, 1993] M. Hogarth. Standard simultaneity is unique in metrical Minkowski spacetime: A generalization of Malament's theorem. unpublished, 1993.

[Janis, 2002] A. Janis. Conventionality of simultaneity. In E. Zalta, editor, *Stanford Encyclopedia of Philosophy*, 2002. URL=http://plato.stanford.edu/archives/fall2002/entries/spacetime-convensimul/.

[Künzle, 1972] H. Künzle. Galilei and Lorentz structures on space-time: Comparison of the corresponding geometry and physics. *Annales Institute Henri Poincaré*, 17:337–362, 1972.

[Künzle, 1976] H. Künzle. Covariant Newtonian limits of Lorentz space-times. *General Relativity and Gravitation*, 7:445–457, 1976.

[Laugwitz, 1965] D. Laugwitz. *Differential and Riemannian Geometry*. Academic Press, 1965.

[Malament, 1977] D. Malament. The class of continuous timelike curves determines the topology of spacetime. *Journal of Mathematical Physics*, 18:1399–1404, 1977.

[Malament, 1995] D. Malament. Is Newtonian cosmology really inconsistent? *Philosophy of Science*, 62:489–510, 1995.

[Malament, unpublished] D. Malament. Lecture Notes on Differential Geometry. available at
http://www.lps.uci.edu/home/fac-staff/faculty/malament/FndsofGR.html, unpublished.

[Norton, 1995] J. Norton. The force of Newtonian cosmology: Acceleration is relative. *Philosophy of Science*, 62:511–522, 1995.

[Norton, 1999] J. Norton. The cosmological woes of Newtonian gravitation theory. In H. Goenner, J. Renn, J. Ritter, and T. Sauer, editors, *The Expanding Worlds of General Relativity*. Birkhäuser, 1999.

[O'Neill, 1983] B. O'Neill. *Semi-Riemannian Geometry*. Academic Press, 1983.

[Page, 1983] D. Page. How big is the universe today? *General Relativity and Gravitation*, 15:181–185, 1983.

[Reichenbach, 1958] H. Reichenbach. *The Philosophy of Space and Time*. Dover, 1958.

[Rindler, 1981] W. Rindler. Public and private space curvature in Robertson-Walker universes. *General Relativity and Gravitation*, 13:457–461, 1981.

[Sachs and Wu, 1977a] R. K. Sachs and H. Wu. General relativity and cosmology. *Bulletin of the American Mathematical Society*, 83:1101–1164, 1977.

[Sachs and Wu, 1977b] R. K. Sachs and H. Wu. *General Relativity for Mathematicians*. Springer Verlag, 1977.

[Sorkin, 1995] R. D. Sorkin. A specimen of theory construction from quantum gravity. In J. Leplin, editor, *The Creation of Ideas in Physics*. Kluwer, 1995.

[Sorkin, forthcoming] R. D. Sorkin. Causal sets: Discrete gravity. In A. Gomberoff and D. Marolf, editors, *Lectures on Quantum Gravity (Proceedings of the Valdivia Summer School)*. Plenum, forthcoming.

[Spirtes, 1981] P. Spirtes. Conventionalism in the philosophy of Henri Poincaré. unpublished Ph. D. thesis, University of Pittsburgh, 1981.

[Stachel, 1980] J. Stachel. Einstein and the rigidly rotating disk. In A. Held, editor, *General Relativity and Gravitation. One Hundred Years After the Birth of Albert Einstein*, volume I. Plenum, 1980.

[Stein, 1991] H. Stein. On relativity theory and the openness of the future. *Philosophy of Science*, 58:147–167, 1991.

[Trautman, 1965] A. Trautman. Foundations and current problems of general relativity. In S. Deser and K. W. Ford, editors, *Lectures on General Relativity*. Prentice Hall, 1965.

[Wald, 1984] R. Wald. *General Relativity*. University of Chicago Press, 1984.

[Weyl, 1921] H. Weyl. Zur Infinitesimalgeometrie: Einordnung der projecktiven und konformen Auffassung. *Nachrichten der Königlichen Gesellschaft der Wissenschaften zu Göttingen; Mathematisch-Physikalische Klasse*, pages 99–112, 1921.

[Weyl, 1950] H. Weyl. *Philosophy of Mathematics and Natural Science*. Princeton University Press, 2nd edition, 1950.