BULLETIN OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 83, Number 6, November 1977

GENERAL RELATIVITY AND COSMOLOGY

BY R. K. SACHS AND H. WU[1]

Contents

CHAPTER 1.

This survey is for mathematicians but is about physics. We have in mind a reader who hasn't worked on physics since sophomore days but is familiar with tensor algebra, differential topology and Riemannian geometry on the introductory graduate level. The theory of Lie groups is needed for discussing examples but not for the fundamental ideas.

This chapter gives some physics background. Chapters 2–6 discuss basic general and special relativity, including a very brief introduction to the theory of black holes. Chapter 7 gives a sample application, cosmology. We should point out that §7.2 summarizes the basic facts of cosmology and is entirely descriptive; it can be understood by a reader who ignores all but the basic definitions of this survey. Chapter 8 gives some examples of mathematics used in current research.

We concentrate on basic current physics at the textbook or folk theorem level. Bibliographical references will be sparse. **[14]**, **[20]** and **[9]** are physics texts which give historical background and far more details. **[18]** is in the same style as this article.

1.1. Conventions. We give a few examples of our notation and terminology. *Smooth* means C^∞. For us, a *manifold* is paracompact, Hausdorff, real, finite dimensional and smooth. M denotes a manifold throughout. TM is the tangent bundle with projection π; thus if M_x denotes the tangent space of M

Expanded version of invited address given by one of us (R. K. S) at the annual meeting of the American Mathematical Society, January 24, 1975, in Washington, D. C. We are grateful to many colleagues, especially J. Arms, J. Beem, K. Sklower and B. O'Neill for detailed comments; received by the editors March 16, 1976.

AMS (MOS) subject classifications (1970). Primary 53C50, 53–02, 83C99, 83F05, 83–02; Secondary 53C20, 53B30, 83C05, 85–02, 85A40.

[1]Research of both authors partially supported by the National Science Foundation.

at x, then $TM = \cup_{x \in M} M_x$. A *tensor field* on a manifold is by definition smooth unless explicitly indicated otherwise. Let V be a vector field on M; thus V: $M \to TM$ such that $\pi \circ V =$ identity. The value of V at $x \in M$ is denoted by Vx, or sometimes $V(x)$; thus $Vx \in M_x$.

Let N be a manifold and ϕ: $N \to M$ be a smooth map. ϕ^* will denote the pullback and ϕ_* the differential. ϕ is a *homeomorphic imbedding* iff ϕ_* is everywhere nonsingular and ϕ: $N \to \phi(N)$ is a homeomorphism with respect to the topology induced on $\phi(N) \subset M$ by that of M.

Suppose the dimension of M is at least two. A *Lorentzian metric* on M is a symmetric (0, 2)-tensor field g on M such that, $\forall x \in M$, the quadratic form gx on M_x has signature $(+, +, \ldots, +, -)$. Thus gx is nondegenerate and

$$\max\{\dim A | A \text{ is a vector subspace of } M_x \text{ and } g|_A \text{ is negative definite}\} = 1.$$

Then (M, g) is a *Lorentzian manifold*. Suppose ϕ: $M \to M$ is smooth. ϕ is an *isometry* for (M, g) iff ϕ is a diffeomorphism and $\phi^* g = g$. Define (M, g) as *time-orientable* iff there is a vector field V on M such that $g(V, V) < 0$ everywhere; this key definition will be examined much more closely in §2.3.

Let γ: $F \to M$ be a continuous *curve*, i.e. $F \subset \mathbf{R}$ is a connected subset which contains more than one point. γ *goes from* x *to* y iff $F = [a, b]$, $x = \gamma a$ and $y = \gamma b$. Let γ: $F \to M$ be a smooth curve. γ_*: $F \to TM$ will denote the tangent vector field. For example, suppose γ is a homeomorphic imbedding. Then there exists a vector field V on M such that $V \circ \gamma = \gamma_*$. du and d/du, respectively, denote the canonical 1-form and vector field on $F \subset \mathbf{R}$. For example, let ϕ and N be as above, ω be a 1-form on M, γ: $F \to N$ be a smooth curve. Then

$$\omega\big((\phi \circ \gamma)_*\big) = (\phi^*\omega)\big(\gamma_*\big) = (\gamma^*\phi^*\omega)(d/du)\text{: } F \to \mathbf{R}.$$

For $\mathbf{R}^n = \mathbf{R} \times \cdots \times \mathbf{R}$, u^i denotes the ith projection. Thus, on any open submanifold of $\mathbf{R}^n$, $(du^1, \ldots, du^n)$ is a basis for the 1-forms. $(\partial_1, \ldots, \partial_n)$ will denote the dual basis.

The rest of this chapter contains little mathematics and can be skimmed.

1.2. General relativity. General relativity is a theory of nature, especially of gravity. Its central assumption is that space, time, and gravity are merely three aspects of one entity, called spacetime, and modeled by a time-orientable Lorentzian 4-manifold (M, g). General relativity analyzes spacetime, electromagnetism, matter and their mutual influences. The models deal with the complete history of a physical process, viewed as a whole. For example, a point particle is modeled by a curve which represents the past, present, and future of the particle. Thus a point $z \in M$ represents, e.g., here-now.

Now in microphysics, gravity counts as a very minor effect. For example, the mutual gravitational attraction between two electrons is believed to be smaller than their electrostatic repulsion by a factor of more than 10^{40}. But gravity is long-range and cumulative. In the realm of stars and galaxies, it can dominate. For example, the discovery of pulsars has shown that there are some stars which avoid gravitational collapse only by a last-ditch effort, at a

circumference of perhaps fifty miles. For such stars, and for the universe as a whole, general relativity is the best available theory. In principle, it applies throughout macrophysics, as we now discuss.

Newtonian physics can handle weak gravitational effects. It cannot adequately handle very strong ones, nor the high-speed effects which occur when relative speeds comparable to the speed of light are involved, nor quantum effects. Here, and throughout, "quantum" is used loosely; it refers to the "fuzzy, jumpy" behavior of small objects. Special relativity can handle high speed and quantum effects, but not gravitational ones. Current general relativity unifies Newtonian theory and nonquantum special relativity into one theory which can handle both high-speed effects and strong gravitational ones. Thus its one known limitation is that quantum effects cannot be handled systematically; they must be neglected or incorporated ad hoc. In particular, the only known interactions in nature are gravity, nonquantum electromagnetism and certain quantum interactions [**21**]. Thus the only interactions which current general relativity can treat in a fully systematic way are nonquantum electromagnetism and gravity.

As just indicated, we presently have two fundamental physical theories: general relativity for macrophysics and special relativistic quantum theory for microphysics. No one really knows how to combine these, although many attempts have been made.

1.3. Past, present and future. Nonquantum special relativity was introduced around 1905 by Einstein, Lorentz, Poincaré, Minkowski and others. Some ten years later, Einstein introduced general relativity, generalizing from flat to curved Lorentzian manifolds to include gravity. The Newtonian limit of general relativity, special relativity and quantum theory have each been checked literally billions of times. But for many years, only small and poorly measured effects within the solar system indicated that general relativity gives better answers than combining Newtonian physics and nonquantum special relativity ad hoc.

Today, more accurate measurements within the solar system, the apparent success of general relativistic models for white dwarf stars and neutron stars (e.g. pulsars), the possible discovery of the black holes and perhaps even of the gravitational radiation predicted by the theory, and the tentative success of general relativistic cosmology have given the theory a somewhat firmer empirical foundation (cf. [**14**], [**20**]). It will eventually be submerged in a theory which somehow unifies microphysics and macrophysics. But its basic ideas will probably be essential in formulating this more accurate theory.

1.4. Time and motion. We shall always use units such that the speed c of light is the dimensionless number 1. Thus a distance L of 2 seconds means $L = 2$ (light-) seconds $\sim 6 \times 10^8$ meters.

We outline how some fundamental concepts of macrophysics have changed during this century. For example, one used to model physical time by Newton's absolute time. Around 1900, it was realized that this model is inaccurate. Nowadays, a very private time called comoving or proper time, modeled by arclength with respect to a Lorentzian metric, is the only corresponding basic concept.

To see roughly what is involved, imagine in elementary Newtonian physics a small body moving in a straight line. Suppose the motion is described by a smooth function x: $\mathbf{R} \to \mathbf{R}$ with $x(t)$ the Euclidean position, $\dot{x}(t)$ the Newtonian velocity, and $|\dot{x}(t)|$ the Newtonian speed for each Newtonian time $t \in \mathbf{R}$. Assume gravity is negligible.

As in freshman calculus, the function $x(t)$ can be analyzed to advantage through its graph, i.e., the smooth curve γ in $\mathbf{R}^2$ such that γ: $\mathbf{R} \to \mathbf{R}^2$ with $\gamma u = (x(u), u)$ $\forall u \in \mathbf{R}$. The game is to supply $\mathbf{R}^2$ with the Lorentzian metric $g = du^1 \otimes du^1 - du^2 \otimes du^2$ and then replace x by γ up to certain reparametrizations of the domain $\mathbf{R}$ and up to isometries of the image space $(\mathbf{R}^2, g)$ (definitions as in §2.1 following).

$(\mathbf{R}^2, g)$ is time-orientable (§§1.1, 2.3). In the present context, time-orientability simply means that the following convention makes sense. For $y \in \mathbf{R}^2$, a vector $V \in (\mathbf{R}^2)_y$ is *future-directed* iff both $g(V, V) \leqslant 0$ and $du^2(V) > 0$ hold; this convention then *time-orients* $(\mathbf{R}^2, g)$. For example, if $y \in \mathbf{R}^2$, then the vectors $\partial_2 y$ and $\partial_2 y + \partial_1 y$ are future-directed, but $\partial_2 y + 2\partial_1 y$ and $-\partial_2 y$ are not.

Throughout the rest of this section, $(\mathbf{R}^2, g)$, the time-orientation, x: $\mathbf{R} \to \mathbf{R}$ and $\gamma = (x,$ identity map of $\mathbf{R})$ are as above. However, each formal definition below generalizes automatically if $(\mathbf{R}^2, g)$ is replaced by a spacetime (M, g) as defined in §2.4 and γ is replaced by a smooth curve into M; the definitions thus apply throughout general and special relativity. γ is *future-directed* iff $\gamma_* u$ is future-directed $\forall u \in \mathbf{R}$.

Now, in our units, the Newtonian speed is no greater than the speed of light iff $|\dot{x}| \leqslant 1$, where equality holds iff the Newtonian speed equals that of light. On the other hand $\gamma_* = (\dot{x}, 1)$, so $g(\gamma_*, \gamma_*) = \dot{x}^2 - 1$. Thus $g(\gamma_*, \gamma_*) \leqslant 0$ iff the Newtonian speed is no greater than the speed of light. This suggests, though it does not prove, how one can replace the criterion $|\dot{x}| \leqslant 1$, which involves the physically inaccurate concept of Newtonian time. Note that $g(\gamma_*, \gamma_*)$ is a function $\mathbf{R} \to \mathbf{R}$.

DEFINITION 1.4.1. γ *models motion at the speed of light* (resp. *at less than the speed of light*) iff γ is future-directed and $g(\gamma_*, \gamma_*) = 0$ (resp. < 0) everywhere.

Now, as long as $\mathbf{R}^2$ is regarded concretely, replacing x by γ as above still does not excise Newtonian time t: if u^2: $\mathbf{R}^2 \to \mathbf{R}$ is regarded as a distinguished function one can identify u^2 and t conceptually. But the crucial Definition 1.4.1 makes sense even if we regard $\mathbf{R}^2$ merely as a manifold, with the C^∞ structure, g, the above time-orientation, and the usual $\mathbf{R}^2$ orientation as the only given structures. One thus says that "the speed of light is absolute".

This shift to a basis-free viewpoint corresponds to genuinely new physics. The natural automorphisms of (M, g) are now those isometries ϕ: $M \to M$ which preserve the orientation and *preserve the time-orientation*, i.e. for every vector V, $\phi_* V$ is future-directed iff V is itself future-directed; note here that if V is future-directed,

$$g(\phi_* V, \phi_* V) = (\phi^* g)(V, V) = g(V, V) \leqslant 0.$$

A standard computation shows that ϕ: $M \to M$ is such a natural automorphism iff there exist $a^1, a^2, \beta \in \mathbf{R}$ such that

$$\begin{pmatrix} u^1 \circ \phi \\ u^2 \circ \phi \end{pmatrix} = \begin{pmatrix} a^1 \\ a^2 \end{pmatrix} + \begin{pmatrix} \cosh \beta & \sinh \beta \\ \sinh \beta & \cosh \beta \end{pmatrix} \begin{pmatrix} u^1 \\ u^2 \end{pmatrix}.$$

In general, $u^2 \circ \phi \neq u^2$, indeed $\phi^* du^2 \neq du^2$. Thus the phyisically inaccurate concept of absolute time has been done away with thereby. One might try the whole set $\{u^2 \circ \phi | a^2, \beta \in \mathbf{R}\}$ as a replacement. This leads to the standard effects discussed in elementary special relativity texts. For example, if $\beta \neq 0$, the level surfaces of $u^2 \circ \phi$ are not those of u^2 ("relativity of simultaneity"), and $(\phi^* du^2)(\partial_2) > du^2(\partial_2)$ ("time dilation"). Similarly, $|(\phi^* du^1)(\partial_1)| > |du^1(\partial_1)|$ for $\beta \neq 0$ ("Lorentz contraction").

But a different replacement for Newton's absolute time is the fundamental one physically. Since arclengths are intrinsic, it makes mathematical sense to define the *proper time interval*,

$$\tau = \int_{u_1}^{u_2} \left| g(\gamma_* u, \gamma_* u) \right|^{1/2} du,$$

for γ *between* γu_1 *and* γu_2 whenever γ is a smooth future-directed curve which models motion at a speed less than that of light (1.4.1). A proper time interval is interpreted as time measured on any good clock moving with the body modeled by γ. This interpretation has been checked empirically very many times; most of the checks are indirect; none is infinitely accurate; all have yielded consistency (cf. §1.5).

Proper time intervals are conceptually very different from absolute time. Suppose we have two curves γ, $\hat{\gamma}$ as above both of which model motion at a speed less than the speed of light and intersect at exactly two points. Then the two relevant arclengths, in general, differ ("one twin ages more than the other"; in the early days of special relativity, this effect was regarded as paradoxical by some people). In particular, suppose γ is a geodesic of the Levi-Civita connection of $(\mathbf{R}^2, g)$, i.e. γ is a straight line. Then $\hat{\gamma}$ is not a geodesic. A short calculation shows that γ ages more than $\hat{\gamma}$ (Riemannian geometry here incorrectly suggests less). For each spacetime, this "twin inequality" has an algebraic counterpart (§2.2) and a global geometric one (§2.6).

Replacing the Newtonian concept of the speed of light by 1.4.1 and replacing absolute time by arclengths as above were perhaps the most important changes effected by relativity. But every other Newtonian concept was either replaced by some intrinsic concept or dropped entirely. We give one more example.

Imagine a small body falling in the earth's gravity. Suppose air friction, changes in the gravity due to the body itself, etc. are negligible. Galileo and Newton knew that the history of such a body is determined by an initial position and an initial Newtonian velocity, independent of the body's mass, composition, etc.–a bullet can have the same orbit as a spaceship. Einstein perceived that geodesics have a similar property: given a point and a tangent vector at the point, one gets a unique inextendible geodesic, independent of mass, composition, etc. Thus in the current theory, one uses future-directed geodesics of the Levi-Civita connection to model small bodies that are *freely falling*. The physical interpretation of free fall is that the net external force

due to electromagenetism and quantum interactions is negligible, so that only gravity is relevant. Using geodesics then makes gravity simply one aspect of spacetime. For example, a short calculation shows γ above is a geodesic iff the Newtonian velocity $\dot{x}$ is independent of Newtonian time. The Newtonian interpretation of $\dot{x}$ = constant (equivalently, $\ddot{x} = 0$, or no acceleration) is that no net external force, gravitational or otherwise, acts on the body. The relativistic interpretation is two-fold: γ geodesic $\leftrightarrow$ only gravity acts on the body; gravity negligible $\leftrightarrow$ $(\mathbf{R}^2, g)$ has zero curvature. Taken together, the two relativistic interpretations are consistent with the Newtonian one.

1.5. Asides on physics vs. mathematics. "As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality" (Einstein, as quoted in [**14**, p. 43]). For later reference we mention some examples and implications of Einstein's comment.

To regard Definition 1.4.1 as a purely mathematical definition would be a swindle, since "speed of light" has an empirical meaning; nature not being a theorem, the empirical meaning can never be made fully precise mathematically. In what follows, definitions will be more important than theorems since many will be, like 1.4.1, physical postulates in disguise. To handle the extra-mathematical connotations, considerable discussion will be needed in some cases.

The discussion above 1.4.1 does not in any sense prove 1.4.1. One reason, among many, is that Newtonian physics was used and Newtonian physics is inaccurate. Generally speaking, physical theories are guessed, not deduced. In particular, general relativity cannot really be proved from anything simpler than itself.

The physical interpretation of arclengths in §1.4 might be regarded as a prescription for checking whether a Lorentzian metric really exists and measuring if it does exist: sufficiently many such arclengths will determine the Lorentzian metric uniquely. But there are many other ways to check for the existence of, and measure, the Lorentzian metric in general relativity. The real point is that one has an overall mathematical theory which is, by the skin of its teeth, mathematically consistent. The theory has a rich supply of extra-mathematical interpretations as in §1.4. These are at least sufficiently precise so that one can check the theory as a whole against nature as a whole a few billion times. In the absence of contradictions, the theory remains acceptable.

CHAPTER 2

We define the models for space, time and gravity, ignoring matter and electromagnetism for the time being.

2.1. Lorentzian manifolds. Let M be an n-manifold, $n \geqslant 2$. There exists a Lorentzian metric g on M iff there exists a nowhere zero vector field on M. (One direction is immediate: If V is such a vector field, we may assume that V is a unit vector field relative to some Riemannian metric h on M. Let ω be the 1-form dual to V relative to h. Then $g \equiv h - 2\omega \otimes \omega$ is a Lorentzian metric. The converse assertion needs some algebraic topology.) In particular,

if M is connected and noncompact, there exists a Lorentzian metric on M.

Let (M, g) be a Lorentzian manifold, D be the *Levi-Civita connection* of g, i.e. $D_X g = 0$ and $D_X Y - D_Y X = [X, Y]$ for all vector fields X, Y on M. The *curvature tensor* is the unique (1, 3)-tensor field R on M which obeys

$$R(\omega, X, Y, Z) = \omega(D_Y D_Z X - D_Z D_Y X - D_{[Y,Z]}X)$$

for each 1-form ω and all vector fields X, Y, Z on M. The *Ricci tensor* is the symmetric (0, 2)-tensor field Ric defined by

$$\mathrm{Ric}(X, Y) = \sum_{\alpha=1}^{n} R(\omega^\alpha, X, X_\alpha, Y),$$

with $(\omega^1, \ldots, \omega^n)$ any local basis of 1-forms and $(X_1, \ldots, X_n)$ the dual basis.

A basis $(X_1, \ldots, X_n)$ for M_x, $x \in M$, is defined as (ordered and Lorentzian) *orthonormal* iff the dual basis $(\omega^1, \ldots, \omega^n)$ obeys

$$g = \sum_{i=1}^{n-1} \omega^i \otimes \omega^i - \omega^n \otimes \omega^n.$$

Let T be a (0, 2)-tensor field on M. We define trace T as that smooth function on M which obeys

$$(\text{trace } T) = \sum_{i=1}^{n-1} T(X_i, X_i) - T(X_n, X_n), \quad \forall x \in M$$

and for an orthonormal basis $(X_1, \ldots, X_n)$ of M_x; the definition is independent of the choice of $\{X_\alpha\}$. The *scalar curvature* is $s = \text{trace Ric}$.

Asides (A) For Lorentzian manifolds, paracompactness need not be postulated independently; indeed the Lorentzian metric (or any affine connection) allows one to construct a Riemannian metric on the principal bundle. (B) In classical notation: the Levi-Civita connection D is characterized by the pair of equations

$$g_{\mu\nu;\rho} = 0 \quad \text{and} \quad X^\mu_{;\nu} Y^\nu - Y^\mu_{;\nu} X^\nu = X^\mu_{,\nu} Y^\nu - Y^\mu_{,\nu} X^\nu;$$

R is characterized by $\eta_\mu R^\mu_{\nu\rho\sigma} = \eta_{\nu;\rho;\sigma} - \eta_{\nu;\sigma;\rho}$; $(\mathrm{Ric})_{\mu\nu} = R^\rho_{\mu\rho\nu}$ by definition; $s = (\mathrm{Ric})_{\mu\nu} g^{\mu\nu}$ by definition; $(X_1, \ldots, X_n)$ is orthonormal iff

$$g_{\mu\nu} X_1^\mu X_1^\nu = \cdots = g_{\mu\nu} X_{n-1}^\mu X_{n-1}^\nu = 1 = -g_{\mu\nu} X_n^\mu X_n^\nu,$$

and $g_{\mu\nu} X_i^\mu X_j^\nu = 0 \; \forall i \neq j$. Though it happens to be best for relativity, we shall not use this classical index notation at all.

2.2. Lorentzian algebra. Let (M, g) be a Lorentzian n-manifold and suppose $x \in M$. We need many details about the vector space M_x together with the quadratic form gx on M_x. Let $A \subset M_x$ be a vector subspace. The *causal character of* A is: *spacelike* iff $g|_A$ is positive definite, *lightlike* iff $g|_A$ is degenerate, *timelike* otherwise. The *causal character of* $V \in M_x$ is that of span V. Thus V is: timelike iff $g(V, V) < 0$; lightlike iff $g(V, V) = 0$ and $V \neq 0$; spacelike iff $g(V, V) > 0$ or $V = 0$. V is *causal* iff V is timelike or lightlike. For example, if V is causal, $V \neq 0$.

The span of a causal vector corresponds to physically realizable motion (§1.4). We shall always assume no information can travel faster than the

speed of light. On this assumption, a nonzero spacelike vector is of little or no physical interest.

The next two results can be obtained by standard, though slightly tedious algebra. We shall use the standard notation:

$$\|V\| \equiv |g(V, V)|^{1/2} \quad \forall V \in M_x,$$

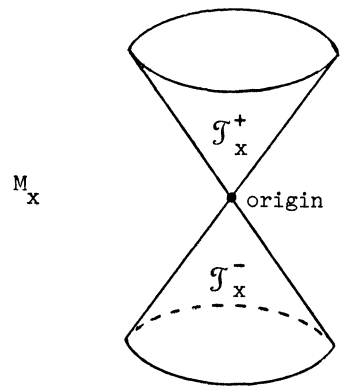
and

$$A^{\perp} \equiv \{V \in M_x | g(V, W) = 0 \, \forall W \in A\} \quad \forall A \subset M_x.$$

2.2.1. *The wrong-way Schwarz inequality.* Suppose $V, W \in M_x$ are causal. Then $|g(V, W)| \geqslant \|V\| \cdot \|W\|$, where equality holds iff span V = span W.

This implies: (A) a timelike vector V and a causal vector W are never orthogonal, i.e. $g(V, W) \neq 0$; (B) two lightlike vectors are orthogonal iff they are proportional (!); (C) A vector V is timelike iff the $(n-1)$-dimensional subspace $V^{\perp}$ is spacelike; (D) V is timelike iff there exists an (ordered, Lorentzian) orthonormal basis $(X_1, \ldots, X_{n-1}, V/\|V\|)$.

2.2.2. *Open solid cones.* (A) The set $\mathcal{T}_x \subset M_x$ of timelike vectors is open and has two connected components, say $\mathcal{T}_x^{\pm}$; $V \in \mathcal{T}_x^{+}$ iff $(-V) \in \mathcal{T}_x^{-}$. (B) Each connected component is a *solid cone*, i.e. $a > 0$, $b \geqslant 0$ and $V, W \in \mathcal{T}_x^{+}$ imply $aV + bW \in \mathcal{T}_x^{+}$.



We illustrate the preceding with a figure, drawn for the case $n = 3$. Locally and globally, physically and formally, this solid cone structure is the heart of relativity, as we shall gradually explain. For example, the existence of two connected components means one can draw a distinction, at least locally, between "heading towards the future" and "heading towards the past" (compare §1.4).

Combining 2.2.2 and 2.2.3 gives the following.

2.2.3. *Wrong-way triangle inequality.* Suppose $V, W \in \mathcal{T}_x^{+}$. Then

$$0 < \|V\| + \|W\| \leqslant \|V + W\|,$$

where equality holds iff span V = span W. The same holds for $V, W \in \mathcal{T}_x^{-}$.

2.3. Time-orientability. Let (M, g) be a Lorentzian manifold, $\mathcal{T} \equiv \{(x, V) \in TM | g(V, V) < 0\}$ be the timelike subset of TM, supplied with the induced topology. Thus $\mathcal{T}_x = (\pi^{-1}x) \cap \mathcal{T}$, $\forall x \in M$.

PROPOSITION 2.3.1. *Suppose M is connected.* (A) $\mathcal{T}$ *is open; it has at most two connected components;* $\forall x \in M$, *the intersection of* $\pi^{-1}x$ *with any connected component is nonempty.* (B) $\mathcal{T}$ *has exactly two connected components iff there is an everywhere timelike vector field on* M.

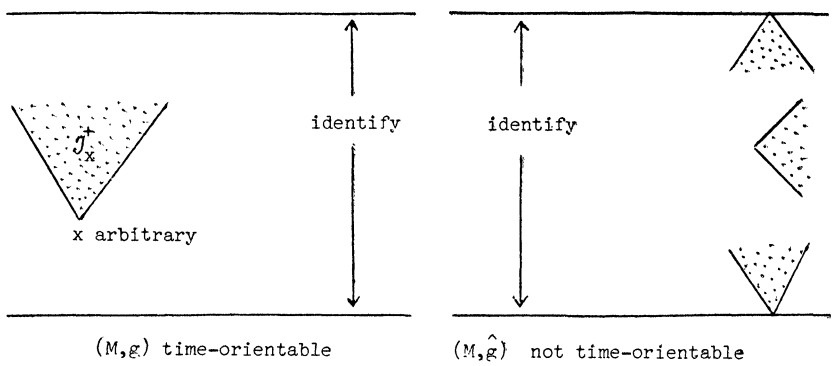
(A) follows from the algebraic counterpart 2.2.2 in a rather routine way on using the bundle isomorphism ψ: $TM \to TM$ given by $\psi(x, V) = (x, -V)$. Now to prove (B), assume $\mathcal{T}$ has two connected components, say $\mathcal{T}^{\pm}$. Then $\forall x \in M$, there is a neighborhood $\mathcal{U}_x$ and a vector field V_x: $\mathcal{U}_x \to T\mathcal{U}_x \cap \mathcal{T}^+$. Let $\{\phi_x\}$ be a smooth partition of unity such that each $\phi_x \geqslant 0$ and has support in $\mathcal{U}_x$. Then $V \equiv \Sigma \phi_x V_x$ is a vector field with value in $\mathcal{T}^+$ since the solid cone property 2.2.2(B) holds pointwise. Thus an everywhere timelike vector field V exists. Conversely, suppose such a V exists. Take $\mathcal{T}^+ = \{(x, W) \in \mathcal{T} \mid g(V, W) < 0\}$ and $\mathcal{T}^- = \psi\mathcal{T}^+$ for ψ as above. Then $\mathcal{T}^+ \cap \mathcal{T}^- = \varnothing$. $\mathcal{T}^+ \cup \mathcal{T}^- = \mathcal{T}$, and each of $\mathcal{T}^{\pm}$ is open. Thus (B) holds.

By 1.1 and 2.3.1(B) (M, g) is time-orientable iff $\mathcal{T}$ has exactly two connected components.

EXAMPLE 2.3.2. Let S^1 be the circle $\mathbf{R}/\pi\mathbf{Z}$, and let $M = \mathbf{R} \times S^1$. Let the natural projections be u: $M \to \mathbf{R}$, ϕ: $M \to S^1$, where we regard $0 \leqslant \phi < \pi$. Thus $d\phi$ is a smooth 1-form on M. With $g = du \otimes du - d\phi \otimes d\phi$, (M, g) is time-orientable. Now define

$$\omega = \cos\phi\, du + \sin\phi\, d\phi, \qquad \eta = -\sin\phi\, du + \cos\phi\, d\phi$$

on the open submanifold $\{\phi \neq 0\}$. Take $\hat{g} = \omega \otimes \eta + \eta \otimes \omega$ and extend $\hat{g}$ to a Lorentzian metric on all of M by continuity at $\phi = 0$. Then $(M, \hat{g})$ is not time-orientable. The following schematic diagrams illustrate both of these possibilities:



In our later definition of a spacetime, we shall demand time-orientability. This is standard: one is interested only in those universes whose beings can distinguish between heading towards the future or the past (cf. [9] for more detailed motivations).

Suppose (M, g) is *time-oriented*, i.e. (M, g) is time-orientable and one component of $\mathcal{T}$ has been designated $\mathcal{T}^+$. Then $V \in M_x$ is *future-directed* iff $V \neq 0$ and $(x, V) \in$ Closure $\mathcal{T}^+$. Thus a future-directed vector is causal, and a causal vector is nonzero and either future-directed or past-directed. Here, as elsewhere, we take for granted dual definitions and results concerning past and future, e.g. the definition of past-directed. The terminology for vectors is

carried over to vector fields and curves in the way indicated by the following examples.

A smooth curve γ: $F \to M$ is *causal* iff $\gamma_* u$ is causal $\forall u \in F$; a vector field V is *future-directed timelike* iff Vx is future-directed timelike $\forall x \in M$; etc. Thus a geodesic γ must be either timelike, lightlike or spacelike, but a general curve need not have a well-defined causal character.

Asides. (A) If a manifold M admits a Lorentzian metric g, then it also admits one, to be called $\hat{g}$, such that $(M, \hat{g})$ is time-orientable. (B) If a Lorentzian manifold either has a lightlike vector field or is simply connected, then it is time-orientable. (C) Neither converse of (B) holds. (D). Time-orientability and orientability are independent conditions.

2.4. Spacetimes. We now give our key definition.

DEFINITION 2.4.1. A spacetime is a connected, oriented, time-oriented, Lorentzian 4-manifold (M, g).

We mention some motivations (cf. §1.5). (A) Four dimensions, rather than three, are needed to model a complete history since intuitively there are three spatial dimensions and one time-dimension. In this context, "disconnected" would connote "always has been, is, and always will be disconnected", so one takes M connected. (B) Some physicists do not regard orientability as essential, but most do. (C) By our definitions, M and g are smooth. As in other physical theories, the motivation for assuming smoothness is obscure. (D) The Lorentzian metric g plays many roles. Going from Newtonian physics to relativity is mainly a matter of excising extraneous structures; somehow g remembers just the right things (cf. §1.4). Concepts of time, distance, gravity, speed of light, acceleration, rotation, causality, etc. are modeled using g, to the extent that they are retained at all.

Literally thousands of modifications of general relativity have been suggested; [**14**] and [**20**] discuss a few. Some of the modifications use a connection with torsion. We shall ignore all the modifications here. In particular, the Levi-Civita connection (§2.1) of a spacetime (M, g) will always be implied.

Henceforth, (M, g) is a spacetime. A geodesic γ: $F \to M$ is *complete* iff $F = \mathbf{R}$; (M, g) is (geodesically) *complete* iff each inextendible geodesic is complete. For an example of a complete spacetime, cf. §2.5 following. For an example of an inextendible geodesic which is not complete, and thus also of a spacetime which is not complete, cf. 3.1.5 following.

The *spacetime equivalence class* of (M, g) is the set $[(M, g)] = \{(M', g') | (M', g')$ is a spacetime and there exists an orientation and time-orientation preserving isometry between M and $M'\}$. An essential connotation of the word "relativity" is that $[(M, g)]$, rather than (M, g) viewed concretely, is the object of interest (cf. §1.4).

2.4.2. *Isometry groups.* Let (M, g) be a spacetime. The *isometry group* of M is

$$\mathcal{G}M \equiv \{\phi: M \to M | \phi \text{ is an isometry}\}$$

(cf. §1.1). $\mathcal{G}M$ is a Lie group of dimension less than or equal to ten (cf. Kobayashi and Nomizu [**12**, p. 238]). Let K be a vector field on M, and L_K be the Lie derivative. K is *Killing* iff $L_K g = 0$. Suppose K is complete. Then K is Killing iff each element of its flow is an isometry. Suppose K is Killing and γ:

$F \to M$ is a geodesic. Then $g(\gamma_*, K)$ is a constant, i.e. a constant function from F to $\mathbf{R}$.

If (M, g) is a really detailed, realistic model, then $\mathcal{G}M$ simply consists of the identity map, just as a faithful model for the surface of the earth, mountains and all, would be a Riemannian 2-manifold without any nontrivial isometries. $\mathcal{G}M$ is nontrivial iff some symmetry idealization–time independence, spherical symmetry, spatial homogeneity, plane wave symmetry, etc.–is being used. For example, suppose K is a complete, future-directed timelike Killing vector field on M. Imagine an observer whose history is modeled by an integral curve of K. Since each element of the flow of K is an isometry, the observer "sees nothing changing" in the geometry as he proceeds towards the future. A spacetime is defined as *time-independent* ($\equiv$ *stationary*) iff there is a timelike Killing vector field on it. This corresponds to the Newtonian concept of a time-independent gravitational field. Somewhat oversimplified in principle, such symmetry idealizations are often essential in practice.

Let $\mathcal{K} \subset \mathcal{G}M$ be a Lie subgroup and suppose $x \in M$. The *orbit of* $\mathcal{K}$ *through* x is $\{y \in M | \phi x = y \text{ for some } \phi \in \mathcal{K}\}$. An orbit is a submanifold: indeed if $\mathcal{K}$ is the subgroup of $\mathcal{K}$ leaving x fixed, then the orbit of $\mathcal{K}$ through x is diffeomorphic to the quotient manifold $\mathcal{K}/\mathcal{K}$.

2.4.3. *Physical equivalence.* For each $y \in M$, let M_y^* denote the dual tangent space; g_y then determines an isomorphism $\psi_y: M_y \to M_y^*$ (the "index-lowering map") via $(\psi_y V)(W) = g(V, W)$, $\forall V, W \in M_y$. V and the covector $\psi_y V$ are by definition *physically equivalent*. The concept of physical equivalence extends in the natural way to tensor fields on M or on subsets of M, as indicated by the following examples. Let $\tilde{\mathcal{F}}$ be a (1, 1)-tensor field on M. The unique (0, 2)-tensor field $\mathcal{F}$ *physically equivalent to* $\tilde{\mathcal{F}}$ is characterized by $\mathcal{F}(V, W) = \tilde{\mathcal{F}}(\psi_y V, W)$, $\forall y \in M$ and $\forall V, W \in M_y$. The unique (2, 0)-tensor field $\hat{\mathcal{F}}$ *physically equivalent to* $\tilde{\mathcal{F}}$ is similarly characterized by $\hat{\mathcal{F}}(\omega, \eta) = \mathcal{F}(\psi_y^{-1}\omega, \psi_y^{-1}\eta)$, $\forall y \in M$, $\forall \omega, \eta \in M_y^*$. If one tensor in a physical equivalence class has a physical interpretation, e.g. as an electromagnetic field (§5.4), the same interpretation is assigned to all other members. For example, suppose f is a smooth function on M. Then the 1-form df is *past-directed timelike* iff, $\forall y \in M$, the vector $\psi_y^{-1}[(df)y]$ is past-directed timelike.

2.5. Minkowski spacetime; special relativity. The most important spacetime in physics is defined as follows.

$$M = \mathbf{R}^4, \qquad g = \sum_{i=1}^{3} du^i \otimes du^i - du^4 \otimes du^4,$$

the orientation is determined by $du^1 \wedge \cdots \wedge du^4$ and the time-orientation is determined by requiring ∂_4 to be future-directed. (M, g) is called *Minkowski spacetime*. Let $[(M, g)]$ be the spacetime equivalence class of Minkowski spacetime. A spacetime (M', g') is in $[(M, g)]$ iff (M', g') is simply connected, is complete, and has zero curvature tensor. This occurs iff the isometry group $\mathcal{G}M'$ is the *Poincaré group*: a ten-dimensional Lie group with four connected components which generalizes the natural automorphism group of §1.4, and can be defined as the semidirect product of $O(3, 1)$ with the group $\mathbf{R}^4$ under the usual representation of $O(3, 1)$ on $\mathbf{R}^4$ [**20**].

$[(M, g)]$ models the trivial gravitational field: no gravity at all. One uses

[(M, g)] iff one is doing quantum or nonquantum special relativity. Note the obvious fact that not every causal curve in M is a geodesic. Thus the assertion, found in some popularizations, that special relativity cannot handle accelerations is (grotesquely) false.

2.6. Elementary global properties. Recall that (M, g) is a spacetime.

Compact spacetimes are almost never used. One reason is this: *Let M be a compact spacetime. Then its universal covering $\tilde{M}$ is a noncompact spacetime.* Here, the spacetime structure of $\tilde{M}$ is obtained by pulling back from M the Lorentzian metric, the orientation and the time-orientation via the covering map. The following topological reasoning shows that $\tilde{M}$ is not compact. If $\tilde{M}$ were compact, its first Betti number would vanish by simple connectivity, and so would its third Betti number by Poincaré duality. The Euler characteristic of $\tilde{M}$ would then be positive. Since $\tilde{M}$ admits a Lorentzian metric, it has a nowhere zero vector field (cf. 2.1). This implies that its Euler characteristic is zero so that $\tilde{M}$ cannot be compact.

(M, g) is defined as *maximal* iff there is no spacetime (N, h) such that M is a proper subset of N and $h|_M = g$. Suppose (M, g) is not maximal and (N, h) is as above. Then one can "see into or out of" M in the following sense: *There is a lightlike geodesic of N which intersects both M and $N - M$.* The proof, here omitted, consists of choosing a point on the boundary of M and using a local argument. In practice, this result gives a convenient way to check spacetimes for maximality, as we shall see. In particular, note that a complete spacetime is maximal.

Whenever possible, one uses maximal spacetimes. But sometimes one is too lazy or insufficiently clever to work out a full model that is maximal. For example, it is often convenient to model (the history of) the earth's exterior, ignoring (the history of) the inside, which is much more complicated and to some extent unknown. In such a case one uses a spacetime which is not maximal (compare 3.2). Even a maximal spacetime is usually not complete. Incompleteness of causal geodesics has rather deep interpretations and implications (Chapters 3, 6 and 7). Incompleteness or completeness of spacelike geodesics is not of much interest physically.

The global structure of spacetime is subtler than that of a Riemannian manifold. Roughly, the key extra question is "who can communicate with whom?" For example, in most cosmological models, not all of spacetime can be observed from the point "here-now", even in principle. Roughly, looking outward in space involves looking backward in time and we get no signals from points which are too distant-late. It has turned out that to analyze such questions, one needs two basic objects, a chronology relation and a chronological distance, which we now define. Henceforth, as a mnemonic device, the use of the specific letter $z \in M$ is usually an invitation to interpret z intuitively as here-now; then $y \in M$ will usually be "earlier" and $x \in M$ "still earlier".

Suppose $(x, z) \in M \times M$. *x chronologically precedes z* iff there is a smooth, future-directed, timelike curve γ: $[a, b] \to M$ from x to z. Here, "chronologically" refers to the fact that the arclength of γ models proper ("comoving") time interval, as explained in §1.4. Define the *chronology relation* $\ll$ by: $x \ll z$

iff x chronologically precedes z. We shall sometimes regard $\ll$ as a subset of $M \times M$. The *chronological past* of z is $I^-\{z\} \equiv \{x \in M | x \ll z\}$. As before, the definitions of dual objects, e.g. the *chronological future* $I^+\{z\}$ of z, will often be taken for granted.

EXAMPLE 2.6.1. Make Minkowski spacetime into a spacetime (N, g) such that N is diffeomorphic to $\mathbf{R}^3 \times S^1$ by identifying $(w, a) \in \mathbf{R}^4$ with $(w, a + 1) \forall w \in \mathbf{R}^3$ and $\forall a \in \mathbf{R}$. Then each point chronologically precedes every point, in particular, itself. $\ll$ is then all of $M \times M$, and $\forall z \in M. I^-\{z\} = M = I^+\{z\}$.

A spacetime *obeys the chronology condition* iff no point chronologically precedes itself. Apart from the present example, we will in this article consider only spacetimes which obey the chronology condition. One reason is that physics would be very confusing if someone could in principle murder his own ancestors. Thus the following comments are merely asides. (A) A spacetime obeys the chronology condition iff each timelike curve is never closed [**16**]. (B) No compact spacetime obeys the chronology condition [**16**]. (C) Certain spacetimes which model rotating black holes violate the chronology condition [**9**]. (D) It seems clear that if one wants to impose the chronology condition for physical reasons, one should go whole hog and impose a more stringent condition (cf. Chapter 8).

We now consider chronological distance. Suppose $x \ll z$. Corresponding to the wrong-way triangle inequality 2.2.3, short, smooth timelike curves from x to z are a dime a dozen. The reader can check that by putting in enough smooth wiggles one can reduce the arclength below any pre-assigned positive value. But if there exists a longest such curve, the curve is a timelike geodesic, by essentially the same argument as in Riemannian geometry.

Take $(0, \infty]$ as $(0, \infty) \cup \{\infty\}$ with the usual order, order topology and addition; thus $a + \infty = \infty \ \forall a \in (0, \infty]$. Regard $\ll$ as a subset of $M \times M$ as before. The *chronological distance function* of spacetime M is $d: \ll \to (0, \infty]$, where $d(x, z) =$ supremum $\{\text{arclength } \gamma | \gamma \text{ is a smooth, future-directed timelike curve from } x \text{ to } z\}$, $\forall (x, z) \in \ll$. In Example 2.6.1, $d(x, z) = \infty \ \forall x, z \in M$, and d is useless. But we shall be mainly interested in cases where d is much better behaved. For example, on Minkowski spacetime,

$$[d(x, z)]^2 = -\sum_{i=1}^{3} (u^i z - u^i x)^2 + (u^4 z - u^4 x)^2 \quad \forall (x, z) \in \ll.$$

The chronology relation and chronological distance have as their basic properties global versions of the algebraic properties 2.2.2 and 2.2.3. The globalization of the solid cone property 2.2.2 (B) is that $\ll$ *is transitive*. This follows from a corner rounding argument [**16**]. So does the fact that d obeys the global wrong-way triangle inequality: whenever $x \ll y \ll z$, $d(x, z) \geq d(x, y) + d(y, z)$. Finally, recall that, $\forall z \in M$, the set $\mathcal{T}_z^-$ of past-directed timelike vectors is open. The global version is the following.

THEOREM 2.6.2 (PENROSE [**16**]). $I^-\{z\}$ *is open* $\forall z \in M$.

Since the constraint that a curve be timelike is an open condition, the theorem is plausible. However, some work is required because $I^-\{z\}$ need not be in the image of the exponential map $\exp_z$. The proof requires two lemmas. Recall that a *geodesically convex* subset $\mathcal{U} \subset M$ is characterized by

the fact that $\forall x, y \in \mathcal{U}$, there is a unique geodesic in $\mathcal{U}$, γ: $[0, 1] \to \mathcal{U}$, from x to y. Further, given $y \in M$, a neighborhood $\mathcal{U}$ of y is called a *normal neighborhood* iff the inverse exponential map restricted to $\mathcal{U}$, $\exp_y^{-1}|_{\mathcal{U}}$, is a well-defined smooth map.

LEMMA 2.6.3. *Given $y \in M$, there is a geodesically convex open neighborhood $\mathcal{U}$ of y which is at the same time a normal neighborhood of each of its points.*

This well-known lemma, due to Henry Whitehead, is valid not just for the Levi-Civita connection but for every affine connection as well (cf. Helgason **[10]** or Kobayashi and Nomizu **[12]**).

LEMMA 2.6.4. *Let $(\mathcal{U}, g)$ be a geodesically convex spacetime such that $\mathcal{U}$ is itself a normal neighborhood of each of its points. Let $x, y \in \mathcal{U}$. Then $x \ll y$ iff the geodesic γ: $[0, 1] \to \mathcal{U}$ from y to x is past-directed timelike.*

PROOF. It suffices to prove the "only if" part. Let $\exp_y$: $\mathcal{Q} \to \mathcal{U}$ be the exponential map, where $\mathcal{Q} \subset \mathcal{U}_y$ is the maximal domain of definition of $\exp_y$. The assumption on $\mathcal{U}$ imply that $\exp_y|_{\mathcal{Q}}$ is a diffeomorphism. Let $\mathcal{T}_y^{\pm} \subset \mathcal{U}_y$ be the two open solid cones of timelike vectors (§2.2), and define $\mathcal{U}^{\pm} \equiv \exp_y(\mathcal{Q} \cap \mathcal{T}^{\pm}) \subset \mathcal{U}$. Then $\mathcal{U}^+$ and $\mathcal{U}^-$ are open and disjoint since $\exp_y$ is a diffeomorphism. Define the "square distance" function f: $\mathcal{U} \to \mathbf{R}$ by

$$fw = g\big(\exp_y^{-1} w, \exp_y^{-1} w\big), \quad \forall w \in \mathcal{U};$$

f is smooth. Now $fw < 0$ iff $w \in \mathcal{U}^+ \cup \mathcal{U}^-$, iff the geodesic from y to w is timelike. Moreover, the vector field R physically equivalent to df is radial **[16]**: $\forall w \in \mathcal{U} - \{y\}$, $Rw = a(\alpha_* 1)$, where $a \in (0, \infty)$ and α: $[0, 1] \to \mathcal{U}$ is the unique geodesic from y to w. In particular, Rw is past-directed timelike iff $w \in \mathcal{U}^-$.

Now suppose $x \ll y$. By definition there exists a smooth past-directed timelike curve γ: $[0, 1] \to \mathcal{U}$ from y to x. A computation gives: $(f \circ \gamma)(0) = 0$, $d(f(\gamma u))/du|_{u=0} = 0$, and $d^2(f(\gamma u))/du^2|_{u=0} < 0$. Thus there is an $\varepsilon > 0$ so small that the image $\gamma[0, \varepsilon] \subset \mathcal{U}^-$. In particular, $f(\gamma\varepsilon) < 0$. Now the point is that γ cannot escape from $\mathcal{U}^-$; specifically, suppose $\gamma 1 \notin \mathcal{U}^-$, and we will derive a contradiction. There exists a $u \in (\varepsilon, 1]$ such that $f(\gamma u) \geqslant 0$. Thus there is a $u_0 \in (\varepsilon, u)$ such that $\gamma u_0 \in \mathcal{U}^-$ and $df(\gamma_* u_0) > 0$. But $R(\gamma u_0)$ and $\gamma_* u_0$ are both past-directed timelike, whence $df(\gamma_* u_0) = g(R, \gamma_* u_0) < 0$ (§2.2). Contradiction. Thus $\gamma 1 \in \mathcal{U}^-$, and the geodesic from y to x is past-directed timelike. □

PROOF OF THEOREM 2.6.2. This is now straightforward. Suppose $x \in I^-\{z\}$ and let $\mathcal{U}$ be a geodesically convex open neighborhood of x which is a normal neighborhood of each of its points (2.6.3). Since there is a smooth future-directed timelike curve from x to z, there is a y on this curve which is in $\mathcal{U}$. Then $x \ll y \ll z$. By Lemma 2.6.4, $x \in \mathcal{U}^-$ where $\mathcal{U}^-$ is as in the proof of that lemma. Thus $\mathcal{U}^-$ is an open neighborhood of x in M. By transitivity of $\ll$, $w \in I^-\{z\}$ $\forall w \in \mathcal{U}^-$. Hence $\mathcal{U}^- \subset I^-\{z\}$. □

2.6.5. *Conformal invariance.* Computing the chronology relation can often be simplified by the following remark. Suppose f: $M \to (0, \infty)$ is smooth and let $g_0 = fg$. Then (M, g_0) is a spacetime in the natural way and clearly $\ll_0 = \ll$.

For physical purposes, one sometimes needs the *causal past* ("observable past") *of* $z \in M$, defined as the set of those points from which there is a smooth, future-directed causal curve to z together with z itself. Clearly $I^-\{z\}$ is a subset of the causal past of z. Causal pasts are much clumsier than chronological ones so one avoids them as much as possible. Fortunately they are not much bigger:

THEOREM 2.6.6 [**16**]. *If x is in the causal past of z, then x is in the closure of the chronological past of z.*

Asides (cf. [**16**]). Suppose $z \in M$; let $I^-\{z\}$ be its chronological past and $J^-\{z\}$ be its causal past. (A) $I^-\{z\}$ is perfectly open, i.e.,

$$\text{Interior}\left[\text{Closure}\left(I^-\{z\}\right)\right] = I^-\{z\}.$$

(B) It now follows from 2.6.6 that Closure $(J^-\{z\})$ = Closure $(I^-\{z\})$, and similarly for the interiors. (C) Sometimes $J^-\{z\}$ is closed, but in general it is neither open nor closed. To get examples, remove one point from Minkowski spacetime. (D) As one might have guessed, the boundary of $I^-\{z\}$ contains lightlike geodesics; however, their behavior is very tricky. (E) A smooth timelike curve intersects the boundary of $I^-\{z\}$ at most once. (F) $x \in$ Closure $I^-\{z\}$ iff $I^-\{x\} \subset I^-\{z\}$.

CHAPTER 3. EXAMPLES OF SPACETIMES

We discuss in this chapter the spacetimes which rank next to Minkowski spacetime in importance.

3.1. Einstein-de Sitter spacetime. The spacetime used most often in current cosmology concisely illustrates most of the points in Chapter 2. For the moment we treat it mainly as a geometric example; a more detailed discussion is given in Chapter 7. Let h be the ordinary Euclidean metric on $\mathbf{R}^3$. Take $M = \mathbf{R}^3 \times (0, \infty)$, with projections $\rho\colon M \to \mathbf{R}^3$ and $t\colon M \to (0, \infty)$. Then $g \equiv t^{4/3}\rho^*h - dt \otimes dt$ is a Lorentzian metric on M. Orient M via

$$\left[\rho^*(du^1 \wedge du^2 \wedge du^3)\right] \wedge dt;$$

take that time-orientation for which ∂_t is future-directed, where ∂_t is defined by $\rho_*\partial_t = 0$ and $dt(\partial_t) = 1$. (M, g) will denote *Einstein-de Sitter spacetime*, defined as above, throughout the rest of this section.

We will shortly show that t and ∂_t are intrinsically attached to the Lorentzian metric and can therefore be defined without reference to the explicit direct product representation of M as $\mathbf{R}^3 \times (0, \infty)$. In Chapter 7, we will motivate the following interpretation rule: t is a cosmological time, increasing from t almost zero near the "big bang" (= the hypothetical "moment of creation" of our universe, or the hypothetical hypersurface $\{t = 0\}$ which is not in M) to a value of roughly 10^{10} years at here-now; moreover, the history of (the center of) any galaxy is modeled by an integral curve of ∂_t. Given these interpretations, the spacetime equivalence class $[(M, g)]$ is a surprisingly accurate model for the history of our universe, at least near here-now. We now consider $[(M, g)]$ geometrically.

3.1.1. *Another representation* (Cf. 2.4). With M, ρ, t, h as above, we define a new Lorentzian metric $\bar{g} = (t/3)^4(\rho^*h - dt \otimes dt)$ and choose the orientation and time-orientation in the obvious manner. Then $(M, \bar{g}) \in [(M, g)]$. To see

this, use the fact that $t \to t^{1/3}$ is a diffeomorphism $(0, \infty) \to (0, \infty)$ which induces an isometry between $(M, \bar{g})$ and (M, g).

All our subsequent discussion will refer to (M, g) rather than to $(M, \bar{g})$ unless otherwise specified. $(M, \bar{g})$ is sometimes useful for technical reasons; compare 3.1.7.

3.1.2. *Curvature. The Ricci tensor of Einstein-de Sitter spacetime* (M, g) *is* $\mathrm{Ric} = (2/9t^2)(g + 2dt \otimes dt)$ *and the scalar curvature is* $s = (4/9t^2)$. The quickest proof consists of introducing the global, orthonormal basis $(t^{2/3}\rho^* du^1, t^{2/3}\rho^* du^2, t^{2/3}\rho^* du^3, dt)$ of 1-forms. Then one can compute the connection forms, curvature forms, curvature tensor, Ric and s much as in Riemannian geometry (cf. §2.1). Note that $t = \frac{3}{2}s^{-1/2}$, so that t is an invariant of the Lorentzian metric, as mentioned earlier. Note also that the scalar curvature obeys $s \to \infty$ for $t \to 0$ ("big bang").

Each level surface of s is a homeomorphically imbedded 3-manifold diffeomorphic to $\mathbf{R}^3$. Let ϕ: $\mathbf{R}^3 \to M$ be such an imbedding; then $\phi\mathbf{R}^3$ is a level surface of t; $\phi\mathbf{R}^3$ is called a *space-slice.* $(\mathbf{R}^3, \phi^* g)$ is isometric to Euclidean 3-space (via $u^i \to t^{2/3}u^i$, for $i = 1, 2, 3$); one thus sometimes says the "spatial curvature" vanishes. Each space-slice corresponds to the popular notion of "our physical universe at a given instant". Though very intuitive, such space-slices are almost completely useless in analyzing actual data: given $z \in M$, e.g. z = here-now, the space-slice through z is disjoint from the chronological past of z (cf. 2.6). Note further that $u^i \circ \phi$ $(i = 1, 2, 3)$ is not in any sense a spatial distance.

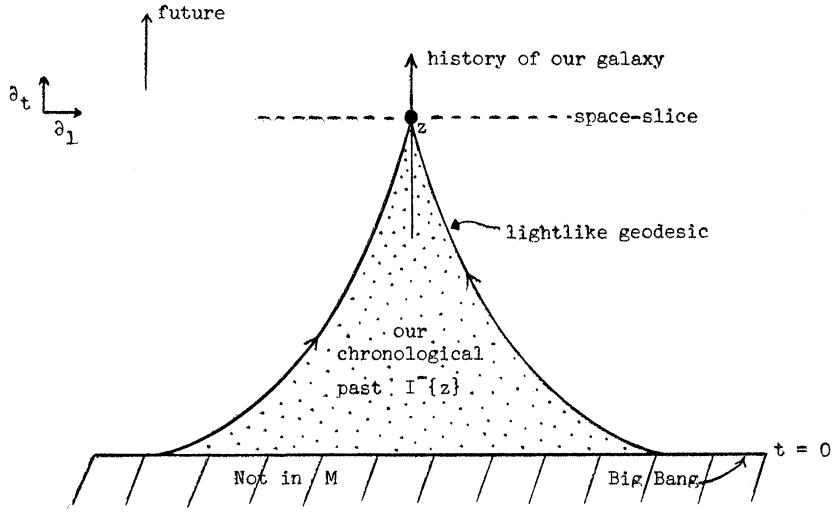


FIGURE 3.1.3

EINSTEIN-DE SITTER SPACETIME. Two dimensions are suppressed. The closed shaded region, corresponding intuitively to $t \leqslant 0$, is not in M. z models here-now. The other structures shown are all described in the text. The fact that the big bang, which is not in M, does not look like a point is consistent with such spacetime boundary constructions as that of Chapter 8. Intuitively, the big bang here is best regarded as a copy of $\mathbf{R}^3$ with zero first fundamental form and infinite second fundamental form.

∂_t can be characterized intrinsically as the vector field physically equivalent to $-dt$. Moreover, let Z be a future-directed vector field such that:

(A) $g(Z, Z) = -1$;

(B) there exists a function μ: $M \to \mathbf{R}$ such that $\mathrm{Ric}(X, Z) = \mu g(X, Z)\ \forall X$.

Then $Z = \partial_t$. This *eigenvector characterization* makes sense in more general cases, e.g., those discussed in Chapter 7.

Fixing an integral curve γ_0 of ∂_t, we shall briefly discuss the behavior of nearby integral curves relative to γ_0. If M ($\equiv \mathbf{R}^3 \times (0, \infty)$) is regarded as the positive half-space $\{u^4 > 0\}$ of $\mathbf{R}^4$, then the integral curves of ∂_t are just the u^4-coordinate curves. Consider the following one-parameter family of integral curves $\{\gamma_s | \leqslant s \leqslant 1\}$ containing γ_0:

$$\gamma_s = \gamma_0 + s(a_1, a_2, a_3, 0),$$

where $(a_1, a_2, a_3, 0)$ is a fixed vector in the hyperplane $\{u^4 = 0\}$ of $\mathbf{R}^4$, and addition is in the sense of vector addition of $\mathbf{R}^4$. The transversal vector field Y ($= \sum_{i=1}^3 a_i \partial_i$) is the infinitesimal version of $\{\gamma_s\}$; it enjoys the property that along γ_0, $g(Y, Y)$ is a monotone increasing function of t, namely, $(\sum_i a_i^2) t^{4/3}$. Recall that each integral curve of ∂_t is used to model the history of a galaxy. As γ_0 and $(a_1, a_2, a_3, 0)$ are arbitrary, $g(Y, Y)$ being monotone increasing then implies that nearby galaxies (i.e., the γ_s's) are running apart. Very loosely, one says "the universe is expanding". Note also that each such integral curve is a geodesic but $(g(Y, Y))^{1/2}$ along such a curve is not a linear function of arclength as it would be if the curvature tensor were zero. In fact $(g(Y, Y))^{1/2}$ has negative second derivative along an integral curve of ∂_t. Intuitively: nonzero curvature $\Leftrightarrow$ nontrivial gravity $\Rightarrow$"slowing down of the expansion rate" in our present case.

3.1.4. *Isometry groups.* Let $\mathcal{G}M$ be the isometry group of (M, g) (cf. 2.4.2), $\mathcal{G}\mathbf{R}^3$ be the usual six-dimensional isometry group of Euclidean 3-space. By inspection of g, $\mathcal{G}M$ contains a subgroup isomorphic to $\mathcal{G}\mathbf{R}^3$. Now by the above, each isometry ϕ must obey $s \circ \phi = s$ and $\phi_* \partial_t = \partial_t$. It follows that two isometries which coincide on a level surface of s coincide everywhere and then that the only isometries are the obvious isometries; thus *$\mathcal{G}M$ is isomorphic to $\mathcal{G}\mathbf{R}^3$* ("spatial homogeneity and isotropy").

3.1.5. *Lightlike geodesics.* Define a smooth curve λ: $(0, \infty) \to M$ by $\lambda u = (3u^{1/5}, 0, 0, u^{3/5})\ \forall u \in (0, \infty)$. A routine computation, e.g. using Killing vector fields (2.4.2), shows λ is a future-directed, inextendible lightlike geodesic. Moreover, λ is fully representative: if $\hat{\lambda}$: $F \to M$ is an inextendible lightlike geodesic, there exists an isometry $\phi \in \mathcal{G}M$ and an affine reparametrization α: $F \to (0, \infty)$ such that $\hat{\lambda} = \phi \circ \lambda \circ \alpha$.

PROPOSITION 3.1.6. (A) *(M, g) is maximal.* (B) *Each inextendible causal geodesic is incomplete.*

PROOF. (A) By our criterion of maximality in §2.6 it suffices to consider lightlike geodesics of M; by 3.1.5 it suffices to consider λ as in 3.1.5. λ cannot be extended beyond ∞ at the upper end. (Roughly: "M cannot be extended beyond $t = \infty$"; this does not follow from the C^∞ structure alone since $(0, 1)$ is diffeomorphic to $(0, \infty)$.) λ cannot be extended to 0 at the lower end since $s \circ \lambda \to \infty$ as $u \to 0$ ("M cannot be extended to include the big bang"). Thus (M, g) is maximal. (B) is fairly routine: every inextendible causal geodesic approaches the big bang. Incidentally, (B) also holds for the spacelike geodesics, but this fact is not of physical interest.

3.1.7. *Chronology*. The quickest way to analyze the chronology relation for (M, g) is to use the representative 3.1.1 and conformal invariance (2.6.6). Suppose $z \in M$. The following hold. (A) $t(z)$ is the chronological distance to the big bang, i.e. $t(z) = \sup\{d(x, z) | x \ll z\}$. (B) $I^-\{z\}$ is the open dotted region in Figure 3.1.3 bounded by lightlike geodesics. For example, if $z = (0, 0, 0, t(z))$, then $x \in I^-\{z\}$ iff

$$t(x)^{1/3} < t(z)^{1/3} - \frac{1}{3}\left(\sum_{i=1}^{3}(u^i x)^2\right)^{1/2}.$$

(C) $\forall x \ll z$, there is a unique timelike geodesic γ: $[0, 1] \to M$ from x to z, and its arclength is $d(x, z)$. This is a very special situation. (D) Let δ: $(0, \infty) \to M$ be an inextendible integral curve of ∂_t such that z is not on the image of δ. Thus with $z =$ (here-now), (part of) δ can model the history of some galaxy different from our own. Then up to reparametrization, there is at most one inextendible lightlike geodesic whose image intersects δ and contains z; there is exactly one iff the image of δ intersects $I^-\{z\}$. This is again very special. Roughly, some galaxies are so distant that they have not yet had a chance to signal z at all ("cosmological horizon").

We now have an example for each key definition and result in Chapter 2.

3.1.8. *Generalizations*. To study the universe, one needs some general theorems and at least one very explicit spacetime; the reasons are outlined in Chapters 7 and 8. Each explicit cosmological spacetime considered in this article will be qualitatively very similar to Einstein-de Sitter spacetime. In particular, each will be topologically trivial. Now the reader has probably read about "closed universes", "spatial curvature", "ultimate collapse", "rotating and shearing universes", "Mach's principle", the cosmological constant, "variable gravitational constant" cosmologies, the steady state cosmology, "tired light" cosmologies, "quantum cosmology", torsion in cosmology, or other famous concepts (cf. [**14**] or [**20**]). Not one of these will be mentioned, let alone discussed, henceforth. We feel they are interesting, but are neither basic nor needed to analyze the main features of the current empirical data and have been grossly overemphasized. Einstein-de Sitter spacetime is basic, very intuitive, mathematically instructive, as accurate physically near here-now as any other explicitly known cosmological spacetime, and–mellowed by 55 years of vigorous give-and-take–free of all inessentials. Regarding it as an exact model of nature would be sheer nonsense (Chapter 7). But it is a truly elegant zeroth order approximation.

3.2. Kruskal spacetimes; black holes. This section will not be needed until Chapter 8, but it belongs here as it nicely illustrates most of the concepts discussed so far.

In 1799, Laplace pointed out that a "heavenly body" might be so massive and small that not even light can escape from the surface [**9**, appendix]; this shows remarkable prescience. Nowadays, "black hole" is the term employed to indicate a region of spacetime in which resides a gravitational field so strong that it allows neither light, nor matter, nor a signal of any kind to escape from the region. It is believed that certain stars with large masses collapse in their final stage of evolution to create black holes. For empirical evidence, pro and con, on whether black holes exist, compare [**4**] and [**19**]. In

this section, we will give the exact definition of a spacetime which contains a subregion from which future-directed causal curves cannot escape. Part of this spacetime is used to model that part of any spherically symmetric situation which is "vacuum", i.e. contains no matter or electromagnetism. A more general and more formal definition of a black hole spacetime is discussed in Chapter 8; there the special assumptions of spherical symmetry and vacuum are unnecessary.

To analyze such a spherically symmetric black hole, or the sun, or a spherically symmetric neutron star, etc., we now introduce a maximal spacetime, called Kruskal spacetime. Kruskal spacetimes will seem rather tricky and rather anti-intuitive at first.

Let S^2 be the ordinary unit 2-sphere. Let h be the usual Riemannian metric on S^2 induced from the Euclidean metric of $\mathbf{R}^3$. Each Kruskal spacetime is diffeomorphic to $\mathbf{R}^2 \times S^2$. Specifically, $u^1u^2 < 1$ defines an open subset $A \subset \mathbf{R}^2$ (Figure 3.2.1). We take $M = A \times S^2$, with projections α: $M \to A$ and σ: $M \to S^2$. Roughly, A will correspond to radius and time while S^2 will correspond to angles. Note that M is connected, simply connected and orientable.

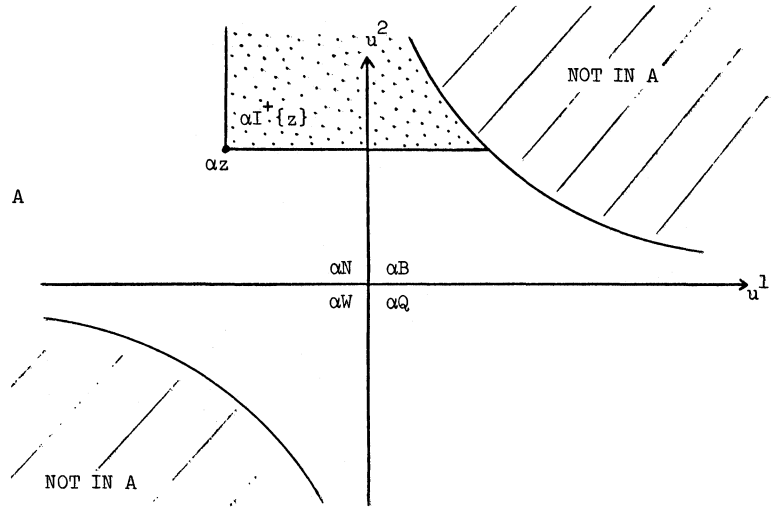


FIGURE 3.2.1

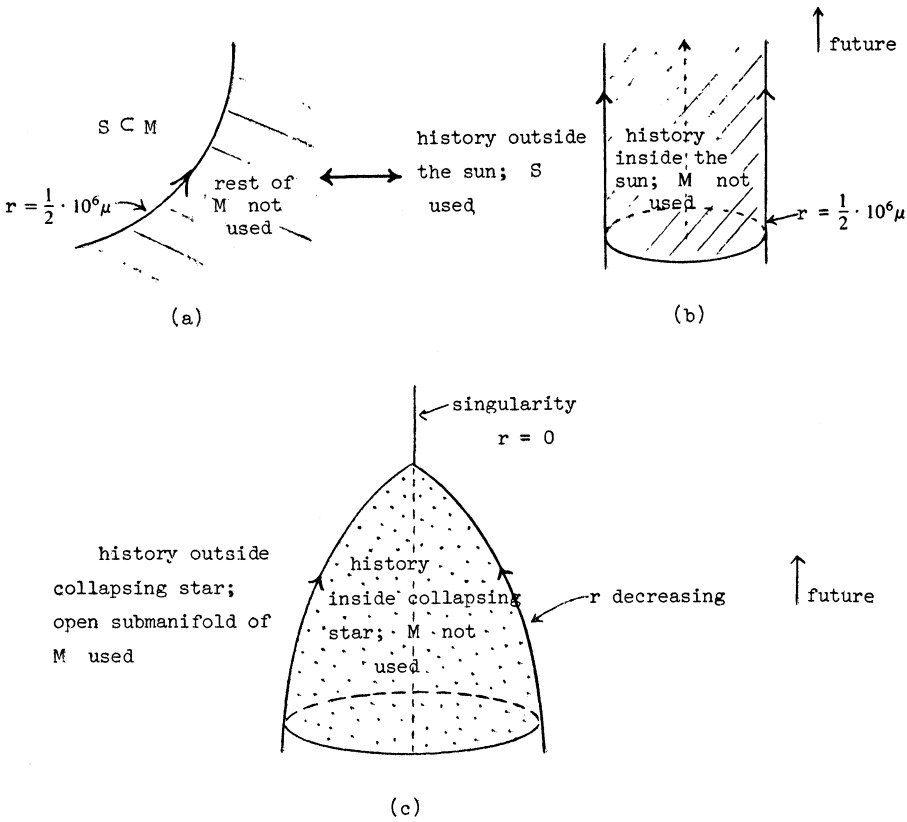
THE KRUSKAL DIAGRAM FOR A. The shaded regions, including the hyperbolae $u^1u^2 = 1$, are not in A. Every other point, including the origin, has as complete inverse image in M a copy of S^2. The figure shows the α-projections of the Schwarzschild black hole B, normal spacetime N, white hole W, and queer duplicate Q; each is defined in the text. $\forall z \in M$, $\alpha I^+\{z\}$ is to the upper right as indicated.

Define $v^i = u^i \circ \alpha$: $M \to \mathbf{R}$ for $i = 1, 2$. Suppose $\mu \in (0, \infty)$ is given; μ will play the role of mass for our black hole, or neutron star, etc. There is a unique, smooth, onto function r: $M \to (0, \infty)$ such that $(r - 2\mu)\exp(r/2\mu) = -2\mu v^1v^2$ because $(x \exp(x/2\mu))' > 0$ $\forall x \in (-2\mu, \infty)$. r is sketched in Figure 3.2.3 below; for reasons to be discussed, it is called the *area-type radius* on M. For the moment interpret r by: for r much bigger than 2μ, r is not too different from Euclidean radius. The above properties of r imply that.

$$g \equiv \left\{-(16\mu^3/r)\exp(-r/2\mu)(dv^1 \otimes dv^2 + dv^2 \otimes dv^1) + r^2\sigma^*h\right\}$$

is a Lorentzian metric on M. This strange form insures Ric = 0, as discussed below. Let ∂_1 be the obvious vector field on M, i.e. $dv^1(\partial_1) = 1$, $dv^2(\partial_1) = 0 = \sigma_* \partial_1$. Then ∂_1 is lightlike and we time-orient M by taking ∂_1 as future-directed. The analogously defined ∂_2 will also be lightlike and future-directed; see Figure 3.2.1. Orient M in the natural way. (M, g) then becomes a spacetime, the *Kruskal spacetime for mass* $8\pi\mu$. For the rest of this section, (M, g) will denote a Kruskal spacetime.

To indicate, in a preliminary way, how such spacetimes are used, we draw some intuitive pictures. (a) indicates the region $\{r > \frac{1}{2} \cdot 10^6 \mu, v^2 > 0\}$ of M. (b) indicates roughly how this region is used. (c) indicates very roughly a spherically collapsing star.



These illustrate fairly well the not-so-simple relationship between the various portions of M and the corresponding real physical situation. The special features of this relationship that are most likely to confuse may be roughly summarized as follows. (A) Usually only a proper subset of M can be used to model a real physical situation; this is because M is a vacuum (cf. 6.2 and 3.2.4 following). (B) Only a part of any real physical situation can be modeled by M; this is again because where matter or electromagnetism is present, Kruskal spacetime M is not applicable.

Black hole theory is heavily concerned with who can communicate with whom. Anticipating the fact that the subregion $\{v^1 > 0, v^2 > 0\}$ of M will be

used to model (the history of) a black hole, we start with a brief general discussion of Kruskal spacetime from the viewpoint of §2.6. The full story is quite complicated, but projecting with α gives very simple results. To see the latter suppose γ: $[a, b] \to M$ is a smooth, future-directed timelike curve. Then from the form of g, $\tilde{\gamma} \equiv (\alpha \circ \gamma, \sigma(\gamma a))$: $[a, b] \to A \times S^2 = M$ is also smooth, future-directed and timelike, and we have $\alpha \circ \gamma = \alpha \circ \tilde{\gamma}$. Thus to judge the α projections we can confine attention to curves which are "radial", i.e. $\sigma \circ \gamma =$ constant (the North Pole, for instance). By virtue of our remark on conformal factors in 2.6.6, this comes down to analyzing $\ll$ for $(A, du^1 \otimes du^2 + du^2 \otimes du^1)$, which is easy. One finds the following: $\forall z \in M$, *the α projection of the chronological future of z is the open upper right quadrant with vertex at αz* (Figure 3.2.1); *in particular, no signal can leave the quadrant αB and none can enter the quadrant αW.* In this sense $B \equiv \{v^1 > 0, v^2 > 0\}$ is "black".

We next consider curvature. By a direct computation of the curvature tensor R, one finds $R \neq 0$, Ric $= 0$ and the following convenient lemma. Let C: $M \to \mathbf{R}$ be the "quadruple trace" of $\hat{R} \otimes \check{R}$, where $\hat{R}$ (resp. $\check{R}$) is the (4, 0)-tensor field (resp. (0, 4)-tensor field) physically equivalent to R. Thus, if $\{\omega^\mu\}$ and $\{X_\mu\}$ are dual bases at $w \in M$,

$$Cw \equiv \sum_{\mu,\nu,\rho,\sigma=1}^{4} \hat{R}(\omega^\mu, \omega^\nu, \omega^\rho, \omega^\sigma)\check{R}(X_\mu, X_\nu, X_\rho, X_\sigma).$$

Roughly, C is the overall square curvature, though the fact g is Lorentzian means that not every term in the sum is positive.

LEMMA 3.2.2. $C = 144\mu^2/r^6$.

Thus r is intrinsically defined. Note $v^1v^2 \to 1$ iff $r \to 0$, iff $C \to \infty$ ("curvature singularity"). $C \to 0$ iff $r \to \infty$ ("spatial infinity"). From the lemma, we find that (M, g) *is maximal.* The proof is much as in §3.1, e.g. M cannot be extended to $r = 0$ because $C \to \infty$ there. We sketch r by showing the α projections of some level surfaces.

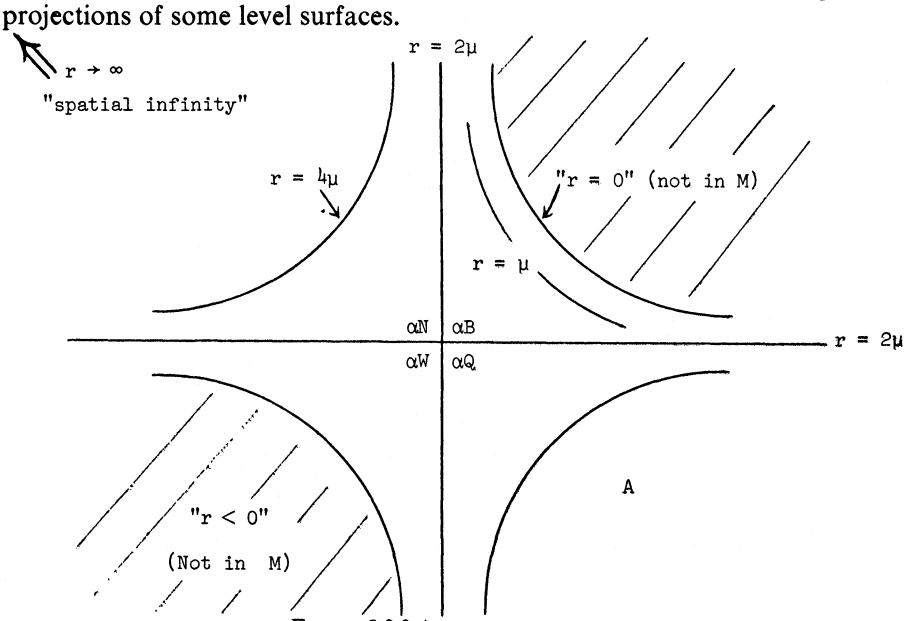


FIGURE 3.2.3. AREA-TYPE RADIUS

3.2.4. *Submanifolds.* We now define some submanifolds of (M, g) which have geometric, physical or historical interest. (A) The *throat* is the set of points on which $dr = 0$; it is the 2-sphere of area-type radius $r = 2\mu$ represented by the origin in Figures 3.2.1 and 3.2.3. (B) The *horizon* is the set of points on which dr is lightlike or zero. It corresponds to the axes in the figures and thus to $r = 2\mu$. Part of the horizon will be interpreted below. (C) The *normal Schwarzschild spacetime* N is the complete inverse image under α of the open upper left quadrant in the figures, i.e. $N = \{v^1 < 0, v^2 > 0\} = \{r > 2\mu, v^2 > 0\}$. Like the next three submanifolds, it is a spacetime in its own right whose equivalence class was found by Schwarzschild in 1916; gluing the four together took physicists nearly 50 years, mainly because the concept of maximality was not well understood. (D) The Schwarzschild *black hole* B, *white hole* W, and *queer duplicate* Q correspond to the remaining three quadrants as indicated. The only intrinsic difference between W and B is furnished by the time-orientation. N and Q are intrinsically identical. But B and N are very different from each other, e.g. chronological futures are trapped by B as above. (E) Having mastered the preceding, the reader will be pained to learn that most physical models use still other submanifolds. For example, let μ be the solar mass in the sense $\mu = (1/8\pi)$ (Newtonian gravitational constant)(mass of the sun as found in tables). Then, in our units (§1.4), μ is roughly $\frac{1}{2} \cdot 10^{-5}$ seconds. The area-type radius r_0 of the sun's actual surface is much bigger, about $\frac{1}{2} \cdot 10^6\mu$. As the black hole B is defined by $r < 2\mu$, the sun is far from being a black hole. To model the history of the outside of the sun, one uses merely that open submanifold of N on which $r > r_0$. This model is in rather good agreement with observations and constitutes one of the main empirical checks in general relativity ([**14**], [**20**]). B, W and Q are irrelevant to the model. (M, g) cannot be used at all to model the inside of the sun since Ric = 0 corresponds to no matter (§6.2). A model valid for both inside and outside has to be more sophisticated.

Of the various pieces of (M, g), N, B and their common boundary are those most often used. Indeed it is primarily via the effect of the curvature of N on other nearby objects, e.g. a companion star, that one hopes to detect black holes empirically. On N, especially for $r \gg 2\mu$, r has (to good approximation, but not exactly) the properties Newtonian intuition assigns to radius in addition to the exact property area $= 4\pi r^2$ (see 3.2.5). By way of contrast, $-dr|_B$ is future-directed timelike. To have an area-type radius also act as a kind of time, as $-r|_B$ does in this case, is very different from anything Newtonian intuition can handle.

The above interpretation for $r|_N$ is more evident in the *Schwarzschild representative* $(\hat{N}, \hat{g})$ of $[(N, g)]$; this is the form of (N, g) discovered by Schwarzschild and is the following. Take $\hat{N} = S^2 \times (2\mu, \infty) \times \mathbf{R}$ with projections $\hat{\sigma}: \hat{N} \to S^2$, $\hat{r}: \hat{N} \to (2\mu, \infty)$ and $t: \hat{N} \to \mathbf{R}$. Let

$$\hat{g} = \hat{r}^2\sigma^* h + \left(1 - \frac{2\mu}{\hat{r}}\right)^{-1} d\hat{r} \otimes d\hat{r} - \left(1 - \frac{2\mu}{\hat{r}}\right) dt \otimes dt,$$

where h is the usual metric on S^2 as before. Then $\hat{\sigma} \leftrightarrow \sigma$, $\hat{r} \leftrightarrow r$ and $t \leftrightarrow 2\mu \ln(-v^2/v^1)$ determine an isometry $\phi: N \leftrightarrow \hat{N}$. r is then clearly displayed as a "radius" via ϕ. Moreover, $\forall a \in \mathbf{R}$, $t \to t + a$ determines an

isometry $\hat{N} \to \hat{N}$. Let ∂_t be the corresponding vector field on N, i.e. $\hat{\sigma}_* \phi_* \partial_t = 0 = dr(\partial_t)$, $dt(\phi_* \partial_t) = 1$. By 2.4.2, ∂_t is Killing; ∂_t is also timelike. Thus (N, g) is time-independent (2.4.2 again). The intrinsic characterization of ∂_t is the following. *Let K be the Killing vector field on M such that $K|_N$ is future-directed, timelike and that $g(K, K) \to -1$ at spatial infinity (i.e. for $r \to \infty$). Then $K|_N = \partial_t$.*

3.2.5. *Spherical symmetry.* Let (M, g) be a Kruskal spacetime. By construction, the isometry group $\mathcal{G}M$ contains at least one subgroup isomorphic to the ordinary rotation group $O(3)$ since each isometry ϕ: $S^2 \to S^2$ induces a unique isometry $\tilde{\phi}$: $M \to M$ which leaves A pointwise fixed, i.e. $\alpha \circ \tilde{\phi} = \alpha$. Call this subgroup $\mathcal{K}$. Each orbit (2.4.2) of $\mathcal{K}$ is a homeomorphically imbedded 2-submanifold diffeomorphic to S^2. Indeed the form of g implies that the orbit through $x \in M$ has the inner geometry of an ordinary 2-sphere with area $4\pi r^2(x)$; hence the name "area-type distance". (Actually the connected component of the identity of $\mathcal{G}M$ is isomorphic to $SO(3) \times \mathbf{R}$, as a result of the existence of the Killing vector field K in 3.2.4.) In general, a spacetime (M_0, g_0) is called *spherically symmetric* iff $\mathcal{G}M_0$ contains at least one subgroup $\mathcal{K}_0$ isomorphic to $O(3)$ such that no orbit of $\mathcal{K}_0$ is more than 2-dimensional and that each 2-dimensional orbit has the intrinsic geometry of an ordinary 2-sphere. Thus Kruskal spacetime is spherically symmetric. Newtonian intuition about a spherically symmetric spacetime such as radius and angle is sometimes useful, but not always, as we have seen in 3.2.4. Specifically, note in the case of Kruskal spacetime that no orbit of $\mathcal{K}$ is merely a single point; intuitively speaking the geometry is so screwed up near a black hole (the portion of M where r is small and therefore one expects to find a point orbit of $\mathcal{K}$) that there are no centers of rotation, even though there are rotations. Thus one certainly cannot interpret r as "distance from the center".

The following characterization of Kruskal spacetime holds. As we have discussed, Kruskal spacetime is not flat, is Ricci flat, maximal, simply connected and spherically symmetric. Conversely, *let (M_0, g_0) be a spacetime which is not flat, is Ricci flat and spherically symmetric; then there is exactly one $\mu \in (0, \infty)$ such that (M_0, g_0) is locally isometric to the Kruskal spacetime of mass $8\pi\mu$. Furthermore, if (M_0, g_0) is simply connected and maximal, then the isometry is a global one* [**9**]. We give an application of this characterization. It was already mentioned in 3.2.4 that a Kruskal spacetime cannot be used to model both the inside and the outside of a star. If one wants such an overall model, one needs a different spherically symmetric spacetime for the inside, as shown intuitively in the pictures (a)–(c) earlier in this section. However, because of the preceding characterization of Kruskal spacetime (M, g), the following makes sense for either the sun, or a star, or a star collapsing towards a black hole. Whatever happens inside the star, Ric = 0 outside so that the outside is always modeled locally by part of (M, g). Thus the star leaves behind its outside gravitational field, like the grin of the Cheshire cat, whenever the whole process is spherically symmetric.

3.2.6. *The black hole B.* In the opening paragraph of this section, we gave some intuitive background information concerning black holes. We now discuss some mathematical properties of the black hole spacetime B which show that indeed B lives up to intuitive expectations.

It has already been shown earlier that no future-directed timelike curves can escape from the black hole B; the same then holds for all causal curves, by Theorem 2.6.6. By putting in more details, one can verify the following: *Suppose* $x \in B$; *then* r *decreases along any future-directed causal curve from* x; *moreover, the chronological distance to the singularity* $\{r = 0\}$, *i.e.*,

$$\sup\{d(x,y) | x \ll y\},$$

exists and is less then $\pi\mu$. For μ the solar value as in 3.2.4, the chronological distance is about $10^{-5}\pi$ seconds. Finally, we remark that B is not time independent (2.4.2); the K of 3.2.4 is spacelike on B.

Now imagine yourself as a future-directed timelike curve in $B \subset M$. You notice "gravity is increasing": r is decreasing and the total square curvature C (Lemma 3.2.2) increases. No matter how you twist and turn, you must head for the future where infinite curvatures ($C \to \infty$) are waiting. You have at most $\pi\mu$ seconds, e.g. less than 10^{-4} seconds, of your own proper time to live (cf. 1.4).

The boundary, $\{r = 2\mu, v^1 > 0\}$, between N and B is part of the horizon. Intuitively, this boundary models the history of the surface of the black hole region in space. It can best be visualized by using the vector field K of 3.2.4. On restricting to the boundary, one finds the integral curves of K lie within the boundary and are future-directed lightlike geodesics. One can regard these integral curves as light signals "running outward as fast as anything can" in a desperate attempt to escape the gravity; they "nonetheless just stay in the same place" in the sense that each element of the flow of K is an isometry. Intuitively, these light signals are trapped in the surface of the black hole region (in space), forever marking time by keeping the same distance from the center of the region. (But one should not push this intuitive comment too far, in view of the remarks about r in B of the preceding sections.) Thus imagine the boundary as the history of a 2-sphere which expands at the speed of light but nevertheless retains constant area-type radius 2μ!

It is fatally easy to fall into B. Indeed, if a spaceship hovering above the North Pole of a black hole at $r = 10\mu$ turns off its motor, it will follow a timelike geodesic γ with an initial tangent $\gamma_*(0)$ for which $dr(\gamma_*) = 0 = \sigma_*(\gamma_*)$. Each such geodesic eventually enters B as one can see by a computation of the geodesics, e.g. via Killing vector fields (2.4.2). Figure 3.2.7 illustrates this situation.

Note that the line segments in αN or αB parallel to the coordmate axes in $\mathbf{R}^2$ are lightlike geodesics; this is a straightforward computation. Thus suppose another spaceship $\hat{\gamma}$ continues to hover above the North Pole of the black hole at $r = 10\mu$. Then the signal s_1 which γ sends out at his proper time t_1 before falling into the black hole ("I seem to be in a funny gravitational field.") will reach $\hat{\gamma}$ in finite time, but the signal s_2 which γ sends out at his proper time t_2 after entering the black hole ("Help!") will never reach $\hat{\gamma}$. See Figure 3.2.7.

Finally, we remark that signals can leave but not enter the white hole W and that the queer duplicate Q has no known application other than simply replacing N by Q.

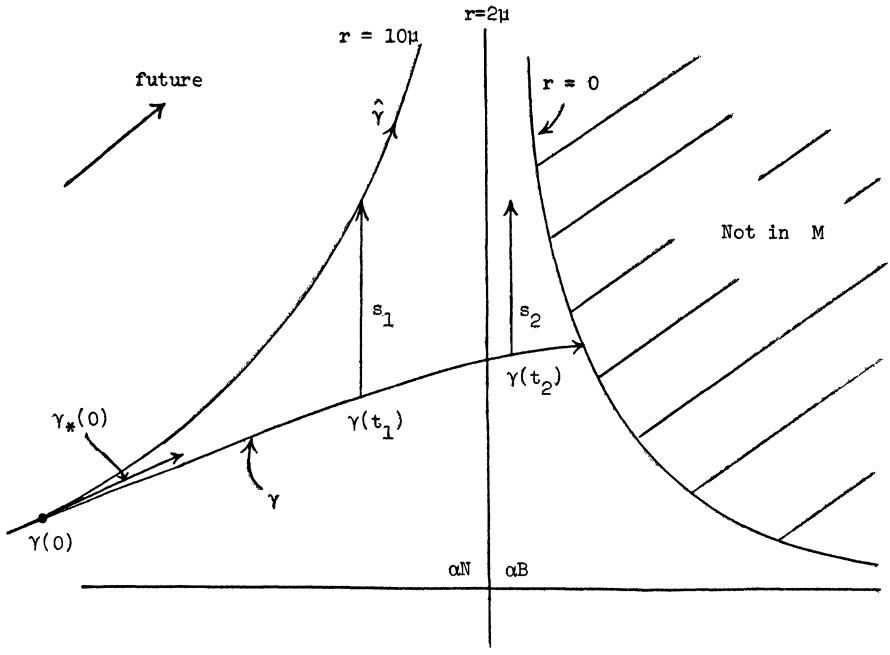


FIGURE 3.2.7

CHAPTER 4. MEASUREMENTS AND PARTICLES

Since general relativity is supposed to cover all of macrophysics, we proceed to deal with observations, matter, and electromagnetism in this chapter and the next. (M, g) is a spacetime (2.4.1) throughout.

4.1. Instantaneous observers. At one instant of her own proper time, an actual observer in M not only determines a point $z \in M$, but also determines a timelike direction at M_z tangent to her history as she heads towards the future (cf. §1.4). Abstracting, and normalizing the direction for convenience, suggests the following definition.

DEFINITION 4.1.1. An *instantaneous observer* (z, Z) on M is a point $z \in M$ and a vector $Z \in M_z$ such that $g(Z, Z) = -1$ and Z is future-directed.

Note that then $M_z = (\text{span } Z) \oplus Z^{\perp}$, with $Z^{\perp}$ 3-dimensional spacelike (§2.2).

Let (M, g) be Minkowski spacetime (2.5). A (global) *inertial frame* on Minkowski spacetime is a vector field Z on M such that: (A) $\forall x \in M$, (x, Zx) is an instantaneous observer; (B) Z is covariant constant, i.e. $D_X Z = 0$ $\forall$ vector X. For example, ∂_4, $\frac{1}{3}(5\partial_4 + 4\partial_1)$ and $\frac{1}{4}(5\partial_4 - 3\partial_3)$ are global inertial frames on (M, g). If Z, $\hat{Z}$ are two such frames, there exists an isometry ϕ: $M \to M$ such that $\phi_* Z = \hat{Z}$. The concept of an inertial frame cannot be generalized to other interesting spacetimes; for example there is no covariant constant nonzero vector field on Einstein-de Sitter spacetime or on a Kruskal spacetime.

Minkowski spacetime together with a preassigned inertial frame Z_0 will be denoted by (M, g, Z_0). The instantaneous observers $(z, Z_0 z)$, $\forall z \in M$, are called the *inertial observers*. Once an inertial frame Z_0 has been fixed on

(M, g), the resulting physics is extremely naive and harks back to Newton and beyond. For instance, there will be absolute time, absolute rest and absolute distance. To see this, let ω_0 be the 1-form physically equivalent to Z_0; the simple-connectivity of M and the covariant constancy of Z_0 imply that $\omega_0 = -dt$ for some t: $M \to \mathbf{R}$. t will then serve as a universal time function for M. Each level hypersurface $\{t = t_0$ for some $t_0 \in \mathbf{R}\}$ is then a Riemannian manifold in the induced metric (2.2) and is "the universe at time t_0". Furthermore, a particle is at absolute rest iff its history is modeled by an integral curve of Z_0. It follows that a spatial distance between any two points $z, z' \in M$ may be defined by: If $t(z) = a$ and $M_a \equiv \{t = a\}$, let the integral curve of Z_0 through z' intersect M_a at z''. The spatial distance between z, z' is by definition the number $d(z, z') =$ the Riemannian distance between z and z'' on M_a. It is elementary to see that this is well defined.

Although (M, g, Z_0) is too naive for doing honest physics, it can be used to give intuitive motivation for some of the definitions in this chapter and the next.

4.2. Observations. Let (z, Z) be an instantaneous observer. How does (z, Z) observe? Often he can measure "just as in special relativity" by "pretending part of M is part of (M_z, gz)". Indeed, (M_z, gz), regarded as a spacetime, is isometric to Minkowski spacetime used in special relativity (cf. §2.5). The transition from (M, g) to a tangent space parallels the prescription in Riemannian geometry for actually measuring, for instance, an angle between two intersecting curves on a nonflat manifold. Postponing temporarily the question of how measurements are made in special relativity, we give an example of the transition $(M, g) \to (M_z, gz)$. Thus given a neighborhood $\mathfrak{U}$ of z, we shall treat $\mathfrak{U}$ as part of Minkowski spacetime (M_z, gz) via the exponential map $\exp_z$. Precisely, suppose $\mathfrak{U}$ is so small that there is a neighborhood $\mathfrak{A}$ of $0 \in M_z$ such that $\exp_z: \mathfrak{A} \to \mathfrak{U}$ is a diffeomorphism; we wish to replace all considerations in $(\mathfrak{U}, g)$ by those in $(\mathfrak{A}, gz)$ via $\exp_z^{-1}$.

$\exp_z$ not being an isometry of $(\mathfrak{A}, gz)$ onto $(\mathfrak{U}, g)$, the amount of distortion incurred in this process must be estimated. For this purpose, first consider the Riemannian case. Suppose N is a 2-dimensional Riemannian manifold, $x \in N$, and $\exp_x: \mathfrak{B} \to \mathfrak{V}$ is a diffeomorphism of a neighborhood $\mathfrak{B}$ of $0 \in N_x$ onto a neighborhood $\mathfrak{V}$ of x. Let $C_0 \subset \mathfrak{V}$ be a rectangular strip and $\exp_x C = C_0$. Further assume that x is (roughly) equidistant from the four vertices of C_0. A fairly precise estimate of the deviation of the region C in the euclidean space N_x from C_0 in the nonflat N can be made via a detailed study of curvature and Jacobi fields. Falling short of such a full analysis, one can nevertheless make the *rough* estimate that if the Gaussian curvature K of N at x is small in absolute value, and if the area A of C_0 is small, then C is probably a good approximation to C_0. Thus the product $|Kx|A$ is a crude numerical invariant to measure the deviation of C from C_0: the smaller the number $|Kx|A$, the smaller the deviation. Note that $|Kx|A$ is independent of units (cm , inches, . . .).

To return to the original situation of $\exp_z: \mathfrak{A}(\subset M_z) \to \mathfrak{U}(\subset M)$, consider the case of observing a distant galaxy for T seconds by using a telescope of length L seconds ($\sim 3 \times 10^8 L$ meters). We will neglect the other dimensions of the telescope. Thus the history of the telescope in T seconds is modeled by

a rectangular strip $\mathcal{R}_0$ in the spacetime M with sides of roughly L units and T units. Let us place z in the "center" of $\mathcal{R}_0$ and assume $\mathcal{R}_0 \subset \mathcal{U}$. Let further $\mathcal{R} \subset \mathcal{Q}$ be such that $\exp_z \mathcal{R} = \mathcal{R}_0$. How does one measure the deviation of $\mathcal{R}$ from $\mathcal{R}_0$? Guided by the Riemannian case, we inspect the number $|sz|LT$, where s is the scalar curvature of M. Note that, in drawing this analogy, we have made two drastic simplifications: (1) instead of examining the behavior of the sectional curvature of M near z, we merely use the scalar curvature, and (2) to replace the area A above, we use the product LT because it corresponds to the intuitive notion of the "area" of $\mathcal{R}_0$ in M. In any case, the size of this dimensionless number $|sz|LT$ will be used as an indicator whether replacing $\mathcal{R}_0$ by $\mathcal{R}$ would lead to catastrophic consequences: the smaller this number, the safer the replacement. In practice, suppose $L = 3$ meters $\sim 10^{-8}$ seconds and $T = 1$ second. If (M, g) is a typical cosmological spacetime, then $|s|$ at z = here-now is about $10^{-35}(\text{seconds})^{-2}$ (cf. Chapter 7). It follows that $|sz|LT \sim 10^{-43}$, which is very small indeed. Thus for the purpose of analyzing the actual measurement (though of course not for analyzing how the light got from the distant galaxy to the telescope), the telescope can, and for convenience should, be modeled by an object in the flat spacetime M_z rather than in the curved spacetime M.

Such use of dimensionless numerics is a basic fact of life in physics.

Granting that measurements often in effect take place on M_z rather than M, how does (z, Z) measure special-relativistically? Until he learns his job properly, he can often measure "just as in Newtonian physics" by "pretending $(Z^\perp, g|_{Z^\perp})$ is Euclidean 3-space". Indeed, $(Z^\perp, g|_{Z^\perp})$, regarded as a manifold, is Euclidean 3-space. We give an example.

Suppose we have two future-directed lightlike vectors $V, \hat{V} \in M_z$, due to two light signals from two distant galaxies. Corresponding to $M_z = (\text{span } Z) \oplus Z^\perp$, there is an orthogonal projection p: $M_z \to Z^\perp$. The *Newtonian angle* (z, Z) *measured between* V *and* $\hat{V}$ is the ordinary Euclidean angle θ between pV and $p\hat{V}$, i.e.,

$$\cos\theta = g(pV, p\hat{V})/\|pV\| \cdot \|p\hat{V}\|.$$

Of course θ depends on Z and not just on V and $\hat{V}$, an effect called *aberration* by astronomers. To get at the intrinsic physics one must amputate Z. Thus a really competent observer would try to tabulate V and $\hat{V}$ directly rather than tabulating $pV, p\hat{V}, Z$, coordinate components, or other nonphysical garbage.

Aside. The Z dependence above corresponds exactly to the set of conformal transformations of the ordinary 2-sphere onto itself.

4.2.1. *Asides on precision, the principle of equivalence, etc.* The reader not puzzled by the preceding example should omit these asides. (A) In any discussion of actual measurements, the physics vs. mathematics problems mentioned in §1.5 become acute. (B) In 4.2, a heuristic argument was given for using the number $|sz|LT$ as a criterion for whether or not to neglect the curvature of M near z. It was pointed out that this number is a very crude estimate. Suppose in the same situation an instantaneous observer (z, Z) is specified in advance. Then this "numerical criterion of empirical accuracy" could be refined by replacing $|sz|$ with bz, where bz is the maximum of the absolute value of the sectional curvature of the 2-planes which are in $Z^\perp$ or which contain Z. In general, as soon as (z, Z) is given, this bz can be used in

other situations to decide whether, empirically, local curvature can be ignored and Minkowski spacetime be employed. When physics texts talk of "local inertial frames", some such region of "negligible curvature" is involved. (C) The principle of equivalence says, very roughly, that in an appropriate region as above, special relativity holds to high accuracy. Attempting to make this fully precise is not only hopeless but also tends to destroy the enormous heuristic power of the principle. We shall not define, or explicitly use, the principle of equivalence here.

4.3. Mass in general relativity. In Newtonian physics, the inertial mass of a particle is measured by collision experiments which do not involve gravity [1]. The corresponding relativistic concept is that each particle is assigned a fixed *rest-mass* $m \in [0, \infty)$. For $m \neq 0$, the term "rest-mass" refers to one special way of measuring m, i.e. in a collision where all relative speeds involved are negligible compared to the speed of light. For $m = 0$, the term "rest-mass" is rather misleading, as indicated in the next section.

In Newtonian physics, one also uses the active-mass ("gravity-producing mass"), conceptually independent of inertial mass [1]. The corresponding general relativistic concept has already been used in §3.2.

We shall henceforth use units in which the constant G of gravity has value $G = 1/8\pi$, in addition to $c = 1$ (§1.4). Then each mass comes out in seconds. For example, the active-mass of the sun is roughly 10^{-5} seconds (3.2.4). For translations of our units to more familiar, less convenient ones, cf. [20] or [14].

Asides. Newtonian physics even uses a third kind of mass, passive-mass ("gravity-responding mass"). But general relativity does not, since the geodesic law (1.4) already specifies how bodies respond to gravity. Older texts on special relativity sometimes also talk of an "inertial mass which depends on speed"; this concept is obsolete and will not be used.

4.4. Particles. To model the history of a small object, one needs a curve. Specifically, a *particle* on spacetime is a smooth, future-directed curve γ: $F \to M$ such that, for some fixed $m \in [0, \infty)$, $g(\gamma_*, \gamma_*) = -m^2$. m is then defined as the *rest-mass* of γ (§4.3). For example, suppose $m \neq 0$. Then $\forall u \in F$, $\gamma_* u$ is timelike.

The tangent vector field γ_* for a particle γ is defined as the *energy-momentum* of γ. Energy-momentum replaces and unifies two Newtonian concepts, namely, energy and momentum. To clarify this and other similar points, we now introduce some auxiliary concepts by using instantaneous observers (4.1). However, it will be useful to keep in mind that γ_* is the only basic object involved and there is only one basic equation needed: $g(\gamma_*, \gamma_*) = -m^2$. The rest follows. It is therefore the energy-momentum γ_*, not the energy defined below, nor the 3-momentum defined below, nor even the pair (energy, 3-momentum), which models something present in nature even when no observers are actually measuring.

Throughout the rest of this section, γ: $F \to M$ is a particle with rest-mass m; $c = 1$ is the speed of light. For simplicity, assume γ has no self-intersections, i.e. for $t \neq t'$, $\gamma t \neq \gamma t'$.

4.4.1. *Auxiliary concepts and results.* Let (z, Z) be an instantaneous observer such that $z = \gamma u$ for some $u \in F$. Then we have the orthogonal

decomposition $\gamma_* u = EZ + p$, $E \in \mathbf{R}$ and $p \in Z^\perp$ (4.1). E is defined as the *energy* (z, Z) *measures* for γ; p is the 3-*momentum* (z, Z) *measures* for γ. The *speed of* γ *relative to* (z, Z) is by definition the number $v = \|p\|/E$. Note that the last is well defined because $E > 0$; this follows from the fact that both $\gamma_* u$ and Z are future-directed and Z is timelike and $E = -g(\gamma_* u, Z)$ (§2.2).

In this paragraph and the next, we shall give an interpretation of the preceding definitions using Minkowski spacetime together with the inertial frame ∂_4, i.e. (M, g, ∂_4), where ∂_4 is the fourth coordinate vector field on $M = \mathbf{R}^4$ (§4.1). With u^4 as the universal time function associated with ∂_4, the spatial distance between any $z, z' \in M$ becomes simply

$$d(z, z') = \left\{ \sum_{i=1}^{3} (u^i z - u^i z')^2 \right\}^{1/2}$$

(§4.1 again). Now let $\gamma\colon F \to M$ be a particle of mass m; we may assume $0 \in F$. At each $\gamma t \in M$, γ is observed by the inertial observer $\partial_4(\gamma t)$; in particular, let $\gamma_* 0 = E_0(\partial_4 z) + p_0$, where $z = \gamma 0$ and $p_0 \in (\partial_4 z)^\perp$. If we let $\gamma^\alpha = u^\alpha \circ \gamma$, $\alpha = 1, \ldots, 4$, then $E_0 = \dot{\gamma}^4 0$ and $p_0 = \Sigma_{i=1}^3 (\dot{\gamma}^i 0)\partial_i z$. Now using Newtonian physics, the inertial observers would observe the following:

Between $z \equiv \gamma 0$ and $z' \equiv \gamma t$ $(t \in F)$, γ has travelled a distance of $d(z, z')$ in the time interval $(u^4 z' - u^4 z)$. Thus the speed of γ at $t = 0$ is

$$\lim_{z' \to z} \frac{d(z, z')}{(u^4 z' - u^4 z)} = \frac{\|p_0\|}{E_0},$$

as a simple calculation shows. This then suggests the definition of v as above. Let then $v_0 = \|p_0\|/E_0$ and assume $v_0 \ll 1$ ($=^{\text{def}}$ much smaller than 1). This immediately implies $\|p_0\| \ll E_0$ and, from $-m^2 = g(\gamma_* 0, \gamma_* 0) = -E_0^2 + \|p_0\|^2$, one gets $m \sim E_0$. Thus,

$$\begin{aligned} E_0 &= \left(m^2 + \|p_0\|^2\right)^{1/2} = m\left(1 + (\|p_0\|/m)^2\right)^{1/2} \\ &\sim m\left(1 + (\|p_0\|/E_0)^2\right)^{1/2} = m(1 + v_0^2)^{1/2} \\ &= m + \tfrac{1}{2} v_0^2 + O(v_0^4) \sim mc^2 + \tfrac{1}{2} m v_0^2, \end{aligned}$$

because in our units $c = 1$, and terms of order v_0^4 or higher are negligible since $v_0 \ll 1$. This means if the speed v_0 is very small (domain of validity of Newtonian physics), and if the inertial observer computes the Newtonian kinetic energy and Newtonian rest-mass energy of γ, he would come up with the number E_0. Similarly, $v_0 \ll 1$ implies

$$\|p_0\| \sim m v_0,$$

so that if the inertial observer computes the magnitude of the Newtonian momentum of γ, he would come up with the number $\|p_0\|$ provided the speed of γ is very small. These give intuitive content to the foregoing definitions.

Returning to the general situation, one always has $E^2 = m^2 c^4 + \|p\|^2 c^2$ $(c = 1)$. Thus Einstein's famous formula $E = mc^2$ holds iff the 3-momentum (z, Z) measures is zero. $E^2 = m^2 + \|p\|^2$ also implies $0 \leqslant v \leqslant 1$. Thus the relative speed is less than the speed of light iff $m \neq 0$, iff γ is timelike, iff γ

models motion at a speed less than the speed of light (§§1.4 and 2.3).

4.4.2. *Photons.* Two kinds of rest-mass zero particles have been found in nature: photons ("particles of light") and neutrinos; in all likelihood a third kind, gravitons, also exists (cf. [**20**]). We shall not need the latter two and thus formally define γ as a *photon* iff $m = 0$. Thus for us γ is a photon iff γ models motion at the speed of light (1.4.1), iff γ is lightlike (§2.3), iff $\gamma_* u$ is lightlike $\forall u \in F$, iff the speed of γ relative to any instantaneous observer is the speed of light (cf. 4.4.1; "even an instantaneous observer who runs as fast as he can away from a photon still measures c as the overhauling speed"). Then no instantaneous observer measures zero 3-momentum for γ. Suppose γ is a photon, (z, Z) and E are as in 4.4.1 and $h \in (0, \infty)$ is Planck's quantum constant; in our units, $h \sim (10^{-43}$ seconds$)^2$, but this will not be relevant except insofar as the ridiculously small value suggests that quantum effects are not essential for large objects. Define the *frequency* (z, Z) *measures for the photon* γ as $f = E/h$. The term "frequency" refers to the fact that waves can also be used to model light (cf. §5.5), and this definition is motivated by the quantum-theoretic equation $E = hf$ for photons found by Planck and Einstein. We shall make no attempt here to derive the latter (cf. [**13**]). Define the *wave-length* (z, Z) *measures for the photon* γ as $\lambda = 1/f$; this leads to the standard relation $\lambda f =$ wave speed $= 1$ in our units.

CHAPTER 5. MATTER AND ELECTROMAGNETISM

We discuss matter, electromagnetism, their mutual influences, and the influence of spacetime on each. The main point will be that matter has a life of its own, at least as rich and interesting as that of the spacetime it inhabits.

(M, g) is a spacetime throughout.

5.1. Divergence and integration. In analyzing a collection of many particles, one needs appropriate integrals. The *volume-form* of spacetime M is the unique 4-form Ω on M such that $\Omega(X_1, X_2, X_3, X_4) = 1$ for every consistently oriented local orthonormal basis $(X_1, \ldots, X_4)$.

Let r be a positive integer, T be an $(r, 0)$-tensor field on M. The *divergence* of T is that $(r-1, 0)$-tensor field div T characterized by:

$$\text{div } T(\psi^1, \ldots, \psi^{r-1}) = \sum_{\mu=1}^{4} (D_{X_\mu} T)(\psi^1, \ldots, \psi^{r-1}, \omega^\mu),$$

$\forall$ local basis $(X_1, \ldots, X_4)$ with dual basis $(\omega^1, \ldots, \omega^4)$ and for all 1-forms $\psi^1, \ldots, \psi^{r-1}$. Thus if T is antisymmetric, $i(\text{div } T)\Omega = d[i(T)\Omega]$, where i denotes the interior product (cf. [**18**]).

A *submanifold* (N, ψ) of (M, g) consists of a manifold N and an immersion $\psi\colon N \to M$. (N, ψ) is *spacelike* iff $\psi^* g$ is a Riemannian metric, iff $\psi_*(N_x)$ is spacelike $\forall x \in N$. When N is homeomorphically imbedded in M (§1.1) and ψ is the inclusion, we write simply: *submanifold* $N \subset M$. For the purpose of integration, we now standardize the meaning of a *compact submanifold with boundary* $N \subset M$: N is an oriented connected compact submanifold whose boundary ∂N is a piecewise smooth manifold with the orientation induced from that of N. If $\dim N = \dim M$, it is understood that the orientations in M and N are consistent.

5.2. Particle-flows. Imagine an enormous number of particles each of the

same rest-mass. Suppose, intuitively speaking, the particles are "streaming smoothly" with "no randomness" in their "velocity pattern". Examples are a very cold gas streaming in space (rest-mass nonzero) or a laser beam (rest-mass zero). Then the following is a useful idealization.

DEFINITION 5.2.1. A *particle-flow* (η, P) on spacetime M is a *particle density* η: $M \to [0, \infty)$ and an *energy-momentum* vector field P on M such that, for some fixed $m \in [0, \infty)$, each integral curve γ of P is a particle of rest-mass m. m is the *rest-mass of* (η, P).

Thus $g(P, P) = -m^2$ and P is future-directed (§4.4).

The main idea here is that η specifies, in a "smoothed-out" way, how many integral curves are actually occupied by particles. More explicitly, let $L \subset M$ be a 3-dimensional compact submanifold with boundary (§5.1) such that each integral curve of P intersects L at most once. Then one defines the *total number of particles for* ("in", "going through") L as $N \equiv |\int_L i(\eta P)\Omega|$, notation being as in 5.1. Thus $N \in [0, \infty)$. One does not insist that N be an integer $\forall L$ since when using a particle-flow model, one has in mind situations where $N \gg 1$ $\forall L$ of interest. Example 5.2.2 below might clarify η and N.

Let (η, P) be a particle-flow on M. Suppose that the particles in the flow do not interact with each other via quantum forces, that electromagnetic effects are negligible, and that there is no other kind of matter present. Then each particle in the flow should be freely-falling (§1.4) so one demands P be *geodesic*, i.e. $D_P P = 0$. One then further demands that particles be *conserved*, i.e. for any 4-dimensional compact submanifold with boundary $L \subset M$, $\int_{\partial L} i(\eta P)\Omega = 0$. By Stokes' theorem and a standard argument, this is equivalent to demanding $\operatorname{div}(\eta P) = 0$ (§5.1). Intuitively speaking, $\operatorname{div}(\eta P) = 0$ iff particles in the flow are nowhere created, e.g. from other kinds of particles, or destroyed, e.g. by decaying into other kinds of particles.

EXAMPLE 5.2.2. Let (M, g, ∂_4) be Minkowski spacetime with inertial frame ∂_4 (§4.1), and η: $M \to [0, \infty)$ be smooth. For $m \in (0, \infty)$, $(\eta, m\partial_4)$ is a particle-flow with energy-momentum vector field $P = m\partial_4$. Intuitively, the particles are at absolute rest relative to the inertial observers. By §5.1, $i(\eta P)\Omega = -\eta m du^1 \wedge du^2 \wedge du^3$. Thus one might call ηm "the number of particles per unit 3-volume", but η and N above are the more useful concepts and are applicable to the rest-mass zero cases as well. A short calculation shows that: $D_P P = 0$; moreover $\operatorname{div}(\eta P) = 0$ iff $\partial_4 \eta = 0$, iff $\eta = \eta(u^1, u^2, u^3)$, "iff the number of particles in each spatial 3-volume is fixed for all time".

When one assumes $D_P P = 0 = \operatorname{div}(\eta P)$, one can state in what sense initial data determine the future for a particle-flow on a given spacetime. Suppose the following data are given: $m \in [0, \infty)$; a spacelike 3-dimensional homeomorphically imbedded submanifold $N \subset M$; a smooth function η_0: $N \to [0, \infty)$ and a future-directed vector field P_0 defined on N such that $g(P_0, P_0) = -m^2$. Since N is spacelike, P_0 is nowhere tangent to N (2.2 and 4.4). By the above restrictions on N, we can find an open connected neighborhood U of N in M and a geodesic vector field P on U such that $P|_N = P_0$ and each inextendible integral curve of P in U intersects N exactly once (cf. [18]). $(U, g|_U)$ supplied with the induced orientation and time-orientation is then a spacetime.

PROPOSITION 5.2.3. *On U there is exactly one particle-flow (η, P) such that $\eta|_N = \eta_0$, $P|_N = P_0$, $D_P P = 0$ and* $\operatorname{div}(\eta P) = 0$.

The proof consists of first showing uniqueness of P, e.g. by using the geodesic spray, and then using standard results on first-order linear partial differential equations to analyze $\operatorname{div}(\eta P) = 0$.

5.3. Matter models and matter equations. A particle-flow (η, P) on M is an example of a matter model $\mathfrak{M}$ on M. $D_P P = 0$ and $\operatorname{div}(\eta P) = 0$ are examples of matter equations.

Matter models are the heart of any physical theory. Unfortunately there is no known, universal, overriding and precise macroscopic matter model from which all others follow as limits or special cases. Instead one works with many intuitively related, mathematically independent matter models; cf. §5.5, [14], [18], or [20]. In lieu of precise, general definitions of "matter models" and "matter equations", we make some general remarks and give some precise examples.

Matter equations model the influence of spacetime, electromagnetism and matter on matter. Sometimes, one can regard the spacetime M and an electromagnetic field (5.5) on M as given *a priori*. Then the matter equations become conditions on the matter model $\mathfrak{M}$ alone (cf. 5.2). But in general the situation is more subtle (§6.2). "Appropriate" matter equations, obtained by analyzing the physics in sufficient detail, always lead to some "present determines the future" theorem similar to Proposition 5.2.3. We now turn to the key example of a matter model.

EXAMPLE 5.3.1. Let $\mathfrak{M} = \{(\eta_A, P_A) | A = 1, \ldots, N\}$ be a finite collection of particle-flows on M; $\mathfrak{M}$ is a matter model. In the absence of all interactions other than gravity, appropriate matter equations for $(M, \mathfrak{M})$ are the following: $D_{P_A} P_A = 0 = \operatorname{div}(\eta_A P_A) \forall A$. These matter equations are interpreted via free-fall and particle number conservation as in §5.2.

EXAMPLE 5.3.2. Now suppose we have just a pair of particle-flows: $\mathfrak{M} = \{(\eta, P), (\hat{\eta}, \hat{P})\}$. Suppose $\hat{P} = P/2$ so that the rest-masses obey $\hat{m} = m/2$. Then it may be appropriate to replace $\operatorname{div}(\eta P) = 0 = \operatorname{div}(\hat{\eta}\hat{P})$ by the following matter equations: $\operatorname{div}(\eta P) = -k\eta = -2\operatorname{div}(\hat{\eta}\hat{P})$, where $k \in (0, \infty)$. Intuitively speaking, the rest-mass m particles are here decaying, e.g. by radioactivity, to make some extra rest-mass $\hat{m}$ particles. From the definition of total particle number, one can show that one here gets two of the latter for one of the former, corresponding to $\hat{m} = m/2$.

EXAMPLE 5.3.3. Though we shall not use it later, we give a more nearly generic example. Suppose one has a hot gas and each particle in the gas has rest-mass m, $m \in [0, \infty)$. Since the temperature is nonzero, there will be some "randomness" in the energy-momenta. To take this "randomness" into account, one might use a large number of particle-flows, $\mathfrak{M} = \{(\eta_A, P_A)\}$, each having the same rest-mass. But usually one smooths out as follows. Let $\mathcal{T}_m^+$ be the following subset of the tangent bundle: $\mathcal{T}_m^+ = \{(x, X) \in TM | g(X, X) = -m^2$, and X is future-directed$\}$. One can check that $\mathcal{T}_m^+$ is a smooth 7-dimensional submanifold of TM (cf. §2.2). Replace $\mathfrak{M} = \{(\eta_A, P_A)\}$ by a smooth function f: $\mathcal{T}_m^+ \to [0, \infty)$; the idea is $f(x, P_A) \leftrightarrow (\eta_A x)$. f is a matter model on M. Let L: $TM \to TTM$ be the geodesic spray [3]. On $\mathcal{T}_m^+$, L is tangent to $\mathcal{T}_m^+$ [18] so $df(L) = 0$ makes sense; this is equivalent to saying that restricted to the tangent vector field of each geodesic, f is a constant. This is called *Liouville's equation* and is a matter equation. Its interpretation

is that the gas is so dilute that collisions can be neglected despite the high temperature.

Other matter models, e.g. perfect fluids, will not be discussed in this article except insofar as they may be regarded as abstractions of the model in Example 5.3.1 when N becomes large.

5.4. Electromagnetism. Nonquantum relativistic electromagnetic theory is perhaps the most elegant part of physics. The formal postulates are simple and the practical applications are well-nigh endless. But we need it here only for background and any attempt to sketch the very rich physics involved would lengthen this article considerably. In this section, we thus proceed very formally, leaving almost all motivations, interpretations, Newtonian analogues, etc. to the references ([**14**], [**18**], [**20**]).

An *electromagnetic field* on spacetime M is a 2-form $\mathcal{F}$ on M. Throughout the rest of this section, $\mathcal{F}$ is an electromagnetic field on M, and $\hat{\mathcal{F}}$ is the physically equivalent (2, 0)-tensor field on M.

Let $\mathfrak{M}$ be a matter model (5.3) on M. Then $\mathfrak{M}$ determines a "charge-current density vector field" J on M (cf. [**14**], [**18**]); again, in place of a precise definition of J, we illustrate with a concrete example. Thus in Example 5.3.1, where one has a finite collection of particle-flows, suppose $\forall A$, (η_A, P_A) models particles all of which have a given electric charge $e_A \in \mathbf{R}$. Then $J = \sum_{A=1}^{N} e_A \eta_A P_A$ ("additivity of electric charge").

The triple $(M, \mathcal{F}, \mathfrak{M})$ *obeys Maxwell's equations* iff $d\mathcal{F} = 0$ and $\operatorname{div} \hat{\mathcal{F}} = 4\pi J$. The former equation replaces and unifies the classical equations

$$\vec{\nabla} \cdot \vec{B} = 0 = \vec{\nabla} \times \vec{E} + (1/c)\partial \vec{B}/\partial t,$$

while the latter equation does the same to Coulomb's law and the Biot-Savart-Maxwell law (see [**14**]). Since $\operatorname{div} \operatorname{div} = 0$, we have $\operatorname{div} J = 0$ ("conservation of electric charge"). In general, Maxwell's equations interrelate (M, g), $\mathcal{F}$ and $\mathfrak{M}$; they model the influence of matter and spacetime on electromagnetism.

Sometimes (M, g) and $\mathfrak{M}$ are given ab initio. Then $d\mathcal{F} = 0$ and $\operatorname{div} \hat{\mathcal{F}} = 4\pi J$ become conditions on $\mathcal{F}$ alone; we give an example. Let (M, g, ∂_4) be Minkowski spacetime with inertial frame ∂_4 (§4.1), and let $f\colon M \to \mathbf{R}$ be smooth. Then $\mathcal{F} = f du^1 \wedge du^2$ is an electromagnetic field on M. The triple $(M, \mathcal{F}$, no matter at all) obeys Maxwell's equation iff $d(f du^1 \wedge du^2) = 0 = \operatorname{div} \hat{\mathcal{F}}$. A computation shows that this is so iff $\partial_1 f = \partial_2 f = 0 = \partial_3 f = \partial_4 f$. Thus $\mathcal{F} = B_0 du^1 \wedge du^2$ for some $B_0 \in \mathbf{R}$. For $B_0 \neq 0$, this $\mathcal{F}$ can be interpreted by saying that each inertial observer $\partial_4 z$, $z \in M$, would measure a constant magnetic field of magnitude $|B_0|$, and zero electric field; cf. §4.1 and [**14**].

Let $\tilde{\mathcal{F}}$ be that (1, 1)-tensor field on M which obeys $\tilde{\mathcal{F}}(\omega, X) = \mathcal{F}(Y, X)$ whenever X, Y are vector fields and ω is the 1-form physically equivalent to Y. Let $\gamma\colon F \to M$ be a particle with electric charge $e \in \mathbf{R}$. Regarding $\tilde{\mathcal{F}}$ as a linear map $M_{\gamma u} \to M_{\gamma u}$ $\forall u \in F$, we obtain a vector field $\tilde{\mathcal{F}}\gamma_*$ along γ, i.e. $\tilde{\mathcal{F}}\gamma_*\colon F \to TM$ such that $\pi \circ \tilde{\mathcal{F}}\gamma_* = \gamma$. Similarly, with D the Levi-Civita connection, we have the curvature (geometric acceleration) $D_\gamma \gamma_*\colon F \to TM$ of γ. The triple $(M, \mathcal{F}, \gamma)$ is said to *obey the Lorentz force law* iff $e\tilde{\mathcal{F}}\gamma_* =$

$D_{\gamma_*}\gamma_*$. The Lorentz force law replaces Newton's $\vec{F} = m\vec{a}$ (in the specific form $\vec{E} + (1/c)\vec{v} \times \vec{B} = m\vec{a}$; cf. [14]). It models the influence of gravity and electromagnetism on particles. Thus $(M, \mathcal{F}, \gamma)$ is to obey the Lorentz force law whenever external quantum influences on γ can be neglected (§1.3). Suppose, in addition, $\mathcal{F} = 0$. Then only gravity acts on γ and we duly have free fall, i.e. $D_{\gamma_*}\gamma_* = 0$.

The tensor S defined by $S(X, Y) = \Sigma_{\mu=1}^{4} \tilde{\mathcal{F}}(\omega^\mu, X)\mathcal{F}(X_\mu, Y)$, $\forall$ vector fields, X, Y, and $\forall$ dual local bases $\{\omega^\mu\}$ and $\{X_\mu\}$, is a symmetric (0, 2)-tensor field on M. Thus trace S (§2.1) is a function and $T \equiv (1/8\pi)\{S - \frac{1}{4}(\text{trace } S)g\}$ is a symmetric (0, 2)-tensor field on M. T is defined as the *stress-energy density* of the electromagnetic field $\mathcal{F}$. A rather detailed physical interpretation and a brief motivation are given in §6.1.

5.5. Models for light. Suppose one has many light signals. Then no less than five models are available: photons (4.4); $m = 0$ particle-flows (5.3.1); the tangent bundle model 5.3.3 for $m = 0$; an electromagnetic field (5.4) or a "statistical superposition" of such fields; and a quantum electrodynamics model. We shall here use only the first two; compare the remarks on matter models in §5.3.

CHAPTER 6. THE EINSTEIN FIELD EQUATION

To complete the discussion of mutual influences in Chapter 5, we must discuss the influence of matter and electromagnetism on spacetime (M, g).

6.1. Stress-energy density. Einstein suggested spacetime is influenced by the stress-energy density of matter and electromagnetism. Now, formally, a stress-energy density on M is simply a symmetric (0, 2)-tensor field T on M. But physically, more is involved, as we now discuss.

6.1.1. *Pre-relativistic concepts.* Somewhat as energy-momentum unifies and replaces the pair (energy, momentum) in §4.4, stress-energy density unifies and replaces the following pre-relativistic quantities: Newtonian inertial mass per unit $\mathbf{R}^3$-volume; electromagnetic and Newtonian kinetic energy per unit $\mathbf{R}^3$-volume; electromagnetic momentum and Newtonian kinetic momentum per unit $\mathbf{R}^3$-volume; flux of energy and of Newtonian inertial mass; and momentum flux, which can be further split into a pressure and an anisotropic stress (cf. [14]). Hence "stress-energy" as short for, e.g., "mass-energy-momentum-pressure-stress". Around 1905, physicists noticed, with great glee, that all these are merely different aspects of one thing, in the sense of the following operational definition.

6.1.2. *Measuring stress-energy.* Let (z, Z) be an instantaneous observer (Definition 4.1.1). Suppose he actually measures the total energy of matter and electromagnetism in any unit volume of $Z^\perp$ (cf. §§4.2 and 6.1.3 below). He is then supposed to get $T(Z, Z)$, where T is some stress-energy density on M as formally defined above.

That the observations actually correspond to $T(Z, Z)$, rather than, for example, to some $S(Z, Z, Z)$ with S a continuous (0, 3)-tensor field, may be regarded as a basic law of nature (cf. §1.5). We have *uniqueness* in the following sense: Suppose T and T' both obey the formal defining conditions for stress-energy density given above. Suppose $T(Z, Z) = T'(Z, Z)$ for all instantaneous observers (z, Z). Then $T = T'$. Indeed, the assumption implies

that $T(X, X) = T'(X, X) \forall X$ which is timelike. Since timelike vectors form an open set of each tangent space while T and T' are symmetric, necessarily $T = T'$.

Now the operational definition 6.1.2 is as general and precise as anything else in nonquantum physics. But it refers to actual measurements and therefore cannot be used in formal proofs (cf.§1.5). But given a mathematically precise matter model $\mathfrak{M}$ and an electromagnetic field on M, the operational definition "leads to" a mathematically precise definition which can then be used in formal proofs. We now illustrate with the one example essential for later purposes. Let $\mathfrak{M} = \{(\eta_A, P_A) | A = 1, \ldots, N\}$ be a finite collection of particle-flows on M (Example 5.3.1) and suppose $\mathfrak{F} = 0$.

DEFINITION 6.1.3. The stress-energy density of $\mathfrak{M}$ is $T = \sum_{A=1}^{N} \eta_A \omega_A \otimes \omega_A$, where ω_A is the 1-form physically equivalent to P_A $\forall A = 1, \ldots, N$.

Motivation. T is a symmetric, (0, 2)-tensor field on M. We must examine its relation to the measurements of 6.1.2. Let (z, Z) be an instantaneous observer. $\forall A = 1, \ldots, N$, he measures energy–$g(P_A, Z)$ for a particle in the Ath particle-flow–by 4.4.1 and 5.2. Now suppose X_1, X_2, X_3 span $Z^{\perp}$. Let Ω be the Lorentzian volume form (5.1). Then the 3-volume of the parallelepiped $K \subset Z^{\perp}$ defined by $\{X_i\}$ is $|\Omega(X_1, X_2, X_3, Z)| > 0$. Moreover, by the definition and interpretation of number density η_A in §5.2, the Ath particle-flow contributes $N_A = |\Omega(X_1, X_2, X_3, \eta_A P_A)|$ particles in this parallelepiped, where we have also made use of the comments in §4.2. "Since energy is additive", the total energy in K is $E = -\sum_{A=1}^{N} N_A g(Z, P_A)$. The energy per unit 3-volume is $E/|\Omega(X_1, X_2, X_3, Z)|$. Algebra gives for the energy per unit 3-volume: $\sum_{A=1}^{N} (\eta_A)[g(P_A, Z)]^2$, which is duly independent of the particular parallelepiped K used. By uniqueness (6.1.2) and the definition of physical equivalence (2.4.3), we get $T = \sum_{A=1}^{N} \eta_A \omega_A \otimes \omega_A$, as was to be motivated. □

From the point of view of physics, the preceding pair, "Definition-Motivation", corresponds to "Theorem-Proof".

Now we are back in business mathematically and can exemplify by a theorem the key-properties which a stress-energy density enjoys in macrophysics. So let $\mathfrak{M} = \{(\eta_A, P_A)\}$ as before and $T = \sum_{A=1}^{N} \eta_A \omega_A \otimes \omega_A$ as in Definition 6.1.3. By §§2.1 and 5.2, trace $T = -\sum_{A=1}^{N} (m_A)^2 \eta_A : M \to (-\infty, 0]$, where $m_A \in [0, \infty)$ and is the rest-mass of the Ath particle-flow. Define the symmetric (0, 2)-tensor $W = T - \frac{1}{2}(\text{trace } T) g$; this tensor plays an important role in some of the "singularity theorems" (cf. §§7.3, 8.3 and [9]). Finally, let $\hat{T}$ be the (2, 0)-tensor field physically equivalent to T (2.4.3). The matter equations in (C) below were discussed in §§5.2 and 5.3.

THEOREM 6.1.4. (A) *For any instantaneous observer* (z, Z), $T(Z, Z) \geqslant 0$ *and* $W(Z, Z) \geqslant 0$, *where either equality holds for one* (z, Z) *iff* $Tz = 0$, *iff* $\eta_A z = 0$ $\forall A$. (B) Trace $T \leqslant 0$. (C) *Suppose* $D_{P_A} P_A = 0 = \text{div}(\eta_A P_A)$ $\forall A$; *then* div $\hat{T} = 0$.

The proofs of (A) and (B) consist of chasing down earlier definitions and using the algebraic results in §2.2. (C) follows from the linearity of div and

$$\text{div}(\eta P \otimes P) = \eta(D_P P) + [\text{div}(\eta P)] P,$$

which are easily verified using 5.1.

Most nonquantum matter models are so similar to the finite collection of particle-flows model above that one usually assumes that Theorem 6.1.4 generalizes "in the obvious way" when any other nonquantum matter model $\mathfrak{M}$ is used or $\mathfrak{F} \neq 0$. We give one more example and leave the rest to the references ([9], [18]).

Let $\mathfrak{M}$ be as in 5.3.1 and 6.1.3, but assume $\mathfrak{F} \neq 0$. Take $T = T_1 + T_2$, where T_1 is the stress-energy density of 6.1.3 and T_2 is that of $\mathfrak{F}$ (§5.4); we shall refer to this T as the *stress-energy density of* $\mathfrak{M}$ *and* $\mathfrak{F}$. Messy algebra shows that 6.1.4(A) remains valid provided we replace "iff $\eta_A z = 0$ $\forall A$" by "iff $\eta_A z = 0$ $\forall A$ and also $\mathfrak{F}z = 0$". Simple algebra shows that 6.1.4(B) remains valid. The equation div $\hat{T} = 0$ in 6.1.4(C) remains valid provided Maxwell's equations hold and, in addition, appropriate matter equations motivated by the Lorentz force law of §5.4 are used (cf. [14] or [18]).

Quite generally, one always demands div $\hat{T} = 0$ whenever the contribution from all forms of matter present and from electromagnetism have been included in T. In special relativity this is equivalent to postulating the very fundamental integral conservation laws for energy-momentum [14]. However, one there uses the fact that on Minkowski spacetime there are appropriate Killing vector fields in defining the relevant integrals. Thus, when gravity is not negligible, this motivation for div $\hat{T} = 0$ fails in an essential way. The Einstein field equation of the next section provides an alternative motivation. Theoretical motivations apart, one always does demand it on empirical grounds. This identity div $\hat{T} = 0$ is a powerful criterion for selecting appropriate matter equations. Popularizations to the contrary, it usually does not determine the full set of matter equations uniquely; 6.1.4(C) with $N > 1$ is one counterexample.

Because of 6.1.4(A) and its generalizations, one interprets $T = 0$ on an open submanifold $\mathfrak{U} \subset M$ to mean $\mathfrak{U}$ is a *vacuum*: no matter or electromagnetism at all in $\mathfrak{U}$ except perhaps "test" quantities which "respond but have negligible influence".

6.2. The Einstein field equation. Let (M, g) be a spacetime and Ric and s be the Ricci tensor and scalar curvature, respectively, as in §2.1. The *Einstein tensor of* M is $G \equiv \text{Ric} - (sg/2)$; it is a symmetric (0, 2)-tensor field on M.

Let $\mathfrak{M}$ be a matter model on M, $\mathfrak{F}$ be an electromagnetic field on M, and T be the stress-energy density of $\mathfrak{M}$ and $\mathfrak{F}$ (cf. the end of §6.1). The triple $(M, \mathfrak{F}, \mathfrak{M})$ *obeys the Einstein field equation* iff $G = T$. For example, suppose M is Minkowski spacetime, $\mathfrak{F} = 0$ and $\mathfrak{M}$ is a single particle-flow (η, P). By 6.1.4(A), $(M, 0, \mathfrak{M})$ obeys the Einstein field equation iff $\eta = 0$. The equation $G = T$ replaces Poisson's equation of Newtonian gravitational theory and indicates how matter and electromagnetism generate gravity. Historical, formal and empirical motivations for Einstein's field equation are given in great detail in every reference, e.g. [14] and [20]. Assuming the reader has already had enough of these, we henceforth focus on the actual content and implications instead. Suppose $(M, \mathfrak{F}, \mathfrak{M})$ obeys the Einstein field equation.

(A) Algebra gives trace $G = -s$. Thus Ric $= 0$ iff $G = 0$, iff $T = 0$, iff we have vacuum (§6.1).

(B) Assume, for the reasons outlined in the previous section, that T obeys the algebraic conditions in Theorem 6.1.4. Algebra shows that $G(X, X) \geqslant 0$

and $\mathrm{Ric}(X, X) \geqslant 0$ $\forall$ causal vector X. This inequality on Ric is perhaps the most important consequence of the Einstein field equation. Roughly, it states that on balance gravity tends to pull things together rather than push them apart. (The corresponding situation in Riemannian geometry is that nonnegative Ricci curvature "on the average" pulls geodesics together; cf. the proof of Myers' theorem.) It underlies the proofs of the singularity theorems in §§7.3 and 8.3.

(C) Let $\hat{T}$ be the (2, 0)-tensor field physically equivalent to T. Then the Bianchi identities for curvature imply $\operatorname{div} \hat{T} = 0$ (cf. [**14**]). Of course one has reason for postulating $\operatorname{div} \hat{T} = 0$ in any case (§6.1).

6.3. General relativistic models. A detailed, fully consistent general relativistic model consists of a triple $(M, \mathfrak{F}, \mathfrak{M})$ as in the last section such that:

(A) $(M, \mathfrak{F}, \mathfrak{M})$ obeys appropriate matter equations (§5.3).

(B) $(M, \mathfrak{F}, \mathfrak{M})$ obeys Maxwell's equations (§5.4).

(C) $(M, \mathfrak{F}, \mathfrak{M})$ obeys Einstein's field equation (§6.2).

Sometimes one can regard at least one member of the triple as given *ab initio*. For example, quantum complications apart, it is consistent to assume no matter anywhere and also take $\mathfrak{F} = 0$. Then Maxwell's equations and the matter equations are vacuous and the stress-energy density vanishes. Thus the problem collapses to finding a spacetime (M, g) such that $G = 0$. Such spacetimes exist; for example, a Kruskal spacetime has $\mathrm{Ric} = 0$, whence $G = 0$.

However, when one is looking for a model to fit an actual physical situation one often cannot regard $\mathfrak{M}$ and $\mathfrak{F}$ as fully given *ab initio*. §7.3 following has been set up as an example of how this actually works in practice. In such a case, one may need to use (A)–(C) jointly, together with empirical arguments, heuristic idealizations, etc. (cf. the discussion at the beginning of Chapter 7).

Aside. Is the Einstein field equation a condition on spacetime, a definition of T, a provable law of nature, or some other darn thing? The last. T has an independent definition (cf. 6.1). But one cannot assume T known *ab initio*, as discussed above. Laws of nature cannot be proved (cf.§1.5). Almost exactly the same question can be asked of any other basic physical law. For example, is Newton's $\vec{F} = m\vec{a}$ a definition of $\vec{F}$, a definition of Newtonian inertial frames, a provable law of nature, or some other darn thing? Really, the last.

CHAPTER 7. COSMOLOGY

Chapters 1–6 cover the basic ideas of nonquantum relativity. As a sample application, we discuss cosmology. The main purpose is to describe the universe. As a byproduct we will get an example of how one builds a model by interweaving the basic equations (§6.3) of macrophysics, empirical data and intuitive guesses. In practice, the construction of a cosmological model goes roughly like this: The relativist confronts the known data, makes a guess at a mathematical model that seems to be approximately consistent with the data, uses the model to analyze the data more closely, correspondingly refines his model, makes further adjustments if necessary when new data come in, etc. When new data do come in, a model may sometimes be completely discarded as beyond repair; often a model is kept but only with a specific qualification of its range of applicability. In the latter case, new models must

now be sought to accomodate data outside the range of applicability of the old model, and the whole zig-zag process begins anew.

The presentation of this chapter intentionally parallels this construction procedure as far as possible. Specifically, the reader will see that more than one model is needed. A mathematician eager to find the most general model applicable to the universe during all epochs from the big bang to the remote future will be disappointed, but in physics this tentative groping character of the models is a way of life.

The outline of this chapter is as follows. §7.1 discusses spatial isotropy. §7.2 outlines some observational results. Broadly speaking the data suggest two things: near here-now, the universe seems to be simpler than was thought likely ten years ago, so the classical cosmological spacetimes are probably better than more sophisticated modern alternatives; but there probably is a hot, dense, region in the history of the universe, "near the big bang", where rather sophisticated matter models are needed. Though the universe's apparent predilection for simple spacetimes and complicated matter strikes us as misguided, §7.3 presents some models which respect it. §7.4 outlines the relation of the models to observations; the last section concerns the early universe.

(M, g) is a spacetime throughout; $z \in M$ connotes "here-now" as always. Einstein-de Sitter spacetime (§3.1) is the canonical example. In verbal comments we shall for brevity sometimes take it for granted that there was a big bang and that the chronological distance from here-now to the big bang is well defined, as in 3.1.7 (cf. also §2.6), and is of order 10^{10} years.

7.1. Spatial isotropy. The concept of spatial isotropy plays a central role in cosmology, though only as an idealization. Intuitively, spatial isotropy means that any one spatial direction is on the same footing as any other. For example, suppose that when you look due East you see the Earth's sky as a certain shade of blue and when you look due West you see exactly the same shade. Thus, for East and West, you observe spatial isotropy. The intuitive concept is made precise in various ways. We give some examples.

EXAMPLE 7.1.1. Suppose (M, g) is Minkowski spacetime, $z \in M, f \in (0, \infty)$ and h is Planck's constant. Then $Y^{\pm} = hf(\partial_4 z \pm \partial_1 z)$ determines two photon energy-momenta at z (§4.4). The instantaneous observer $(z, \partial_4 z)$ measures frequency f for both (4.4.2). In this sense, $(z, \partial_4 z)$ observes "East-West" spatial isotropy.

7.1.2. *Pointwise spatial isotropy.* Let (z, Z) be an instantaneous observer in spacetime M. Let $O_Z(3)$ be the rotation group for $Z^{\perp} \subset M_z$, i.e.

$$O_Z(3) = \{\phi: M_z \to M_z | \phi Z = Z \text{ and } g(\phi X, \phi X) = g(X, X)\ \forall X \in M_z\}.$$

Then $O_Z(3)$ is isomorphic to $O(3)$ (= the rotation group in $\mathbf{R}^3$) and each $\phi \in O_Z(3)$ is linear. Now suppose f: $M_z \to \mathbf{R}$ is a function. f is (pointwise) *spatially isotropic for* (z, Z) iff $f(\phi X) = f(X)\ \forall X \in M_z$, and $\forall \phi \in O_Z(3)$. In particular, this definition applies to any symmetric $(0, r)$-tensor T at z, on regarding T as a function f with $f(X) = T(X, \ldots, X)$. The extension to other tensors at z is straightforward. For example, $X \in M_z$ is (pointwise) *spatially isotropic for* (z, Z) iff $\phi X = X\ \forall \phi \in O_Z(3)$, iff $X \in \text{span } Z$.

7.1.3. *Global spatial isotropy.* Let (x, X) be an instantaneous observer. Let

$\mathcal{G}M_X$ be the set of isometries of M (2.4.2) which leave (x, X) fixed, i.e. $\mathcal{G}M_X = \{\phi \in \mathcal{G}M | \phi x = x \text{ and } \phi_* X = X\}$. $\mathcal{G}M_X$ is a subgroup of $\mathcal{G}M$. Spacetime is *spatially isotropic for* (x, X) iff $\mathcal{G}M_X$ is isomorphic to $O(3)$. For example, Einstein-de Sitter spacetime is spatially isotropic for an instantaneous observer (x, X) iff $X = \partial_t x$ (cf. §3.1).

Let (M, g) be a spacetime such that, $\forall x \in M$, (M, g) is spatially isotropic for exactly one instantaneous observer (x, X). Let T be a $(0, r)$-tensor field on M. T is (globally) *spatially isotropic* iff $\phi^* T = T$ $\forall \phi \in \mathcal{G}M_X$ $\forall (x, X)$ as above. The extension of this concept to other tensor fields is again straightforward, e.g. a vector field Y is (globally) *spatially isotropic* iff $\phi_* Y = Y$ $\forall$ such ϕ.

Suppose T is a (0, 2)-tensor field on Einstein-de Sitter spacetime. An instructive computation shows that T is spatially isotropic iff $T = \mu(t) g + \nu(t) dt \otimes dt$, where μ, ν: $(0, \infty) \to \mathbf{R}$ are smooth; for example, such a T cannot be a nonzero 2-form. Let T be a spatially isotropic (0, 2)-tensor field on Einstein-de Sitter spacetime; then $\forall x \in M$, Tx is pointwise spatially isotropic for $(x, \partial_t x)$ (but the converse need not hold).

7.1.4. *Uniqueness.* Suppose T is as above, with ν nowhere zero. By algebra, one finds that T is spatially isotropic for (x, X) iff $X = \partial_t x$. Thus T singles out ∂_t as a distinguished vector field.

Generally speaking, spatial isotropy not only indicates some kind of intrinsic symmetry but also selects distinguished instantaneous observers. For example, in 7.1.1, $(z, (\cosh \beta)\partial_4 z + (\sinh \beta)\partial_1 z)$ is $\forall \beta \in \mathbf{R}$ an instantaneous observer who measures frequencies (see 4.4.2) $f^{\pm} = f(\cosh \beta \mp \sinh \beta)$ for the two photons. Thus $(z, \partial_4 z)$ is the only member of the family who measures "East-West" spatial isotropy. Similarly, in 7.1.2, one can check that f is spatially isotropic for at least one instantaneous observer (z, Z) iff there are functions μ, ν: $\mathbf{R} \to \mathbf{R}$ such that $f(X) = \mu g(X, X) + \nu g(X, Z)$ $\forall X$; then f is spatially isotropic for more than one instantaneous observer at z iff $\nu = 0$, iff f is spatially isotropic for every instantaneous observer at z, which is a very special case. Similarly, in 7.1.2, suppose $X \neq 0$ is given. Then X is spatially isotropic for at most one (z, Z), being spatially isotropic for exactly one iff X is timelike.

7.2. Observational cosmology. Most of the data relevant to cosmology is low precision, but so much is now available, owing mainly to work during the last decade, that at least the spacetime region near here-now seems to be reasonably well understood. We summarize the most important empirical results. [15], [17] and [20] contain more details on each topic discussed below.

In presenting data, astronomers normally use certain concepts naively: "observable universe", "spatial isotropy", "recession speed", etc. Given a detailed model, which is in any case sometimes needed to reduce the data systematically, each such term can be assigned a formal meaning. For example, if one assumes a spacetime without any piecewise smooth closed causal curves, "observable universe" can normally be taken to mean the causal past of here-now (as defined above Theorem 2.6.6). Since we don't yet have a detailed model, and can't write one down ad hoc without giving a misleading impression of how such models arise, we shall sometimes proceed naively in this section.

7.2.1. *Galaxies.* The observable universe contains about 10^{11} galaxies. Naively, imagine all these now distributed more or less uniformly within a big sphere having here as center and a radius of about 10^{10} (light-) years. There is some clumping. The biggest clumps of galaxies seem to be about 10^8 light years across and contain perhaps a million galaxies. Our own galaxy is part of a small local group which in turn is part of a big clump. The "random motion" of galaxies is rather small: a pair of nearest neighbors may have a relative speed of up to 0.005 the speed of light but usually the number is smaller. Thus we shall here ignore these random motions throughout. Quasars are probably just unusual galaxies and we shall regard them as such.

The most important physical property of an individual galaxy is beauty; the reader should look at some slides if he can. A typical galaxy has a rest-mass of the order of 10^6 seconds in our units (§4.3), and a diameter of perhaps 50,000 years. It contains several billion stars, some gas, some dust, and other constituents minor in the sense that their contribution to the total rest-mass is small. Whether a significant admixture of black holes is present is not known. Hydrogen is the predominant element. But about 30% of the rest-mass is in helium, which seems to be rather uniformly distributed, and traces of most elements are present.

7.2.2. *Actual observers.* Thus our own galaxy is very small compared to the observable universe. Moreover, the speed of the earth relative to the center of our galaxy is less than 0.001 the speed of light. Thus we can and shall idealize as follows. (z, Z) will denote an instantaneous observer on spacetime M, with z interpreted not only as here-now but also as an appropriate point on the history of the center of our galaxy; Z will be interpreted not only as tangent to the history of an actual telescope but also to the history of the center of our galaxy. We use (z, Z) or the heuristic phrase "actual observer" iff we have this interpretation in mind.

7.2.3. *Local physics there-then.* Let x represent a moderately distant-early point in the observable universe, e.g. x is $\frac{1}{2}$ or less of the way back in time toward the big bang in the sense of Figure 3.1.3. There is considerable evidence that the basic laws of local physics at x are the same as those at z; general relativity assumes this, as indicated by the fact that in stating the laws we have never referred to a distinguished spacetime point; we assume it throughout.

7.2.4. *The Hubble law and Hubble time.* Suppose an actual observer measures the wavelength λ_z of a photon from a distant galaxy as in 4.4.2. Using 7.2.3 one can usually infer what wavelength λ_x an instantaneous observer at $x \in M$ at rest with respect to the distant galaxy would have measured for that same photon at the emission event x. One systematically finds $\lambda_z > \lambda_x$. This is called a red shift since on the two ends of the visible spectrum, red light has longer wavelength than violet light. Formally, define the *red shift ratio* for the photon as $r = \lambda_z/\lambda_x$. We shall assume, and our models will predict, $r \in (1, \infty)$. In the references, $r - 1$ is called the *cosmological red shift*; we regard r as "directly measurable" (but compare §1.5). On a naive interpretation, $r > 1$ indicates that we and the distant galaxy are running away from each other (the Doppler effect, see [14]). As a temporary definition, call $v = (r^2 - 1)/(r^2 + 1)$ the *recession speed* of the galaxy. Thus $v \in (0, 1)$; for example, if one pretends spacetime is Minkowski spacetime one can motivate the term recession speed [18].

In a similarly naive way one can assign a distance L to the galaxy by observing its apparent brightness or the area it appears to subtend on the sky. For example, on a naive view, apparent brightness is proportional to {(actual brightness)/L^2}. Assuming the actual brightness known, e.g. by comparing with nearby galaxies and assuming 7.2.3, this acts as a temporary definition of L. Thus one can also assign a time T via $L = vT$; naively, T indicates how long ago we and the distant galaxy were right on top of each other.

The empirical *Hubble law* states that there is some single Hubble time $T_H \in (0, \infty)$ such that $T = T_H$, to good approximation, for all moderately distant-early galaxies. In particular, the pattern is spatially isotropic in the sense that T does not depend very much on the direction from which our photon comes. For technical reasons, the numerical value of T_H is less accurately known than the fact that there is just a single number involved. Current estimates give, roughly, $T_H = 1.5 \cdot 10^{10}$ years $\pm 20\%$. T_H^{-1} is known as the Hubble constant.

7.2.5. *Other time scales.* One can obtain time scales by other methods: radioactive dating of old rocks in the solar system; estimating the age of old stars in our galaxy; applying a dimensional argument to the observed stress-energy density discussed below; etc. Some of these measurements are very difficult and controversial. However, each gives a time of roughly 10^{10} years, so there is some kind of rough, overall consistency.

7.2.6. *The microwave photons.* We observe many photons with measured wavelengths between 0.1 and 10 cm. These are called microwave photons. They have three remarkable properties, whose discovery, interpretation, and implications have been the focus of attention in cosmology during the last decade. First, they do not come from identifiable discrete sources such as stars or galaxies. Probably the ones we see were created no later than 10^5 years after a big bang. In this sense observing them probably involves looking backward in time 99.999% or more of the way and thus also almost to the very edge of the observable universe; compare Figure 3.1.3.

Second, the observed pattern is spatially isotropic to an accuracy of considerably better than 0.1%. This counts as extremely high precision in cosmology. In view of the first property, it seems to indicate a surprisingly high uniformity of the whole observable universe.

Finally, the microwave photons have what is called a thermal spectrum. The term refers to a result found by Planck in the days before relativity. He considered a box containing gas molecules in complete thermal equilibrium at some temperature **T**. He pointed out that there must then also be photons in the box and discovered the following remarkable law, discussed in more detail in, e.g. [**20**]. Suppose an observer at rest with respect to the box measures the number $N \in [0, \infty)$ of photons whose measured energy (4.4) is greater than a given $E \in (0, \infty)$ in any unit measured 3-volume of the box. He finds a graph $N(E, \mathbf{T}) = \mathbf{T}^3 \int_{E/\mathbf{T}}^{\infty} b(u)\, du$, where $b(u)$ is a certain universal function independent of the kind of gas present, the size of the box, the temperature, etc.

Now when an actual observer measures the microwave photons and constructs the corresponding graph he finds, to good approximation, $N(E,$ 2.7° Kelvin). Near here-now there seems to be no photon source

sufficiently strong and close to thermal equilibrium to account for the fact that the observed graph has the characteristic thermal ("Planck", "black body") shape mentioned. However, big bang models can explain the shape in a reasonably plausible way; §7.4 will give an example. Thus the observed graph is generally regarded as the most nearly convincing of a number of observational results which indicate that something like a big bang actually occurred.

7.2.7. *Stress-energy density.* Recall that it is the stress-energy density T of matter and electromagnetism which governs their influence on spacetime (Chapter 6). With (z, Z) as in 7.2.2 the observed value is about $T(Z, Z) = (10^{10} \text{ years})^{-2}$ in our units, or probably rather less. Here it is understood that $T(Z, Z)$ has been "averaged over a very small spacetime volume, say 10^7 years across". The dominant contribution to the observed $T(Z, Z)$ comes from the rest-mass of the galaxies. The contribution of the microwave photons is about 10^{-4} of this galactic contribution. That of the macroscopic electromagnetic field is likewise negligible. However, there may be forms of matter, even near here-now, which cannot be directly detected at present even if they contribute significantly to $T(Z, Z)$.

7.2.8. *The cosmological reference frame.* Almost every current cosmological model postulates a distinguished future-directed timelike vector field, to be called the *cosmological reference frame.* For example, the vector field ∂_t in §3.1 is one such. The history of each galaxy is modeled by an inextendible integral curve of this vector field with two qualifications: (A) when the galaxy is regarded as an extended region, it is the history of its center which is modeled; and (B) for very early times, it is the history of the (nebulous) matter which will eventually form the galaxy that is modeled. The main motivation for postulating such a vector field is the spatial isotropy observed by an actual observer (z, Z) together with the uniqueness argument in 7.1.4 and the assumption that z = here-now is not very special. Other motivations include the possibility of an eigenvector characterization as in 3.1.2.

7.3. Basic cosmological models. We need a spacetime (M, g), a matter model $\mathfrak{M}$ on M and an electromagnetic field $\mathfrak{F}$ on M such that $(M, \mathfrak{F}, \mathfrak{M})$ approximates the history of our universe (§6.3).

7.3.1. *Strategy.* We will first make some assumptions on M and $\mathfrak{F}$ motivated by the data, leaving $\mathfrak{M}$ rather general. Then we work out the mathematical consequences, mainly by using Einstein's field equation. Third, we compare the resulting models to observations. Then we will have enough information to formulate more specific assumptions on $\mathfrak{M}$ and compare again to the data in 7.2. This section is devoted to the first two steps; the succeeding two sections deal with the third.

7.3.2. *Assumptions.* When large regions are concerned the overall influence of $\mathfrak{F}$ on M and $\mathfrak{M}$ seems to be negligible (7.2.7). Assume therefore: (α) $\mathfrak{F} = 0$. Thus we only need a model $(M, \mathfrak{M})$. Assume further: (β) (M, g) is maximal (§2.6).

We next make the assumptions: (γ) $\mathfrak{M}$ consists of a finite, possibly very large, collection of particle-flows (5.3.1). (This is very general; see §6.1.) (δ) Einstein's equation $G = T$ holds, where T is the stress-energy density of $\mathfrak{M}$ (6.1.3). It will be convenient to assume also: (ε) T is nowhere zero; this is

suggested, e.g. by Theorem 6.1.4 and by 7.2.7. At this stage, it would not be appropriate to try to specify exactly which particle-flows are in $\mathfrak{M}$ or specify the matter equations in detail (cf. 7.4.3(C) below).

Up to this point, our assumptions are too broad to give a sharp confrontation between observations and the model. We next make some very specific assumptions on (M, g). The main idea is to insist on an isometry group large enough to take into account the observed spatial isotropy (7.1 and 7.2). Other assumptions could be made at this point (3.1.8). All have some drawbacks (7.4.3 below).

Precisely, the remaining assumptions are: (ζ) $M = \mathbf{R}^3 \times F$, with $F \subset \mathbf{R}$, F open and connected. (η) $g = \mathfrak{R}^2(t)\rho^* h - dt \otimes dt$, where $\mathfrak{R}$: $F \to (0, \infty)$ is smooth, ρ: $M \to \mathbf{R}^3$ and t: $M \to F$ are the projections, and h is the Euclidean metric on $\mathbf{R}^3$. Denote the first and second derivatives of $\mathfrak{R}$ by $\dot{\mathfrak{R}}$ and $\ddot{\mathfrak{R}}$. Replacing $\mathfrak{R}(t)$ by $\mathfrak{R}(-t)$ if necessary, we may assume: (θ) $\dot{\mathfrak{R}}$ is somewhere nonnegative. In a moment, we will use the Einstein field equation to get information on F and $\mathfrak{R}$. But first orient M via $\rho^*(du^1 \wedge du^2 \wedge du^3) \wedge dt$, and time-orient (M, g) by defining a causal vector V to be future-directed iff $dt(V) > 0$. This time-orientation insures we don't get a model which is, intuitively speaking, everywhere contracting, in gross contradiction to the observations 7.2.4 (see 7.3.4 below).

One can show that, together with any standard matter equations, our above assumptions imply: (ι) F is not bounded from above. To avoid a detailed discussion of matter equations here we postulate (ι) separately. In any case the postulate concerns only the ultimate fate of the universe, which is (despite the enormous fuss made about it in popularizations) irrelevant to a discussion of the observable past of here-now.

Throughout the rest of this chapter, $(M, \mathfrak{M})$ will denote a *basic cosmological model*, i.e., one which satisfies all the assumptions (α)–(ι) above. We have finished the first step of our overall plan 7.3.1 and we next turn to the second.

PROPOSITION 7.3.3. *The Einstein tensor G of (M, g) is*

$$G = -\left(2\ddot{\mathfrak{R}}/\mathfrak{R}^3 + \dot{\mathfrak{R}}^2/\mathfrak{R}^4\right)\rho^* h + 3\left(\dot{\mathfrak{R}}/\mathfrak{R}\right)^2 dt \otimes dt,$$

where the notation is as in (ζ) and (η).

The proof is quite similar to that outlined in 3.1.2 and is thus omitted. The following corollary places an immediate restriction on $\mathfrak{R}$. For its purpose and for later purposes, let Z be the vector field on M physically equivalent to $-dt$; equivalently, Z satisfies $dt(Z) = 1$ and $\rho_* Z = 0$. Note that $g(Z, Z) = -1$ and Z is future-directed.

COROLLARY 7.3.4. *$\dot{\mathfrak{R}}$ is everywhere positive.*

PROOF. $3(\dot{\mathfrak{R}}/\mathfrak{R})^2 = G(Z, Z) = T(Z, Z) > 0$, by assumptions (γ)–(ε) and Theorem 6.1.4(A). Thus $\dot{\mathfrak{R}}$ is nowhere zero. By assumption (θ), $\dot{\mathfrak{R}} > 0$ everywhere. □

Recall that $g = \mathfrak{R}^2(t)\rho^* h - dt \otimes dt$. $\mathfrak{R}$ therefore measures the "spatial size" of g. Since the preceding corollary says $\mathfrak{R}$ is a strictly increasing function of t, one interprets this to mean our model must always expand; see 7.2.4. The next theorem states roughly that our model must also have a big

bang. For its statement, first define a future-directed causal geodesic γ: $I \to M$ to be *past-incomplete* iff its domain of definition I is bounded from below in $\mathbf{R}$.

THEOREM 7.3.5. *Each integral curve of Z is a future-directed timelike geodesic which is past-incomplete.*

If we regard $M = \mathbf{R}^3 \times F$ as a subset of $\mathbf{R}^4$, i.e. $M = \{(a, t) | a \in \mathbf{R}^3, t \in F\}$, then the integral curves of Z are just the t-coordinate curves oriented in the positive direction. Theorem 7.3.5 then implies that going in the negative direction along each t-coordinate curve must end in a "singularity" after a finite t-value. Thus according to assumptions (β) and (ι), *we may, and will, take $F = (0, \infty)$ without any loss of generality*. The hypothetical hypersurface $\mathbf{R} \times \{0\}$ (not in M) then corresponds to the big bang (cf. 3.1). The preceding theorem is an example of a singularity theorem under very special circumstances; a more general singularity theorem is given in §8.3.

PROOF OF THEOREM 7.3.5. In the above representation of M as a subset of $\mathbf{R}^4$, we may write

$$g = \mathfrak{R}^2(t) \sum_{i=1}^{3} du^i \otimes du^i - dt \otimes dt.$$

Each isometry $\sigma \in \mathcal{G}\mathbf{R}^3$ (cf. 2.4.2) then extends to an isometry $\tilde{\sigma} \in \mathcal{G}M$ by $\tilde{\sigma}(a, t) = (\sigma(a), t)\ \forall a \in \mathbf{R}^3$. If γ is the t-coordinate curve $t \to (a, t)$ for a fixed $a \in \mathbf{R}^3$, let G be the subgroup of $\mathcal{G}\mathbf{R}^3$ having exactly $\{a\}$ as the set of fixed points. Then γF is the set of fixed points of $\tilde{G} \equiv \{\tilde{\sigma} | \sigma \in G\}$. The fixed-point set of a group of isometries being a totally geodesic submanifold (cf. Kobayashi and Nomizu [**12**, II, p. 61]), γ is a geodesic. Equivalently, each integral curve γ of Z is a geodesic. The fact that γ is future-directed and timelike follows from the definitions.

To show that each integral curve γ of Z is past-incomplete, we first make a series of observations. The first one is a restatement of the preceding paragraph.

(i) $D_Z Z = 0, g(Z, Z) = -1$.

(ii) $g(D_V Z, W) = g(V, D_W Z)\ \forall$ vector fields V, W orthogonal to Z.

We have

$$g(D_V Z, W) = V[g(Z, W)] - g(Z, D_V W) = -g(Z, D_V W),$$

and similarly $g(V, D_W Z) = -g(D_W V, Z)$. Since $D_V W = D_W V + [W, V]$, it suffices to show $g(Z, [W, V]) = 0$. This is so because V, W are everywhere tangent to the hypersurfaces $\{t = \text{constant}\}$ and hence so is $[V, W]$. This implies $g(Z, [V, W]) = 0$.

(iii) $Z(\operatorname{div} Z) = -\operatorname{Ric}(Z, Z) - \frac{1}{3}(\operatorname{div} Z)^2$. (Special case of Raychaudhuri's equation.)

This is a straightforward computation using (i), (ii) and the following immediate consequences of the definitions:

$$\operatorname{div} Z = \sum_{i=1}^{3} g(D_{X_i} Z, X_i),$$

where $\{X_1, X_2, X_3, Z\}$ is locally an orthonormal basis (5.1, 2.1), and with the

same notation,

$$\mathrm{Ric}(Z, Z) = \sum_{i=1}^{3} -g\big(D_Z D_{X_i} Z + D_{[X_i, Z]} Z, X_i\big) \quad \text{(cf. 2.1)}.$$

(iv) $\mathrm{Ric}(Z, Z) \geqslant 0$.

Indeed, by assumptions (γ), (δ), Theorem 6.1.4(A) and remark (A) in 6.2,

$$0 \leqslant W(Z, Z) = T(Z, Z) - \tfrac{1}{2}(\text{trace } T)\, g(Z, Z)$$

$$= G(Z, Z) + \tfrac{1}{2}(\text{trace } G) = G(Z, Z) - \tfrac{1}{2}s.$$

By the definition of G, this gives

$$0 \leqslant \mathrm{Ric}(Z, Z) - \tfrac{1}{2}sg(Z, Z) - \tfrac{1}{2}s = \mathrm{Ric}(Z, Z).$$

(v) $\mathrm{div}\, Z > 0$.

As we observed in the proof of (iii), $\mathrm{div}\, Z = \Sigma_{i=1}^{3} g(D_{X_i} Z, X_i)$, where $\{X_1, X_2, X_3, Z\}$ is a local orthonormal basis. Using $g(X_i, X_i) = 1$, we get

$$g\big(D_{X_i} Z, X_i\big) = g\big(D_Z X_i + [X_i, Z], X_i\big)$$

$$= \tfrac{1}{2} Z g(X_i, X_i) + g\big([X_i, Z], X_i\big) = g\big([X_i, Z], X_i\big).$$

Thus $\mathrm{div}\, Z = \Sigma_i g([X_i, Z], X_i)$. Letting $\{\partial_i\}$ be the usual coordinate vector fields in $\mathbf{R}^3$, we choose $X_i = (1/\mathcal{R})\partial_i$, $i = 1, 2, 3$. It follows that $\mathrm{div}\, Z = 3\dot{\mathcal{R}}/\mathcal{R}$. By Corollary 7.3.4 and assumption (η), we see that $\mathrm{div}\, Z > 0$.

The proof of the theorem can now be simply completed. Let γ: $(-a, 0] \to M$ be an integral curve of Z. We have to show $-\infty < -a$. Let $f = (\mathrm{div}\, Z) \circ \gamma$; then f is a smooth function on $(-a, 0]$. (iii)–(v) now imply that

$$\dot{f} \leqslant -f^2/3, \qquad f > 0.$$

Let $b = f(0)$. Then we will prove that $(-3/b) \leqslant -a$. For, letting $g = 1/f$, we have $g > 0$ and $\dot{g} \geqslant \frac{1}{3}$ from the above. Now take $t \in (-a, 0]$; integrating this last inequality from t to 0 gives $(1/b) - (1/f(t)) \geqslant -t/3$. Thus $f(t) \geqslant 3b/|3 + bt|$. If $-a < -3/b$, then $f(-3/b)$ should be well defined. But the preceding inequality implies $f(-3/b) = +\infty$, contradiction. □

We make three observations about this proof: (A) $T(Z, Z) = \frac{1}{3}(\mathrm{div}\, Z)^2$. This follows from the proof of step (v), Proposition 7.3.3 and assumption (δ). (B) $\mathrm{div}\, Z \to \infty$ as one goes backward in time along the integral curves of Z. This follows from the last paragraph of the proof. (C) The instantaneous observer (z, Zz), as he goes backward in time along any integral curve of Z, will observe infinite total energy in finite proper time. This follows from 6.1.2 and (A), (B) above.

EXAMPLE 7.3.6 (THE EINSTEIN-DE SITTER MODEL). We give the key example. Suppose (M, g) is Einstein-de Sitter spacetime (§3.1), i.e. $\mathcal{R}(t) = t^{2/3}$. We take over the notation of 3.1 without comment. Let $\mathfrak{M}$ consist of a single particle-flow (η, P) with m = average rest-mass of a galaxy, $\eta = 4t^2/3m^2$: $M \to (0, \infty)$, and $P = m\partial_t$. Then (η, P) obeys the simple matter equations $\mathrm{div}(\eta P) = 0 = D_P P$, interpreted in §5.3. A short computation shows that $(M, \mathfrak{M})$ is in fact a basic cosmological model as defined above. $(M, \mathfrak{M})$ is called the Einstein-de Sitter model.

Conversely, let $(M, \mathfrak{M})$ be a basic cosmological model and suppose $\mathfrak{M}$

consists of a single particle-flow (η, P) with rest-mass m. We will show that $(M, \mathfrak{M})$ is essentially the Einstein-de Sitter model. Let ω be the 1-form physically equivalent to P. By assumption (δ) of 7.3.2, $G = \eta\omega \otimes \omega$ (see Definition 6.1.3). Comparison with Proposition 7.3.3 and elementary algebra gives $m^2\eta = 3(\dot{\mathcal{R}}/\mathcal{R})^2$, $2\ddot{\mathcal{R}}\mathcal{R} + \dot{\mathcal{R}}^2 = 0$, and $\chi = -mdt$. The first equation and Corollary 7.3.4 imply $m > 0$ and $\eta > 0$. The second equation is equivalent to $2(\ln \mathcal{R})^{\cdot} + (\ln \mathcal{R})^{\cdot} = 0$. Thus we obtain

$$\mathcal{R}(t) = kt^{2/3}, \quad \text{for some } k \in (0, \infty),$$

$$\eta = 4t^2/3m^2 \quad \text{and} \quad P = m\partial_t.$$

Upon choosing m as the average galaxy rest-mass, we have an isometry of $(M, \mathfrak{M})$ with the Einstein-de Sitter model. In this sense, *a basic cosmological model $(M, \mathfrak{M})$ is the Einstein-de Sitter model iff $\mathfrak{M}$ consists of a single particle-flow which models the galaxies.*

Apart from m above, which is irrelevant in many geometric arguments, the Einstein-de Sitter model has exactly one adjustable parameter, namely the time t assigned to z = here-now. Now, by §3.1, no star or rock at here-now can have a proper age greater than $t(z)$. On the basis of the data in 7.2.5, let us agree to take $t(z) = 10^{10}$ years, as a specific value, when using the Einstein-de Sitter model. Every other general-relativistic cosmological model has more adjustable parameters. Next to vague philosophy, gratuitous adjustable parameters are the biggest curse of theoretical cosmology and a really satisfying model should have none.

The next two corollaries of Proposition 7.3.3 apply to any basic cosmological model. The ideas involved have already been discussed in §§2.6 and 3.1 and 7.1.4; we omit the proofs. For the rest of this chapter, let ∂_t be the vector field physically equivalent to $-dt$; thus this is the same vector field as the Z of Theorem 7.3.5.

COROLLARY 7.3.7. *t is the cosmological distance from the big bang, i.e., $\forall y \in M$, $t(y) = 1.u.b.\{s(x, y) | x \ll y\}$ where s is the chronological distance.*

COROLLARY 7.3.8. *The isometry group $\mathcal{G}M$ is isomorphic to $\mathcal{G}\mathbf{R}^3$ ("spatial homogeneity and isotropy"). (M, g) is spatially isotropic for an instantaneous observer (x, X) iff $X = \partial_t x$.*

Thus we shall take ∂_t as the cosmological reference frame, with galaxy histories modeled by integral curves of Z (7.2.8). This is consistent with the Einstein-de Sitter matter model (7.3.6).

7.4. Confronting the data. As an example of how a basic cosmological model (M, g) can be compared to observations, consider the following experiment. The red shift ratio r of a photon which comes to us from the center of a moderately distant-early galaxy is measured as in 7.2.4. In addition, suppose the galaxy appears as a slightly extended haze in the sky, rather than merely as a point the way a star does. By direct observation, one can assign a solid angle $\Delta\Omega$ to the galaxy (the actual measurement can be very difficult, cf. [**20**]). Thus $\Delta\Omega \in (0, 4\pi)$ and corresponds to that portion of the sky under observation. Doing this experiment for many different galaxies, one obtains a graph $\{r, \Delta\Omega\} \leftrightarrow \Delta\Omega(r)$. The job of our model is to predict this graph.

To analyze the red shift ratio r, let λ: $[a, b] \to M$ be the observed photon, with $z = \lambda b$ and emission point $x = \lambda a \in M$. By §4.4, λ is future-directed and lightlike. One takes λ as a geodesic on the grounds that quantum interactions are either negligible or would annihilate the photon before it reaches us and electromagnetism is also negligible (both the electric charge and $\mathcal{F}$ are zero). The energy measured at emission is $-g(\partial_t x, \lambda_* a)$ and the measured energy at here-now is $-g(\partial_t z, \lambda_* b)$; these results follow from §§4.4 and 7.2.3 and our interpretation of ∂_t in the last section. By analyzing the geodesics, e.g. by using Killing vector fields as in 2.4.2, one can compute the red shift ratio to get:

7.4.1. $r = \mathcal{R}(t(z))/\mathcal{R}(t(x))$.

Since λ is future-directed and $\dot{\mathcal{R}}$ is everywhere positive (Corollary 7.3.4), we have $r \in (1, \infty)$, in qualitative agreement with the observations in 7.2.4 which give systematic red shifts, rather than any shift to violet. Note here that the more distant-early the emission point x, the bigger r. Textbooks thus refer to the measured quantity $(r - 1)$ as a "distance indicator".

The appropriate model for the measurement of $\Delta\Omega$ at here-now is the following. Let $\mathbb{S}^2$ be the *celestial sphere* in $Z^\perp$, i.e.

$$\mathbb{S}^2 = \{U \in Z^\perp \mid g(U, U) = 1\},$$

where $(z, Z) \equiv (z, \partial_t z)$ is an actual observer 7.2.2. Every photon from the distant galaxy determines a point U on the celestial sphere, $U =$ its direction, via the orthogonal decomposition $Y = E(Z - U)$ where Y is its energy-momentum at here-now. Then $\Delta\Omega$ is simply the 2-area of the interpolated set $\{U \in \mathbb{S}^2 \mid U$ as above$\}$.

Now let us assume the intrinsic 2-dimensional cross-sectional area ΔA of the distant-early galaxy is somehow known, e.g. by comparing with nearby galaxies and assuming 7.2.3; in practice this step can also be quite difficult. Let $\mathcal{G}_Z M$ be the group of isometries of M leaving (z, Z) fixed, as in 7.1.3. Let S^2 be the orbit of $\mathcal{G}_Z M$ through x (2.4.2). In the figure below, one dimension is suppressed. By the form of g, the homeomorphically imbedded 2-manifold S^2 has the intrinsic geometry of an ordinary 2-sphere. Let A be its intrinsic area. Since $\mathcal{G}_Z M$ sends lightlike geodesics through z into the same, the particular way S^2 corresponds to $\mathbb{S}^2$ then implies that $\Delta\Omega/(4\pi) = \Delta A/A$. If we can get A in terms of r, we are finished.

To take a specific example, suppose $(M, \mathcal{M})$ is the Einstein-de Sitter model (7.3.6) with $t(z) = 10^{10}$ years as before. Then we have the explicit formulas for lightlike geodesics λ as in 3.1.5. A computation gives

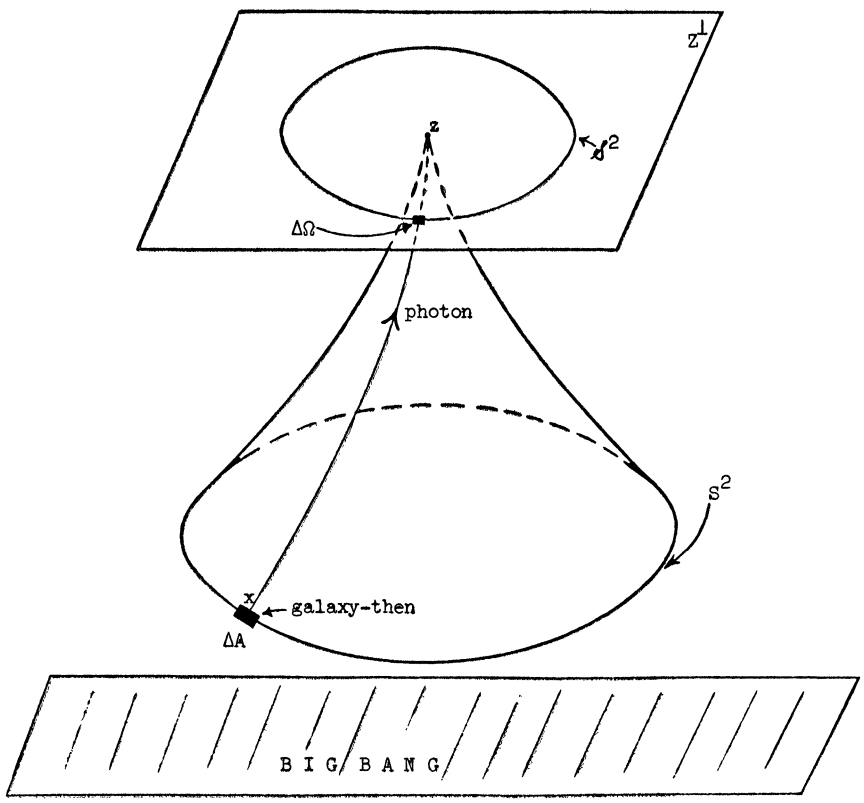
$$\begin{aligned} A &= 4\pi\left\{3(t(x))\left[\mathcal{R}^{1/2}(t(z)) - \mathcal{R}^{1/2}(t(x))\right]\right\}^2 \\ &= 4\pi\left[3t(z)(r^{-1} - r^{-3/2})\right]^2. \end{aligned}$$

Thus we have our desired prediction:

7.4.2. $\Delta\Omega = \Delta A/[3t(z)(r^{-1} - r^{-3/2})]^2$.

We make some comments on the precision of this equation relative to the data. (A) Naively, with ΔA and $t(z)$ fixed, one expects $\Delta\Omega$ to decrease as r increases. For $r < 9/4$, 7.4.2 duly predicts this ("the more distant-early a galaxy is, the smaller it appears"). But for $r > 9/4$, $\Delta\Omega$ is an increasing function of r. The intuitive interpretation then is that gravity is focusing the

light, so that the galaxy looks larger than it should. The actual observations do not extend in any convincing way out to the break point $r = \frac{9}{4}$. (B) For smaller values of r, the curve 7.4.2 fits the data about as well as the curve predicted by any other model. Neither the data nor the fit are high precision in any case except for $r - 1 \ll 1$. (C) In particular, working with a naive definition of distance as discussed in 7.2.4, one would here define distance L via $\Delta\Omega = \Delta A / L^2$. For $r - 1 \ll 1$, i.e. for rather nearby galaxies, comparison with 7.4.2 gives a prediction for the naively defined Hubble time 7.2.4. By using a Taylor expansion and 7.2.4, the reader can check that the predicted value of the Hubble time is $T_H = 1.5 \times 10^{10}$ years. The exact agreement with the measured value is spurious, in view of the observational uncertainties, but the agreement in the order of magnitude is gratifying.



7.4.3. *Others tests.* A number of similar classical tests can be applied to the basic cosmological models, as discussed in [**15**] and [**20**]. However, the one we have discussed is typical in the following ways. First the data is low precision and one does not get a high precision fit (of the model for the data). Second, the fit is about as good as for any other models even though all the other models have extra adjustable parameters. Third, on a qualitative level, the fit is really very good. Many different kinds of data can be fitted in without forcing or gross discrepancies. Then one has a simple overall context within which to analyze each particular observation. These models really are basic.

7.4.4. *Diseases.* However, the models also have serious diseases. (A) The universe has no exact symmetries; for example, there are fairly big clumps of galaxies (7.2.1). In our basic models, however, $\mathcal{G}M$ is nontrivial. Every explicitly known cosmological spacetime, general relativistic or not, suffers from this deadly disease. No real cure is known. To analyze more nearly generic models, one must use general theorems or linearization stability results (cf. [6]). Then one has so much leeway in the models that the comparison to observation becomes less tense and thus less interesting. The next disease is also almost universal and is extremely interesting. (B) Our models are self-defeating in the following sense: they predict a big bang (Theorem 7.3.5), but no matter how one chooses the matter model and matter equations one finds that sufficiently close to the big bang they become unrealistic. For example, the Einstein-de Sitter model is probably quite good near here-now and for, say, 90% of the time back to the big bang. But we shall see in the next section that the Einstein-de Sitter model is self-defeating, exactly in the sense just mentioned, for early times, e.g. for $t < 10^4$ years. Similarly, there is empirical evidence, discussed briefly in the next section, that the basic cosmological models may be qualitatively correct for quite early times, e.g. 1 second after the big bang. But if, as is frequently done nowadays, one starts speculating on much earlier times still, e.g. 10^{-23} seconds after the big bang (!), then the particle-flow model (7.3.2) itself breaks down. If there was such an epoch, quantum theoretical matter models must be used to analyze it.

This self-defeating disease is important because in trying to explain observed phenomena near here-now, e.g. the microwave photons and helium abundances mentioned in §7.2, one is sometimes driven back, willy nilly, to earlier epochs. It is probably not a mortal disease: hopefully many features near here-now are not too sensitive to the existence, let alone the details, of an ultra dense epoch. Moreover, there is a reasonably straightforward standard cure. The further back in time one has to go, the more one refines the matter model stepwise, using previous steps as a guide. Einstein-de Sitter is here the 0th step; the next section briefly discusses the first.

In addition to these essentially universal diseases, the basic models (7.3.2) have a special one: (C) They predict that the observed Hubble time T_H and observed energy density (7.2.7) must be related by $T(Z, Z) = (\frac{4}{3})T_H^2$ [17]. That the model interrelates these two different measurements is a virtue and the prediction is barely consistent with current data, but a smaller predicted value of $T(Z, Z)$ would fit the data somewhat better.

Finally, if one wants to make theoretical cosmology seem harder, deeper, and more accurate than it really is, the basic cosmological models are the worst possible models to use.

7.5. The early universe. Let (M, g) be a basic cosmological model. We briefly discuss here the microwave photons (7.2.6), the self-defeating disease 7.4.4(B), and the early universe.

By assumption (γ) in 7.3.2, we must model our microwave photons by a finite collection of particle-flows. Let (η, P) be one. Then P is lightlike future-directed and the rest-mass is zero (§§4.4 and 5.2). On the basis of the argument used for the single photon λ in the preceding section (cf. also §§5.3

and 5.4), we assume P geodesic as before; thus $D_P P = 0$. We shall here assume photons are conserved, i.e. $\operatorname{div}(\eta P) = 0$. Probably this is a reasonable assumption for times later than 10^5 years, though the arguments in support of this are rather difficult and to some extent controversial ([**15**], [**20**]). For much earlier times still, the matter equation $\operatorname{div}(\eta P) = 0$ in turn becomes self-defeating, but we shall not explicitly go back that far here.

Since P is lightlike, we cannot insist on spatial isotropy (7.1) for the individual particle-flow (η, P). However, since our large isometry group $\mathcal{G}M$ was in part motivated by the microwave photons we shall insist that (η, P) has *maximal symmetry*, i.e. there is a subgroup $\mathcal{K} \subset \mathcal{G}M$ such that $\eta \circ \phi = \eta$ and $\phi_* P = P\ \forall \phi \in \mathcal{K}$, with the dimension of $\mathcal{K}$ as large as possible. It turns out that the *relevant dimension is* 4. Moreover, suppose $N \in (0, \infty)$ and $E \in (0, \infty)$ are given; let r be the red shift ratio (7.4.1), i.e.

$$r = \left[\mathcal{R}(t(z))/\mathcal{R} \circ t\right]: M \to (0, \infty).$$

Define

$$\eta = (N/E)r^2: M \to (0, \infty),$$

$$P = Er\left(\partial_t + \left[1/\mathcal{R} \circ t\right]\partial_1\right): M \to TM.$$

Then (η, P) *is a photon-particle flow of maximal symmetry and obeys* $D_P P = 0 = \operatorname{div}(\eta P)$; *conversely, if* $(\bar{\eta}, \bar{P})$ *is a photon particle-flow of maximal symmetry which obeys these matter equations and* $\bar{\eta} > 0$, *then there exists an isometry* $\phi \in \mathcal{G}M$ *and positive numbers* N *and* P *such that* $\bar{\eta} = \eta \circ \phi$ *and* $\bar{P} = \phi_* P$, *where* η *and* P *are defined in terms of* N *and* E *as above.* Note that since $r = 1$ at here-now, by §§4.4 and 5.2, E is the energy measured at here-now for a photon in the particle-flow (η, P); similarly, by §5.2, N has the interpretation of number per unit 3-volume of $Z^{\perp}$ at here-now.

Now an observer at here-now measures many different energy-momenta, in particular, many different energies, for the various microwave photons (7.2.6). Thus we shall have to use many maximally symmetric particle-flows as above. Ideally, one would like to insist on global spatial isotropy (7.1.3) at least for the set as a whole. This cannot actually be achieved unless one uses an "infinite" number, i.e. uses the tangent bundle model of 5.3.3. However, it can be approximated to any desired degree of accuracy by using sufficiently many. That will suffice for present purposes and the exact number used will be irrelevant.

This model explains the observed thermal spectrum (7.2.6) as follows. Let (η, P) be a photon particle-flow with maximal symmetry, $N > 0$. Suppose $x \in M$ models an early point, i.e. $t(x) \ll 10^{10}$ years. Then a computation using the definition in §§4.4 and 5.2 shows that the number density N_x and energy E_x measured by $(x, \partial_t x)$ are $N_x = N(r(x))^3 > N$ and $E_x = Er(x) > E$. Suppose $(x, \partial_t x)$ measures many such flows. Suppose his graph for $N_x(E_x)$ does have the thermal form 7.2.6 with $\mathbf{T}_x \in (0, \infty)$ as the temperature. Since the situation near x is very dense (cf. the observations after the proof of Theorem 7.3.5), assuming a thermal shape at x is quite possible (cf. [**15**] and [**20**] for detailed arguments). Now our equations above show directly that at here-now we must also observe a thermal spectrum equal to

$$N(E) = r^{-3}N_x(E_x) = r^{-3}\mathbf{T}_x^3\int_{E_x/\mathbf{T}_x}^{\infty} b\,du = \mathbf{T}_z^3\int_{E/\mathbf{T}_z}^{\infty} b\,du$$

with $\mathbf{T}_z = \mathbf{T}_x/r < \mathbf{T}_x$.

[This famous and beautiful argument does not explain the particular value $\mathbf{T}_z = 2.7°\mathrm{K}$. By going back still further in time, and assuming the observed helium abundance (7.2.1) is due to the creation of helium out of hydrogen at early times, one can let a self-consistent explanation of the number 2.7°K for the temperature and the 30% figure quoted in 7.2.1 ([**15**], [**20**]). Since that argument involves times of order 1 to 1000 seconds after the big bang it is spectacularly correct if, as at present seems likely, it is correct.]

Finally, we analyze in what sense the Einstein-de Sitter model is self-defeating. Let $(M, \mathfrak{N})$ be a basic cosmological model. Let T_1 be the galactic contribution to the total stress-energy density T of §6.1, and let T_2 be the microwave photon contribution. By 7.2.7, we take $T_2(Z, Z) = 10^{-4}T_1(Z, Z)$, i.e. the photons are negligible near (z, Z) = here-now. Now suppose T_2 is spatially isotropic in the sense of 7.1.3. Since each particle-flow which contributes to T_2 has zero rest-mass we find from §§6.1 and 7.1.3 that $T_2 = (\mu \circ t)(g + 4dt \otimes dt)$ for some smooth function μ: $(0, \infty) \to [0, \infty)$; by using six or more particle-flows of maximal symmetry, one can construct a T_2 of this form. Using the Einstein-de Sitter matter equations for T_1 and our above matter equations for the photon particle-flows, one finds $\operatorname{div} \hat{T}_1 = 0 = \operatorname{div} \hat{T}_2$ by Theorem 6.1.4(C). Explicit integration of these two equations now gives $T_2(\partial_t, \partial_t) = 10^{-4}rT_1(\partial_t, \partial_t)$. Now $r \to \infty$ as $t \to 0$ because $r = t(z)^{2/3} \cdot t^{-2/3}$ in the case of Einstein-de Sitter. This means that for sufficient early times, $T_2(\partial_t, \partial_t) \gg T_1(\partial_t, \partial_t)$. This in turn means that the Einstein-de Sitter model, which neglects the effect on spacetime of the photon particle-flows entirely, becomes unrealistic at early times. The cure is merely to include the photon particle-flows. Near here-now, it makes no essential difference; for early times, it gives a more nearly realistic model. Thus we have given an example of disease 7.4.4(B) and its stepwise cure. The next steps are discussed, e.g. in [**15**] and [**20**].

Summary. The actual universe is beautiful. Some of that rubs off on the various models used in cosmology. But the models must be taken with a grain or more of salt.

CHAPTER 8. CHRONOLOGY, SINGULARITIES AND BLACK HOLES

In current research in general relativity, the most interesting topic conceptually is that of combining the theory with quantum physics. Equally important is finding detailed models, especially matter models, for a variety of observed astrophysical systems. Neither these nor many other physically motivated current investigations as yet lend themselves readily to description in reasonably precise mathematical terms. We will not discuss them.

Among relevant advanced mathematical topics, the study of the initial value problem and linearization stability is one of the most important. In [**6**], mathematicians should find discussions of these topics in terms that are accessible. Also very important is the theory of causal vector fields; cf. [**5**]. However, lightlike vector fields are so anti-intuitive for anyone trained in Riemannian geometry that we have here (somewhat artificially) avoided this theory whenever possible and shall continue to do so. Leaving aside a host of

minor topics, the third major mathematical area of current interest is chronology theory (cf. §2.6). This chapter discusses some examples of it. Many of the results are due to Geroch, Hawking and Penrose. [9] and [16] are the canonical references.

For reasons of space, the presentation of this chapter has been made more condensed than that of the preceding seven. As usual, (M, g) denotes a spacetime.

8.1. Mean curvature. Let $N \subset M$ be a 3-dimensional spacelike submanifold of M (§5.1). We define the *mean curvature* K: $N \to \mathbf{R}$ *of* N as follows. Given $x \in N$, let $\{X_1, X_2, X_3, Z\}$ be vector fields in a neighborhood $\mathcal{U}$ of x such that Z is future-directed, at each $y \in N \cap \mathcal{U}$ $\{X_1, X_2, X_3, Z\}$ is an orthonormal basis of N_y consistent with the orientation of M, and $\{X_1 y, X_2 y, X_3 y\}$ spans N_y. Then by definition,

$$Kx \equiv \left\{ \sum_{i=1}^{3} g(D_{X_i} Z, X_i) \right\}(x).$$

That this definition is independent of the choice of $\{X_1, X_2, X_3, Z\}$ is a standard calculation which we omit; this calculation is in fact implicit in the proof of step (ii) in the proof of Theorem 7.3.5. The following observation follows directly from the definition of the divergence of a vector field in 5.1 (see also the proof of step (iii) in the proof of Theorem 7.3.5): If Z is a unit vector field, future-directed and orthogonal to N at every point of N, then

$$K = \operatorname{div} Z|_N.$$

The readers familiar with Riemannian geometry will recognize K as the trace of the second fundamental form of N, and as such, this definition coincides with the Riemannian definition of mean curvature (cf. Kobayashi and Nomizu [12, II], p. 33). One has the following geometric interpretation of K. Suppose N is, in addition, a compact submanifold with boundary (§5.1). $\forall x \in N$, let γ_x: $[0, \varepsilon] \to M$ be a future-directed geodesic orthogonal to N and satisfying $\gamma_x 0 = x$ and $\|(\gamma_x)_*\| = 1$. Also $\forall t \in [0, \varepsilon]$, define $N^t \equiv \{\gamma_x(t) | x \in N\}$. If $v(t)$ denotes the Riemannian volume of N^t, then a straightforward computation gives

$$v'(0) = \int_N K \Omega_N,$$

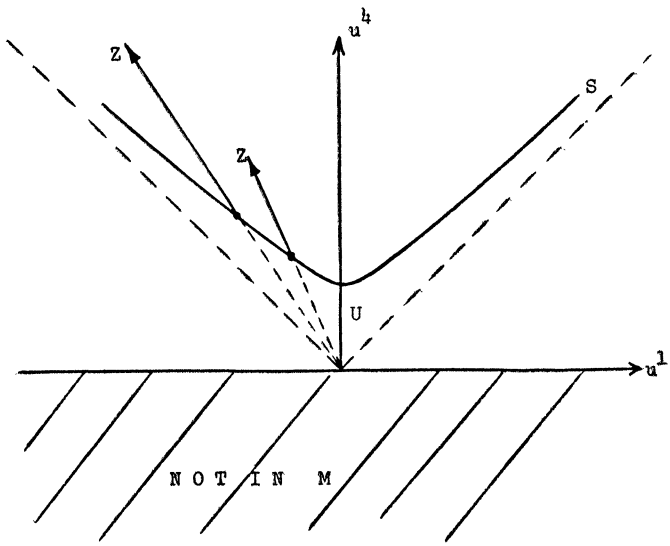
where Ω_N denotes the Riemannian volume element of N. Thus $K > 0$ roughly means that the future-directed geodesics orthogonal to N are, on the average, spreading out near N so as to increase the volume of N. The corresponding interpretation for $K < 0$ underlies the proofs of both singularity theorems of this article (Theorem 7.3.5 and Theorem 8.3.2 following).

EXAMPLE 8.1.1. Suppose $M = \mathbf{R}^3 \times (0, \infty)$ and $g = \sum_{i=1}^{3} du^i \otimes du^i - du^4 \otimes du^4$. With the usual orientation and time-orientation, (M, g) becomes the *upper-half space of Minkowski spacetime*. Define r: $\mathbf{R}^3 \to (0, \infty)$ by $r^2 = \sum_{i=1}^{3} (u^i)^2$ and write $t = u^4$: $M \to (0, \infty)$; then both are smooth positive functions on M. Let $S \subset M$ be the *unit hyperboloid* defined by $r^2 = t^2 - 1$. S is then a spacelike 3-dimensional submanifold of M. In the open neighborhood $U \equiv \{r^2 < t^2\}$ of S, define a vector field Z by $Z \equiv (t^2 - r^2)^{-1/2} \sum_{\mu=1}^{4} u^\mu \partial_\mu$. Z satisfies $g(Z, Z) = -1$, is future-directed, and $Z|_S$

is everywhere orthogonal to S. Moreover, div $Z = 3(t^2 - r^2)^{-1/2}$. Thus the mean curvature of S equals $K = \text{div } Z|_S = 3$.

Note that Z is just the restriction of the radial vector field of $\mathbf{R}^4$ to M. Consequently, its integral curves are just radial geodesics (= straight lines). From the geometric interpretation of K given above, we expect from $K = 3 > 0$ that in following the geodesic integral curves of Z in the past direction they should converge. Indeed, these geodesics converge at the origin (not in M), thereby giving rise to a "singularity"; compare the proofs of Theorems 7.3.5 and 8.3.2.

In the figure below, the orthogonality of Z to S might look peculiar; it is because g is Lorentzian.



8.2. Cauchy surfaces. The global structure of a spacetime (M, g) can be quite subtle. However, if there exists a Cauchy surface, as defined below, the global structure is not really more complicated than that of a Riemannian 4-manifold.

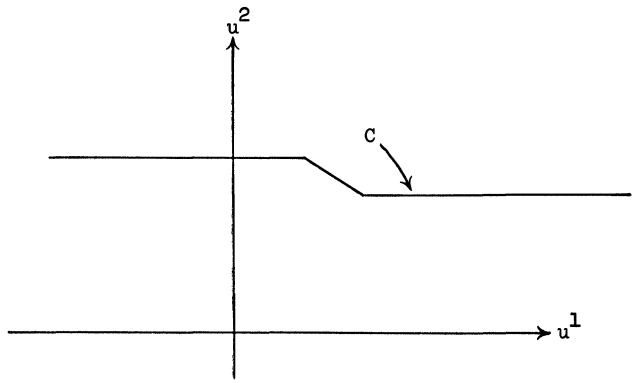
Let γ: $(a, b) \to M$ be a continuous curve, where $-\infty \leqslant a < b \leqslant \infty$. An $x \in M$ is an *upper endpoint* for γ iff $\lim_{u \to b} \gamma u = x$, i.e. $\forall$ neighborhood U of x in M, there exists a $u_0 \in (a, b)$ such that $\gamma(u_0, b) \subset U$. Intuitively: γ ultimately enters and stays in U. *Lower endpoints* are defined dually. γ is *endless* iff it has no endpoints.

EXAMPLES 8.2.1. (A) Let γ: $F \to M$ be an inextendible geodesic which is not a constant map. Then γ is endless; this follows from the existence of geodesically convex neighborhoods (Lemma 2.6.3). (B) For any timelike vector field X on M (which always exists by Proposition 2.3.1), the inextendible integral curves of X are endless. (C) The existence of upper or lower endpoints cannot be judged just by looking at the domain F. For example, suppose $M = N \times (0, \infty)$ with N a 3-manifold, $n \in N$, $F = (0, \infty)$, and $\gamma u = (n, \tanh u)$ $\forall u \in F$. Then γ has no lower endpoint and has an upper endpoint $(n, 1)$, although the opposite is true of the canonical imbedding $F \hookrightarrow \mathbf{R}$. (D) As always, causal curves (§2.3) play a special role. For example,

suppose γ: (0, 1) $\to M$ is smooth and the sequence $\gamma(\frac{1}{2}), \gamma(\frac{3}{4}), \ldots$ converges to $x \in M$. Then if γ is causal, a bounded variation argument shows x is an upper endpoint for γ; however, for a general γ, x need not be an upper endpoint.

Let $C \subset M$ be a subset. C is defined as a *Cauchy surface* iff each smooth endless timelike curve γ: $F \to M$ intersects C exactly once, i.e. there exists a unique $u_0 \in F$ such that $\gamma u_0 \in C$. This definition is conformally invariant in the sense of 2.6.5 since a curve is endless timelike for $(M, \alpha g)$ iff it is endless timelike for (M, g). A spacetime is *globally hyperbolic* iff there exists a Cauchy surface.

EXAMPLES 8.2.2. (A) Let (M, g) be the upper-half space of Minkowski spacetime (8.1.1). Define C as the level surface $t = 2$, $a > 0$. Then C is a Cauchy surface, as an elementary argument shows. (B) Minkowski spacetime, a Kruskal spacetime (§3.2), any basic cosmological spacetime (§7.3), and, in particular, Einstein-de Sitter spacetime (§3.1) are all globally hyperbolic. (C) In (A), the Cauchy surface is smooth and spacelike. Neither condition holds in general. For example, in the 2-dimensional Minkowski spacetime ($\mathbf{R}^2$, $du^1 \otimes du^1 - du^2 \otimes du^2$), the line indicated diagramatically below is a Cauchy surface C:



Proposition 8.2.3 below usually obviates the need to worry about Cauchy surfaces which are not smooth and spacelike. (D) Let $N \subset M$ be a 3-dimensional spacelike submanifold and i: $N \to M$ the natural injection. Then (N, i^*g) is a Riemannian manifold. Even if (N, i^*g) is complete and (M, g) is maximal, N need not be a Cauchy surface. For instance, let (M, g) be Minkowski spacetime with the origin deleted, and let N be the hyperplane $\{u^4 = 1\}$ in M. N is not a Cauchy surface in (M, g). Now choose a positive function α on M which is constant outside the Euclidean unit ball of $\mathbf{R}^4 - \{0\}$ $(= M)$ such that $(M, \alpha g)$ is (geodesically) complete. By the conformal invariance of a Cauchy surface, N is still not a Cauchy surface in the maximal spacetime $(M, \alpha g)$ although $(N, i^*(\alpha g))$ is a complete Riemannian manifold and M is maximal.

PROPOSITION 8.2.3. *Suppose M possesses a Cauchy surface. Then M possesses a Cauchy surface $N \subset M$ which is a connected spacelike 3-dimensional submanifold of M.*

This is one of the folk theorems of the subject. It is not difficult to prove that every Cauchy surface is in fact a Lipschitzian hypersurface in M [16]. However, to our knowledge, an elegant proof that his Lipschitzian submanifold can be smoothed out to such an N above is still missing.

Now let N be as in Proposition 8.2.3 and let $x \in M - N$.

THEOREM 8.2.4. (A) *There exists at least one smooth timelike curve* γ: $[0, 1] \to M$ *such that* $\gamma 0 = x$, $\gamma 1 \in N$, *and the arclength of* γ, $\int_0^1 \|\gamma_* u\|\, du$, *is at least as large as the arclength of any other smooth, timelike curve from* x *to* (*a point in*) N. (B) *This* γ *is a geodesic which hits* N *orthogonally*.

We recall explicitly that $\|\gamma_* u\| \equiv \{-g(\gamma_* u, \gamma_* u)\}^{1/2}$, so that the arclength of γ is positive. Proving part (A) requires considerable machinery. We here remark only that the main step consists of showing (M, g) is globally hyperbolic (i.e. possesses a Cauchy surface) iff it is globally hyperbolic in the sense of Leray [16]. Part (B) follows directly from part (A) together with the standard Riemannian argument and the wrong-way triangle inequality 2.2.3.

8.2.5. *Spacetimes which have no Cauchy surface.* To lend perspective, we mention some cases where no Cauchy surface exists. (A) Suppose there exists a piecewise smooth timelike curve γ: $F \to M$ which is closed. Then no Cauchy surface exists. For by a standard "rounding the corner" argument, one can smooth out γ to a smooth closed timelike curve $\hat{\gamma}$: $\mathbf{R} \to M$. Then $\hat{\gamma}$ intersects each set either infinitely many times or none at all. To get a concrete example, use the 2-dimensional time-orientable Lorentzian manifold in §2.3. (B) Amputate the origin from Minkowski spacetime as in 8.2.2 (D). (C) Amputate the half-space $u^1 \leqslant 0$ from Minkowski spacetime. Then by a conformal change of the metric as in 8.2.2(D), one can also get a geodesically complete example.

Very roughly speaking, these three examples indicate the three main ways in which a spacetime can fail to be globally hyperbolic: there are "causality violations or near causality violations" (case (A)); there are "gaps" (case (B)); or "infinity has the wrong shape" (case (C)).

8.3. A Singularity Theorem. To have a detailed example, we shall prove one of the simplest singularity theorems (Theorem 8.3.2); it is due primarily to Hawking. The main physical motivation for discussing the theorem has in effect been outlined in Chapter 7: one probably needs something like a big bang to account for the observational data of cosmology (§7.2ff.). The explicit models which predict a big bang (e.g. Theorem 7.3.5) are all somewhat suspect because such strong idealizations have been made (7.3.2). So one wants some general results. In addition, Theorem 8.3.2 gives some insight into the phenomenon of matter collapsing towards a singularity, e.g. a black hole; this follows by merely reversing the time-orientation below. However, it happens not to be among the more interesting theorems for this "time reversed" situation.

8.3.1. *Assumptions.* Let (M, g) be a spacetime throughout. Assume: (A) $\mathrm{Ric}(V, V) \geqslant 0$ $\forall$ causal vector V. The motivation has been discussed in considerable detail in §6.2. Almost any nonquantum matter model implies this condition via inequalities such as those in Theorem 6.1.4 on the total stress-energy density of matter and electromagnetism and the Einstein field

equation (§6.2). Thus, if assumption (A) breaks down, that in itself would presumably indicate the existence of a very hot, dense region, and this would be accomplishing the same purpose as a singularity theorem Note that assumption (A) is merely an inequality on Ric and not something more stringent, such as a partial differential equation.

Next we assume: (B) There exists a Cauchy surface for (M, g). Unfortunately, this may well be an unrealistically strong assumption for our actual universe. Thus many theorems use considerably weaker versions instead. But then one necessarily gets entangled in Lorentzian subtleties as in the next section. Making this assumption here will avoid lengthy explanations of such points. It follows from assumption (B) that there exists a connected smooth spacelike hypersurface $N \subset M$ which is a Cauchy surface (Proposition 8.2.3).

Finally assume: (C) For one such Cauchy surface N, there exists a $c > 0$ such that the mean curvature of N obeys $K \geqslant c$ (§8.1). The motivation is twofold. Almost all explicit cosmological models, in particular, all the basic cosmological models of §7.3, imply this. (To see this for the basic cosmological models, recall from the proof of step (v) in the proof of Theorem 7.3.5 that $\operatorname{div} Z = 3\dot{\mathfrak{R}}/\mathfrak{R}$. One checks that for each $b > 0$, $M_b \equiv \{t = b\}$ is a Cauchy surface. By 8.1, the mean curvature of M_b is $K = \operatorname{div} Z|_{M_b} = 3\dot{\mathfrak{R}}(b)/\mathfrak{R}(b)$. Thus we may let $c = 3\dot{\mathfrak{R}}(b)/\mathfrak{R}(b)$, by Corollary 7.3.4.) Second, we are again dealing merely with an inequality so the assumption is not too sensitive to detail.

To state the theorem, we fix a smooth connected spacelike Cauchy surface N whose mean curvature $K \geqslant c$ for some positive constant c (assumptions (B) and (C)). Recall from Theorem 7.3.5 that a future-directed causal geodesic $\gamma\colon F \to M$ is past-incomplete iff its domain of definition F is bounded from below in $\mathbf{R}$.

THEOREM 8.3.2. *Every future-directed timelike geodesic which hits N orthogonally is past-incomplete. In fact, if $\xi\colon (-u, 0) \to M$ is such a geodesic satisfying $\|\xi_*\| = 1$ and $\xi 0 \in N$, then $-u \geqslant -3/c$.*

REMARKS. (1) Of course one has in mind the case where M is maximal. (2) Unfortunately the theorem does not state that these geodesics are past-incomplete because when going backward in time, they would encounter infinite curvature or infinite energy density, etc. The canonical examples, e.g. basic cosmological models, do give such stronger statements (see the observations after the proof of Theorem 7.3.5). This lack of a detailed description of the precise nature of incompleteness for timelike geodesics is the main disease of all the general singularity theorems. There should be stronger results one can obtain without losing too much generality. (3) This theorem is atypical among singularity theorems in that it predicts the incompleteness of many inextendible time-like geodesics. In the more general theorems, one can only assert the existence of one incomplete inextendible timelike geodesic.

PROOF OF THEOREM 8.3.2. We begin with some preliminary material. Let $\gamma\colon [-a, 0] \to M$ be a future-directed timelike geodesic hitting N orthogonally such that $\gamma 0 \equiv x \in N$, and $\|\gamma_*\| = 1$; thus $g(\gamma_* 0, A) = 0\ \forall A \in N_x$. Define a vector space $\mathcal{V}$ of vector fields along γ by: $X \in \mathcal{V}$ iff (i) $Xt \in M_{\gamma t}$ $\forall t \in [-a, 0]$, (ii) $g(Xt, \gamma_* t) = 0\ \forall t \in [-a, 0]$, (iii) $X0 \in N_x$, and (iv)

$X(-a) = 0$. Note that, in particular, $\mathcal{V}$ consists only of spacelike vector fields along γ. For convenience, we introduce the curvature transformation R_{XY}: $\forall$ vector fields X, Y, Z, $R_{XY}Z$ is the unique vector field which satisfies $\omega(R_{XY}Z) = R(\omega, Z, X, Y)$ $\forall$ 1-forms ω, where the right-hand side is the curvature tensor as in §2.1. Now we define a quadratic function I: $\mathcal{V} \to \mathbf{R}$ by

$$I(X) = \int_{-a}^{0} \left\{ g\left(D_{\gamma_*}X, D_{\gamma_*}X\right)(t) - g\left(R_{X\gamma_*}\gamma_*, X\right)(t)\right\} dt + g\left(D_X X, \gamma_* 0\right),$$

where the last term is understood in the sense that if $\tilde{X}$ is any extension of $X0$ in the Cauchy surface N, then

$$g\left(D_X X, \gamma_* 0\right) \overset{\text{def}}{=} g\left(D_{\tilde{X}}\tilde{X}, \gamma_* 0\right).$$

This definition is independent of the particular extension so specified and depends solely on $X0$, as a calculation shows. In the Riemannian case, $I(X)$ is, of course, the quadratic form associated with the Morse index bilinear form on N-vector fields along γ (cf. Bishop and Crittendon [**2**, §11.2]; Kobayashi and Nomizu [**12**, II, pp. 71–88]; Hicks [**11**, §§10.0–10.2]).

Suppose $X \in \mathcal{V}$. Let τ: $[-a, 0] \times [0, \varepsilon] \to M$ be any smooth map such that: (1) γ coincides with the curve τ_0: $[-a, 0] \to M$ given by $\tau_0(t) = \tau(t, 0)$; (2) $\tau(0, s) \in N$ $\forall s \in [0, \varepsilon]$; (3) $\forall t \in [-a, 0]$, Xt is equal to the initial tangent vector of the curve ξ: $[0, \varepsilon] \to M$ defined by $\xi s = \tau(t, s)$; and (4) $\forall s \in [0, \varepsilon]$, the curve τ_s: $[-a, 0] \to M$ given by $\tau_s(t) = \tau(t, s)$ is timelike and future-directed. Then we call τ a *rectangle which induces* X. The last requirement (4) is not crucial because it can always be satisfied if we take ε sufficiently small. Then, just as in the Riemannian case, one proves:

(α) $\forall X \in \mathcal{V}$, there exists a rectangle which induces it.

Now let γ, X, τ be as above, and let l: $[0, \varepsilon] \to \mathbf{R}$ be the arclength function:

$$l(s) = \int_{-a}^{0} \left\|(\tau_s)_*(t)\right\| dt.$$

Then $l(0)$ = arclength of γ. Again, carrying over the Riemannian proof except for sign changes, we have:

(β) If τ is a rectangle which induces an $X \in \mathcal{V}$, then the second derivative of the arclength function at 0 is given by $\ddot{l}(0) = -I(X)$.

We can now give the proof of the theorem proper. Suppose ξ: $(-u, 0] \to M$ is a future-directed timelike geodesic which hits N orthogonally and satisfies $\|\xi_*\| = 1$ and $\xi 0 \in N$. Let $-v \in (-u, 0]$, and let $\xi(-v) = z \in M$. Theorem 8.2.3 implies that there is a future-directed timelike geodesic γ: $[-a, 0] \to M$ hitting N orthogonally such that $\gamma(-a) = z$, $\gamma 0 \in N$, $\|\gamma_*\| = 1$ and the arclength of γ maximizes those of all the smooth timelike curves from z to N. In particular, since the arclength of γ is a and the arclength of $\xi|_{[-v,0]}$ is v, we have

$$a \geqslant v.$$

We now carry over the notation of the initial segment of the proof. Let (Y_1, Y_2, Y_3) be parallel vector fields along γ (i.e. $D_{\gamma_*}Y_i = 0$ $\forall i$) such that $\forall t \in [-a, 0]$, $(Y_1 t, Y_2 t, Y_3 t, \gamma_* t)$ is an orthonormal basis of $M_{\gamma t}$. If $\{X_i\}$ are the vector fields along γ defined by $X_i t = ((a + t)/a) Y_i t$, $\forall t \in [-a, 0]$ and $\forall i = 1, 2, 3$, then each $X_i \in \mathcal{V}$. By (α) and (β) above and the choice of γ, we

have $I(X_i) \geqslant 0\ \forall i$. Since $D_{\gamma_*} X_i = (1/a)Y_i$, one obtains

$$0 \leqslant \sum_{i=1}^{3} I(X_i)$$

$$= \frac{3}{a} - \int_{-a}^{0} \left\{ \sum_i g\left(R_{X_i\gamma_*}\gamma_*, X_i\right)(t) \right\} dt + \sum_i g\left(D_{X_i}X_i, \gamma_* 0\right).$$

However, it follows from the definitions in §2.1 that

$$\sum_i g\left(R_{X_i\gamma_*}\gamma_*, X_i\right)(t) = \left(\frac{a+t}{a}\right)^2 \sum_i g\left(R_{Y_i\gamma_*}\gamma_*, Y_i\right)(t)$$

$$= ((a+t)/a)^2 \operatorname{Ric}\left(\gamma_* t, \gamma_* t\right) \geqslant 0,$$

where the last inequality is by assumption (A) on (M, g). Thus $0 \leqslant (3/a) + \Sigma_i g(D_{X_i}X_i, \gamma_* 0)$.

We now estimate the last sum. Note that $(X_1 0, X_2 0, X_3 0, \gamma_* 0)$ is an orthonormal basis of M_x which may be assumed to be consistent with the orientation of M, and $(X_1 0, X_2 0, X_3 0)$ spans N_x. Extend $(X_1 0, \ldots, \gamma_* 0)$ to vector fields (X_1', X_2', X_3', Z) in a neighborhood of x in M such that at each $y \in N$ inside the neighborhood, $(X_1' y, X_2' y, X_3' y, Zy)$ is an orthonormal basis of M_y and $N_y = \operatorname{span}\{X_i' y\}$. Thus, according to §8.1,

$$\sum_i g\left(D_{X_i}X_i, \gamma_* 0\right) = \sum_i g\left(D_{X_i'}X_i', \gamma_* 0\right) = -\sum_i g(X_i', D_{X_i}Z)(x) = -Kx,$$

where K is the mean curvature of N. By assumption (C), $-K \leqslant -c$. Thus $0 \leqslant (3/a) + \Sigma_i g(D_{X_i}X_i, \gamma_* 0) \leqslant (3/a) - c \Rightarrow -(3/c) \leqslant -a$. Together with the inequality $a \geqslant v$ above and the fact that $(-v)$ in $(-u, 0]$ is arbitrary, we have $-(3/c) \leqslant -u$. □

8.4. Black holes in general. The chronology relation of §2.6 is used in a number of ways: proving singularity theorems; defining and analyzing black holes; defining and analyzing spacetime boundaries with special reference to the concept of asymptotic flatness; discussing naked singularities; generating thesis problems for relativists who like global arguments; etc.

We now indicate how it is used in analyzing black holes. Again the attitude is that explicit spacetimes, such as those of §3.2, are very useful for detailed discussions but may not give adequate insight into the general case. For example, the assumption of spherical symmetry is very special, and the assumption of Ricci flatness means one cannot directly discuss collapsing matter at all (§6.2).

We thus start by discussing how the definition of the black hole region B in a Kruskal spacetime (§3.2) can be reworded so that it makes sense, formally, in any spacetime. Roughly speaking, the essential property of B that we shall use is the fact that light signals cannot escape from B to infinity.

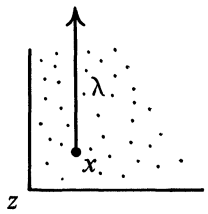
Let (M, g) be a spacetime with chronology relation $\ll$ (§2.6). We define the *chronological past of* a curve λ: $F \to M$ by $I^-\lambda = \cup_{u \in F}\{x \ll \lambda u\}$. Recall that a future-directed lightlike geodesic λ: $F \to M$ is *future-complete* iff its domain of definition F is not bounded from above in $\mathbf{R}$.

Define $C = \operatorname{Interior}(M - \cup_{\lambda \in \Lambda} I^-\lambda)$, where Λ is the collection of future-

directed, future-complete lightlike geodesics in M. Roughly, speaking, the definition means C is a region in which all light signals are trapped and short-lived. We give two examples.

In Minkowski spacetime, every inextendible geodesic is complete. It follows that C is empty.

On the other hand, in Kruskal spacetime (3.2), C is simply the black hole B. For example, consider a point z in the normal Schwarzschild submanifold N of Kruskal spacetime M. To show that z is not in C, refer to Figure 3.2.1. By confining attention to points and lines which have the same angles as z (i.e. have the same σ projection) we may regard A in the figure as a 2-dimensional submanifold of M and regard the projection αz as z itself. Choose an x in $\alpha I^+\{z\} \cap \alpha N$.



Consider the vertical line λ directed toward the top of the page, starting at x. λ is lightlike (§3.2). A short computation shows that, suitably parameterized, λ is a future-directed, future-complete lightlike geodesic. By construction $z \in I^-\lambda$. Thus z is not in C. Similar arguments show no point in N, W, or Q is in C. On the other hand, a separate computation shows that every future-directed lightlike geodesic which intersects B is future-incomplete. Since B is open, and contains the chronological future of each of its own points (§3.2), the definition of C now implies $B \subset C$. Since $M = B \cup \text{Closure}(N \cup W \cup Q)$ we thus have $B = C$ as claimed.

If one insisted on having available a definition of black hole region applicable to every spacetime, C above might be the most reasonable candidate. However, in physics one usually takes the view that the concept of a black hole should be introduced only for spacetime which "obey suitable causality conditions" and are "asymptotically flat" [9]. It is then also more appropriate to replace Λ in our definition above by a suitable subset of Λ, consisting, roughly speaking, of those lightlike geodesics which actually "escape to infinity" rather than merely being future-complete. The next two sections indicate some of the methods involved.

Aside. In the above definition, one can show

$$\text{Closure } C = \text{Closure}\left(M - \bigcup_{\lambda \in \Lambda} I^-\lambda\right);$$

in particular, C is nonempty if $M - \bigcup_{\lambda \in \Lambda} I^-\lambda$ is. In this sense, the fact that "Interior" has been inserted into the definition makes no essential difference; it is done for technical convenience.

8.5. Stable causality (cf. [9]). Empirically, there are no known violations of the chronology condition (2.6). Many physicists feel that one should impose the chronology condition, or some very similar requirement, as a basic physical requirement on all spacetimes considered. However, merely imposing

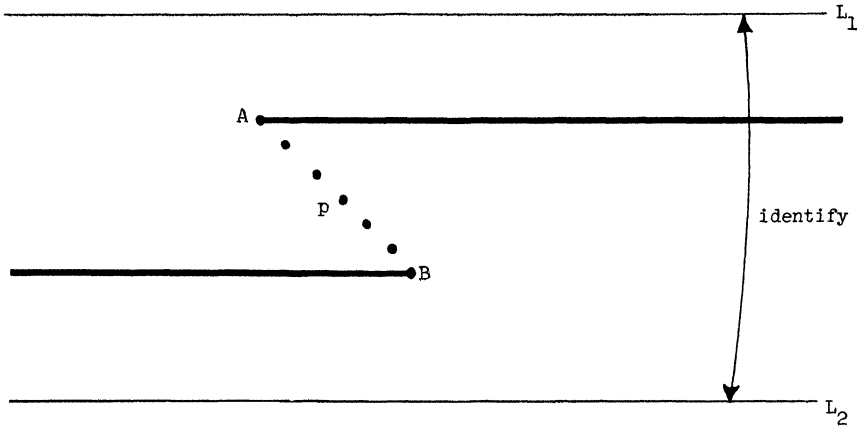
the chronology condition would still allow a certain marginal cases, such as that given in Example 8.5.2 below. We now define a more stringent restriction, which is generally regarded as being more plausible, mainly because it also excludes all such marginal cases.

Recall that (M, g) is a spacetime. A Lorentzian metric g_0 on M is defined as *wider* than g iff each vector W which is causal for (M, g) is timelike for (M, g_0), i.e. $g_0(W, W) < 0$ whenever $W \neq 0$ and $g(W, W) \leqslant 0$; intuitively, the lightcones for (M, g_0) are then wider than those for (M, g) (cf. §2.2). A *narrower* Lorentzian metric is defined dually.

8.5.1. REMARK. Given (M, g) there exists a spacetime (M, g_0) with g_0 wider than g. For let V be a vector field with $g(V, V) < 0$ everywhere (cf. 2.3.1). Let ω be the 1-form physically equivalent to V via g (2.4.3). Define $g_0 = g - \omega \otimes \omega$. By algebra (e.g. by using at a given point x a basis for M_x, orthonormal with respect to g, whose fourth vector is proportional to V) one finds g_0 is a Lorentzian metric on M. Moreover, if W is causal with respect to g, $\omega(W) \neq 0$ (§2.2); it follows that g_0 is wider than g. $g_0(V, V) < 0$ so (M, g_0) is time-orientable (§2.3). Time orienting gives the required spacetime (M, g_0).

(M, g) is defined as *stably causal* iff there is a spacetime (M, g_0) which obeys the chronology condition (2.6), with g_0 wider then g. Thus a stably causal spacetime obeys the chronology condition. But, intuitively speaking, the point is that, in addition, any "sufficiently small" perturbation of g in a stably causal spacetime will lead to a Lorentzian metric which is also narrower than g_0 and thus to a spacetime which is again stably causal; in this sense the restriction of stable causality has an agreeable stability which analogous restrictions lack. An example may clarify the point.

EXAMPLE 8.5.2. On the 2-dimensional version of Minkowski spacetime, $(R^2, du^1 \otimes du^1 - du^2 \otimes du^2)$, let L_1 and L_2 be two infinite horizontal straight lines. Take two points A and B between L_1 and L_2 lying on a lightlike geodesic (= straight line with a 45° slant). We construct a new (2-dimensional) spacetime by identifying L_1 and L_2, and further deleting the closed semi-infinite line segments emanating from A and B, as shown:



The resulting 2-dimensional spacetime obeys the chronology condition. But if g_0 is any Lorentzian metric on the manifold such that g_0 is wider than g, then there are closed curves through p, timelike with respect to g_0. Thus the 2-dimensional spacetime is not stably causal.

THEOREM 8.5.3. *(M, g) is stably causal iff there is a smooth, real valued function h on M such that dh is everywhere timelike.*

In one direction the proof is simple. Suppose such an h exists. Take ω in 8.5.1 to be $\omega = dh$ and let V be physically equivalent to ω as before. Then (M, g_0) is a spacetime, with g_0 wider than g. Algebra shows dh is timelike with respect to g_0. Thus if γ is a curve in M such that γ is timelike with respect to g_0, then $dh(\gamma_*) = \omega(\gamma_*) \neq 0$ (cf. §2.2). This implies γ is one-one. Thus (M, g_0) obeys the chronology condition. Thus (M, g) is stably causal. For the idea of the (harder) proof of the converse, see [9].

Any function h which obeys the conditions of Theorem 8.5.3 is called a *global time function*. Usually, when discussing black holes, one assumes spacetime is stably causal. Theorem 8.5.3 indicates one way in which stable causality can then be applied: even though h is not canonically determined, the existence of at least one global time function means it makes sense to speak of a black hole being created at some time: e.g. one could assume that C in §8.4 is nonempty but that $h|_C > t_0$ for some t_0 in the image of h ("at times earlier than t_0, there was no black hole"). In the absence of a global time function such concepts need not make sense, as Example 8.5.2 perhaps suggests. We now turn to another way in which the assumption of stable causality can be used when analyzing black holes.

Aside. A globally hyperbolic spacetime is stably causal but the converse need not hold.

8.6. Causal boundaries. We now indicate roughly how, in a stably causal spacetime, one can replace Λ in the construction of §8.4 by a more appropriate subset of Λ, consisting of those lightlike geodesics which, physically speaking, actually escape to infinity. The method to be used also gives some idea of current activities in attaching boundaries to spacetimes. One main idea is to assign a future "ideal endpoint" to each endless causal curve, thereby grouping such curves into the equivalence classes of curves which have the same ideal endpoint. We start with a simple observation about the chronological pasts of points of an arbitrary spacetime (M, g) (cf. §2.6 for terminology and notation).

LEMMA 8.6.1. *Suppose $z \in M$. Then $I^-\{z\}$ has the three properties*: (A) *it is nonempty*; (B) *if $y \in I^-\{z\}$, then $I^-\{y\} \subset I^-\{z\}$*; (C) *if $x, w \in I^-\{z\}$, then there exists $y \in I^-\{z\}$ such that both $x, w \in I^-\{y\}$.*

(B) is equivalent to the transitivity of $\ll$, and (C) is proved by noting that $I^+\{x\} \cap I^+\{w\}$ is an open neighborhood of z (cf. Theorem 2.6.2 and its proof). The lemma suggests that we investigate the following subset $\overline{M}$ of the power set of M: $P \in \overline{M}$ iff

(A) P is nonempty.

(B) $y \in P \Rightarrow I^-\{y\} \subset P$.

(C) $x, w \in P \Rightarrow$ there exists $y \in P$ such that both $x, w \in I^-\{y\}$.

From Theorem 2.6.2 and (B) and (C), we know that $\overline{M}$ consists of open sets. Moreover, Lemma 8.6.1 implies that $I^-\{z\} \in \overline{M}\ \forall z \in M$. We therefore have a natural mapping $\zeta\colon M \to \overline{M}$ defined by $\zeta z = I^-\{z\}\ \forall z \in M$.

Throughout the remainder of this section, we assume (M, g) is stably

causal. The relevant consequence is that then the mapping ζ above is one-one. Indeed, suppose $I^-\{q\} = I^-\{p\}$. Then $p \in$ Closure $I^-\{q\}$ and $q \in$ Closure $I^-\{p\}$ (§2.6). If $p \neq q$ this clearly implies $p \ll q$ and $q \ll p$ for any spacetime (M, g_0) with g_0 wider than g; thus assuming $p \neq q$ contradicts the fact that (M, g) is stably causal. *Thus ζ is one-one and we shall agree henceforth to identify M with $\zeta M \subset \overline{M}$.* We define the *future causal boundary* of M to be $M^+ \equiv \overline{M} - M$. M^+ is intuitively the ultimate future of M and is nonempty (cf. Example 8.6.2 below). Thus a consequence of stable causality is that it allows M to be naturally imbedded in a bigger space with a future boundary.

We now define a *causal structure* on $\overline{M}$. $\forall Q \in \overline{M}$, we shall define its causal past $J^-\{Q\}$ and its chronological part $I^-\{Q\}$ (see end of §2.6):

$P \in J^-\{Q\}$ iff $P \subset Q$,

$P \in I^-\{Q\}$ iff there exists $x \in Q$ such that $P \subset I^-\{x\}$.

One can check that each $I^-\{Q\} \subset \overline{M}$ enjoys the three properties of Lemma 8.6.1, and that $I^-\{Q\} \subset J^-\{Q\}$ $\forall Q \in \overline{M}$. Moreover, this chronology relation, i.e. $P \ll Q$ iff $P \in I^-\{Q\}$, extends that of M; in other words, $\forall x, y \in M$, $x \ll y$ in M iff $x \ll y$ in $\overline{M}$. We shall also need the following characterization of M^+ before we can outline the standard definition of a black hole. For a proof, see [9].

LEMMA 8.6.2. *$P \in M^+$ iff $P = I^-\gamma$ for some smooth future-directed causal curve γ in M which has no upper endpoint* (*cf.* 8.2).

To pin down the various concepts introduced above, we give a simple example.

EXAMPLE 8.6.3. Consider the 2-dimensional version of Minkowski spacetime $(N, h) \equiv (\mathbf{R}^2, du^1 \otimes du^1 - du^2 \otimes du^2)$. Let $a \in \mathbf{R}$ be arbitrary; using Lemma 8.6.2 one can show that N^+ consists of the following elements:

$\mathbf{R}^2$ $(= I^-\gamma$, where γ is any complete future-directed timelike geodesic).

$R_a \equiv \{(u^1, u^2) | u^2 - u^1 < a\}$ $(= I^-\lambda_1$, where λ_1 is the complete lightlike geodesic $u \to (u, u + a)$ $\forall u \in \mathbf{R})$.

$L_a \equiv \{(u^1, u^2) | u^2 + u^1 < a\}$ $(= I^-\lambda_2$, where λ_2 is the complete lightlike geodesic $u \to (-u, u + a)$ $\forall u \in \mathbf{R})$.

Note that, with λ_1 as above, $I^-\lambda_1$ $(= R_a)$ also equals $I^-\gamma_1$, where γ_1 is the endless future-directed timelike curve $u \to (u, u - e^{-u} + a)$ $\forall u \in \mathbf{R}$. Thus the causal character of the γ guaranteed by Lemma 8.6.2 is not unique.

One sees directly from the definition that $\overline{N}$ is in one-one correspondence with the causal past $J^-\{0\}$ of the origin 0 in N, i.e.

$$J^-\{0\} = \left\{(u^1, u^2) | u^2 \leqslant 0, (u^2)^2 - (u^1)^2 \geqslant 0\right\}.$$

The tip of this solid cone $J^-\{0\}$ corresponds to $\mathbf{R}^2 \in N^+$, and the rest of the boundary of $J^-\{0\}$ corresponds to the future lightlike infinity of N, i.e. the points R_a and L_a, $\forall a \in \mathbf{R}$, of N^+. Moreover this correspondence between the boundary of $J^-\{0\}$ and N^+ preserves the causal structure, in the sense that, just as in $J^-\{0\}$, every point of N^+ is in $J^-\{\mathbf{R}^2\}$ but not in $I^-\{\mathbf{R}^2\}$, and if $a \leqslant b$, $R_a \in J^-\{R_b\} - I^-\{R_b\}$ (resp. $L_a \in J^-\{L_b\} - I^-\{L_b\}$).

Returning to the case of a general stably causal spacetime we remark that, given M^+, it is possible to identify which points of M^+, if any, are genuinely

at infinity. A necessary condition for $P \in M^+$ to be at infinity is that $P = I^-\lambda$ for some future-complete lightlike geodesic λ. In our above Example 8.6.1, this is also sufficient, so infinity consists of all of M^+ with the exception of the single element $\mathbf{R}^2 \in M^+$. For necessary and sufficient conditions in the general case, cf. [9].

We thus get the desired subset Λ' of Λ mentioned in §8.4, namely $\Lambda' = \{\lambda$: λ is a future-directed, future-complete lightlike geodesic and $P = I^-\lambda$ for some $P \in M^+$ which is at infinity$\}$. We thus also arrive at the most nearly standard definition of a black hole in a stably causal spacetime: there is a black hole iff $M - \bigcup_{\lambda\in\Lambda'} I^-\lambda$ is nonempty and then the black hole region B in M is the interior of $M - \bigcup_{\lambda\in\Lambda'} I^-\lambda$. For the case of a Kruskal spacetime this general definition again coincides with the definition of §3.2.

Under appropriate restrictions, the most important of which are the Einstein field equation (6.2) and inequalities on the stress-energy density appropriate for nonquantum matter (cf. Theorem 6.1.4), one can prove a number of theorems on black holes as defined above [9].

For example one can prove a result interpreted as saying that, in the absence of quantum effects, a black hole cannot disappear. More specifically the area is a nondecreasing function of time in the following sense. Let h be any global time function, S_a be the level surface $h = a$, Σ_a be a connected component of the intersection of Boundary B with S_a; here we assume $B \cap S_a$ is nonempty. One shows Σ_a is a C^0 2-submanifold with a well-defined 2-dimensional area A_a. Moreover if $b > a$ and b is in the image of h then there is a corresponding connected component Σ_b in the level surface $h = b$ and the area A_b of Σ_b obeys $A_b \geqslant A_a$. For the detailed statement and proof see [9]. In our example, Kruskal spacetime (3.2), A_a is $4\pi\mu^2$, independent of h and of a, as is perhaps plausible from the fact that Boundary B contains integral curves of a future-directed Killing vector field (K in §3.2) and can be checked by a direct computation.

Similarly, one can prove a theorem interpreted as saying that, in the absence of quantum effects, a black hole can never bifurcate, though coalescence with another black hole is allowed [9].

CHAPTER 9. CONCLUSION.

Einstein's theory has severe limitations. In its current form, it is at best only a nonquantum approximation. Even within macrophysics, only its Newtonian limit and the nonquantum special relativity subcase have the kind of overwhelming empirical support one demands of a fundamental theory. As with other current physical theories, trying to regard it as pure mathematics results at best in piecewise clumsy and piecewise ill-motivated pure mathematics.

But it is a genuine bridge between mathematics and nature. These are both beautiful; both have "less is more" as a motto. Often the theory reflects the beauty of one or the other. Then it really comes to life. For sixty years, it has stood basically unaltered as the fundamental theory of macrophysics.

REFERENCES

1. M. Alonso and E. J. Finn, *Fundamental university physics*. I, II, Addison-Wesley, New York, 1970.

2. R. L. Bishop and R. J. Crittenden, *Geometry of manifolds*, Academic Press, New York and London, 1964. MR 29 #6401.

3. R. L. Bishop and S. I. Goldberg, *Tensor analysis on manifolds*, MacMillan, New York, 1968. MR 36 #7057.

4. C. Dewitt and B. S. Dewitt (editors), *Black holes*, Gordon and Breach, New York, 1973.

5. G. F. R. Ellis, *Relativistic cosmology*, General Relativity and Cosmology, Academic Press, New York, 1971. MR 49 #8567.

6. A. E. Fischer and J. E. Marsden, *Linearization stability of nonlinear partial differential equations*, Proc. Sympos. Pure Math., vol. 27, part 2, Amer. Math. Soc., Providence, R. I., 1975, pp. 219–263.

7. R. P. Geroch, *Domain of dependence*, J. Mathematical Phys. 11 (1970), 437–449. MR 49 #5585; erratum, 42, p. 1825.

8. R. P. Geroch and E. H. Kronheimer and R. Penrose, *Ideal points in space-time*, Proc. Roy. Soc. London Ser. A. 327 (1972), 545–567. MR 47 #4583.

9. S. Hawking and G. F. R. Ellis, *The large scale structure of spacetime*, Cambridge Univ. Press, London and New York, 1973.

10. S. Helgason, *Differential geometry and symmetric spaces*, Academic Press, New York and London, 1962. MR 26 #2986.

11. N. J. Hicks, *Notes on differential geometry*, Van Nostrand, Princeton, N. J., 1965. MR 31 #3936.

12. S. Kobayashi and K. Nomizu, *Foundations of differential geometry*. I, II, Interscience, New York and London, 1963, 1969. MR 27 #2945; 38 #6501.

13. E. Merzbacher, *Quantum mechanics*, 2nd ed., Wiley, New York, 1970. MR 41 #4912.

14. C. W. Misner, K. S. Thorne and J. A. Wheeler, *Gravitation*, Freeman, San Francisco, Calif., 1973.

15. P. J. E. Peebles, *Physical cosmology*, Princeton Univ. Press, Princeton, N. J., 1971.

16. R. Penrose, *Techniques of differential topology in relativity*, SIAM Publications, Philadelphia, 1972.

17. R. K. Sachs, *Cosmology*, Relativity, Astrophysics and Cosmology, Reidel, Holland, 1973, pp. 197–236.

18. R. K. Sachs and H. Wu, *General relativity for mathematicians*, Springer-Verlag, New York and Berlin, 1977.

19. K. S. Thorne, *The search for black holes*, Scientific American, vol. 231, no. 6, Dec. 1974.

20. S. Weinberg, *Gravitation and cosmology*, Wiley, New York, 1972.

21. _____, *Unified theories of elementary particle interaction*, Scientific American, vol. 231, no. 1, July 1974.

Department of Mathematics, University of California, Berkeley, California 94720