6
Causal structure

By postulate (*a*) of § 3.2, a signal can be sent between two points of $\mathscr{M}$ only if they can be joined by a non-spacelike curve. In this chapter we shall investigate further the properties of such causal relationships, establishing a number of results which will be used in chapter 8 to prove the existence of singularities.

By § 3.2, the study of causal relationships is equivalent to that of the conformal geometry of $\mathscr{M}$, i.e. of the set of all metrics $\tilde{\mathbf{g}}$ conformal to the physical metric $\mathbf{g}$ ($\tilde{\mathbf{g}} = \Omega^2\mathbf{g}$, where Ω is a non-zero, C^r function). Under such a conformal transformation of the metric a geodesic curve will not, in general, remain a geodesic curve unless it is null, and even in this case an affine parameter along the curve will not remain an affine parameter. Thus in most cases geodesic completeness (i.e. whether all geodesics can be extended to arbitrary values of their affine parameters) will depend on the particular conformal factor and so will not (except in certain special cases described in § 6.4) be a property of the conformal geometry. In fact Clarke (1971) and Siefert (1968) have shown that, provided a physically reasonable causality condition holds, any Lorentz metric is conformal to one in which all null geodesics and all future-directed timelike geodesics are complete. Geodesic completeness will be discussed further in chapter 8 where it forms the basis of a definition of a singularity.

§ 6.1 deals with the question of the orientability of timelike and spacelike bases. In § 6.2 basic causal relations are defined and the definition of a non-spacelike curve is extended from piecewise differentiable to continuous. The properties of the boundary of the future of a set are derived in § 6.3. In § 6.4 a number of conditions which rule out violations or near violations of causality are discussed. The closely related concepts of Cauchy developments and global hyperbolicity are introduced in § 6.5 and § 6.6, and are used in § 6.7 to prove the existence of non-spacelike geodesics of maximum length between certain pairs of points.

In § 6.8 we describe the construction of Geroch, Kronheimer and

https://doi.org/10.1017/CBO9780511524646.007 Published online by Cambridge University Press

Penrose for attaching a causal boundary to space–time. A particular example of such a boundary is provided by a class of asymptotically flat space–times which are studied in § 6.9.

6.1 Orientability

In our neighbourhood of space–time there is a well-defined arrow of time given by the direction of increase of entropy in quasi-isolated thermodynamic systems. It is not quite clear what the relationship is between this arrow and the other arrows defined by the expansion of the universe and by the direction of electrodynamic radiation; the reader who is interested will find further discussion in Gold (1967), Hogarth (1962), Hoyle and Narlikar (1963) and Ellis and Sciama (1972). Physically it would seem reasonable to suppose that there is a local thermodynamic arrow of time defined continuously at every point of space–time, but we shall only require that it should be possible to define continuously a division of non-spacelike vectors into two classes, which we arbitrarily label future- and past-directed. If this is the case, we shall say that space–time is *time-orientable*. In some space–times it is not possible to define such a time-orientation. An example is the space–time obtained from de Sitter space (§ 5.2) in which points are identified by reflection through the origin of the five-dimensional imbedding space. In this space there are closed curves, non-homotopic to zero, on going round which the orientation of time is reversed. However this difficulty could clearly be resolved by simply unidentifying the points again, and in fact this is always the case: if a space–time $(\mathscr{M}, \mathbf{g})$ is not time-orientable, then it has a double covering space $(\tilde{\mathscr{M}}, \mathbf{g})$ which is. $\tilde{\mathscr{M}}$ may be defined as the set of all pairs (p, α) where $p \in \mathscr{M}$ and α is one of the two orientations of time at p. Then with the natural structure and the projection π: $(p, \alpha) \rightarrow p$, $\tilde{\mathscr{M}}$ is a double covering of $\mathscr{M}$. If $\tilde{\mathscr{M}}$ consists of two disconnected components then $(\mathscr{M}, \mathbf{g})$ is time-orientable. If $\tilde{\mathscr{M}}$ is connected, then $(\mathscr{M}, \mathbf{g})$ is not time-orientable but $(\tilde{\mathscr{M}}, \mathbf{g})$ is. In the following sections we shall assume that either $(\mathscr{M}, \mathbf{g})$ is time-orientable or we are dealing with the time-orientable covering space. If one can prove the existence of singularities in this space–time then there must also be singularities in $(\mathscr{M}, \mathbf{g})$.

One may also ask whether space–time is *space-orientable*, that is whether it is possible to divide bases of three spacelike axes into right handed and left handed bases in a continuous manner. Geroch (1967*a*) has pointed out that there is an interesting connection between this and time-orientability which follows because some experiments on elementary particles are not invariant under charge or parity reversals, either singly or together. On the other hand there are theoretical reasons for believing that all interactions are invariant under the combination of charge, parity and time reversals (CPT theorem; see Streater and Wightman (1964)). If one believes that the non-invariance of weak interactions under charge and parity reversals is not merely a local effect but exists at all points of space–time, then it follows that going round any closed curve either the sign of a charge, the orientation of a basis of spacelike axes, and the orientation of time must all reverse, or none of them does. (The ordinary Maxwell theory, in which the electromagnetic field has a definite sign at every point, does not allow the sign of a charge to change on going around a closed curve non-homotopic to zero unless the orientation of time changes. However one could have a theory in which the field was double-valued and changed sign on going round such a curve. This theory would agree with all existing experimental evidence.) In particular if one assumes that space–time is time-orientable then it must also be space-orientable. (This in fact follows on using the experimental evidence alone without appealing to the CPT theorem.)

Geroch (1968*c*) has also shown that if it is possible to define two-component spinor fields at every point then space–time must be parallelizable, that is it must be possible to introduce a continuous system of bases of the tangent space at every point. (Further consequences of the existence of spinor structures are obtained in Geroch (1970*a*).)

6.2 Causal curves

Taking space–time to be time-orientable as explained in the previous section, one can divide the non-spacelike vectors at each point into future- and past-directed. For sets $\mathscr{S}$ and $\mathscr{U}$, the *chronological future* $I^+(\mathscr{S}, \mathscr{U})$ *of* $\mathscr{S}$ *relative to* $\mathscr{U}$ can then be defined as the set of all points in $\mathscr{U}$ which can be reached from $\mathscr{S}$ by a future-directed timelike curve in $\mathscr{U}$. (By a curve we mean always one of non-zero extent, not just a single point. Thus $I^+(\mathscr{S}, \mathscr{U})$ may not contain $\mathscr{S}$.) $I^+(\mathscr{S}, \mathscr{M})$ will be denoted by $I^+(\mathscr{S})$, and is an open set, since if $p \in \mathscr{M}$ can be reached by a future-directed timelike curve from $\mathscr{S}$ then there is a small neighbourhood of p which can be so reached.

This definition has a dual in which 'future' is replaced by 'past', and the $+$ by a $-$; to avoid repetition, we shall regard dual definitions and results as self-evident.

The *causal future of $\mathscr{S}$ relative to $\mathscr{U}$* is denoted by $J^+(\mathscr{S}, \mathscr{U})$; it is defined as the union of $\mathscr{S} \cap \mathscr{U}$ with the set of all points in $\mathscr{U}$ which can be reached from $\mathscr{S}$ by a future-directed non-spacelike curve in $\mathscr{U}$. We saw in §4.5 that a non-spacelike curve between two points which was not a null geodesic curve could be deformed into a timelike curve between the two points. Thus if $\mathscr{U}$ is an open set and $p, q, r \in \mathscr{U}$, then

$$\left.\begin{array}{ll}\text{either} & q \in J^+(p,\mathscr{U}),\, r \in I^+(q,\mathscr{U}) \\ \text{or} & q \in I^+(p,\mathscr{U}),\, r \in J^+(q,\mathscr{U})\end{array}\right\} \text{ imply } \quad r \in I^+(p,\mathscr{U}).$$

From this it follows that $\overline{I^+(p,\mathscr{U})} = \overline{J^+(p,\mathscr{U})}$ and $\dot{I}^+(p,\mathscr{U}) = \dot{J}^+(p,\mathscr{U})$ where for any set $\mathscr{K}$, $\overline{\mathscr{K}}$ denotes the closure of $\mathscr{K}$ and

$$\dot{\mathscr{K}} \equiv \overline{\mathscr{K}} \cap (\overline{\mathscr{M} - \mathscr{K}})$$

denotes the boundary of $\mathscr{K}$.

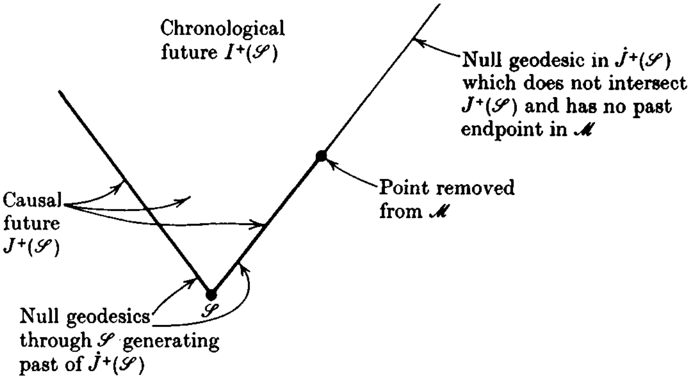
Chronological future $I^+(\mathscr{S})$
Null geodesic in $\dot{J}^+(\mathscr{S})$ which does not intersect $J^+(\mathscr{S})$ and has no past endpoint in $\mathscr{M}$
Point removed from $\mathscr{M}$
Causal future $J^+(\mathscr{S})$
$\mathscr{S}$
Null geodesics through $\mathscr{S}$ generating past of $\dot{J}^+(\mathscr{S})$

FIGURE 34. When a point has been removed from Minkowski space, the causal future $J^+(\mathscr{S})$ of a closed set $\mathscr{S}$ is not necessarily closed. Further parts of the boundary of the future of $\mathscr{S}$ may be generated by null geodesic segments which have no past endpoints in $\mathscr{M}$.

As before, $J^+(\mathscr{S}, \mathscr{M})$ will be written simply as $J^+(\mathscr{S})$. It is the region of space–time which can be causally affected by events in $\mathscr{S}$. It is not necessarily a closed set even when $\mathscr{S}$ is a single point, as figure 34 shows. This example, incidentally, illustrates a useful technique for constructing space–times with given causal properties: one starts with some simple space–time (unless otherwise indicated this will be Minkowski space), cuts out any closed set and, if desired, pastes it together in an appropriate way (i.e. one makes identifications of points of $\mathscr{M}$). The result is still a manifold with a Lorentz metric and therefore still a space–time even though it may look rather incomplete where points have been cut out. As mentioned above, however, this incompleteness can be cured by an appropriate conformal transformation which sends the cut out points to infinity.

The *future horismos of* $\mathscr{S}$ *relative to* $\mathscr{U}$, denoted by $E^+(\mathscr{S}, \mathscr{U})$, is defined as $J^+(\mathscr{S}, \mathscr{U}) - I^+(\mathscr{S}, \mathscr{U})$; we write $E^+(\mathscr{S})$ for $E^+(\mathscr{S}, \mathscr{M})$. (In some papers the relations $p \in I^+(q)$, $p \in J^+(q)$ and $p \in E^+(q)$ are denoted by $q \ll p$, $q < p$ and $q \to p$ respectively.) If $\mathscr{U}$ is an open set, points of $E^+(\mathscr{S}, \mathscr{U})$ must lie on future-directed null geodesics from $\mathscr{S}$ by proposition 4.5.10, and if $\mathscr{U}$ is a convex normal neighbourhood about p then it follows from proposition 4.5.1 that $E^+(p, \mathscr{U})$ consists of the future-directed null geodesics in $\mathscr{U}$ from p, and forms the boundary in $\mathscr{U}$ of both $I^+(p, \mathscr{U})$ and $J^+(p, \mathscr{U})$. Thus in Minkowski space, the null cone of p forms the boundary of the causal and chronological futures of p. However in more complicated space–times this is not necessarily the case (e.g. see figure 34).

For the purposes of what follows it will be convenient to extend the definition of timelike and non-spacelike curves from piecewise differentiable to continuous curves. Although such a curve may not have a tangent vector we can still say that it is non-spacelike if locally every two points of the curve can be joined by a piecewise differentiable non-spacelike curve. More precisely, we shall say that a continuous curve $\gamma: F \to \mathscr{M}$, where F is a connected interval of R^1, is *future-directed and non-spacelike* if for every $t \in F$ there is a neighbourhood G of t in F and a convex normal neighbourhood $\mathscr{U}$ of $\gamma(t)$ in $\mathscr{M}$ such that for any $t_1 \in G$, $\gamma(t_1) \in J^-(\gamma(t), \mathscr{U}) - \gamma(t)$ if $t_1 < t$, and $\gamma(t_1) \in J^+(\gamma(t), \mathscr{U}) - \gamma(t)$ if $t < t_1$. We shall say that γ is *future-directed and timelike* if the same conditions hold with J replaced by I. Unless otherwise specified, we will in future mean by a timelike or non-spacelike curve such a continuous curve, and shall regard two curves as equivalent if one is a reparametrization of the other. With this generalization we can establish a result that will be used repeatedly in the rest of this chapter. We first give a few more definitions.

A point p will be said to be a *future endpoint* of a future-directed non-spacelike curve $\gamma: F \to \mathscr{M}$ if for every neighbourhood $\mathscr{V}$ of p there is a $t \in F$ such that $\gamma(t_1) \in \mathscr{V}$ for every $t_1 \in F$ with $t_1 \geqslant t$. A non-spacelike curve is *future-inextendible* (respectively, *future-inextendible in a set* $\mathscr{S}$) if it has no future endpoint (respectively, no future endpoint in $\mathscr{S}$). A point p will be said to be a *limit point* of an infinite sequence of non-spacelike curves λ_n if every neighbourhood of p intersects an infinite number of the λ_n. A non-spacelike curve λ will be said to be a *limit curve* of the sequence λ_n if there is a subsequence λ'_n of the λ_n such that for every $p \in \lambda$, λ'_n converges to p.

Lemma 6.2.1

Let $\mathscr{S}$ be an open set and let λ_n be an infinite sequence of non-spacelike curves in $\mathscr{S}$ which are future-inextendible in $\mathscr{S}$. If $p \in \mathscr{S}$ is a limit point of λ_n, then through p there is a non-spacelike curve λ which is future-inextendible in $\mathscr{S}$ and which is a limit curve of the λ_n.

It is sufficient to consider the case $\mathscr{S} = \mathscr{M}$ since $\mathscr{S}$ can be regarded as a manifold with a Lorentz metric. Let $\mathscr{U}_1$ be a convex normal coordinate neighbourhood about p and let $\mathscr{B}(q, a)$ be the open ball of coordinate radius a about q. Let $b > 0$ be such that $\mathscr{B}(p, b)$ is defined and let $\lambda(1, 0)_n$ be a subsequence of $\lambda_n \cap \mathscr{U}_1$ which converges to p. Since $\dot{\mathscr{B}}(p, b)$ is compact it will contain limit points of the $\lambda(1, 0)_n$. Any such limit point y must lie either in $J^-(p, \mathscr{U}_1)$ or $J^+(p, \mathscr{U}_1)$ since otherwise there would be neighbourhoods $\mathscr{V}_1$ of y and $\mathscr{V}_2$ of p between which there would be no non-spacelike curve in $\mathscr{U}_1$. Choose

$$x_{11} \in J^+(p, \mathscr{U}_1) \cap \dot{\mathscr{B}}(p, b)$$

to be one of these limit points (figure 35), and choose $\lambda(1, 1)_n$ to be a subsequence of $\lambda(1, 0)_n$ which converges to x_{11}. The point x_{11} will be a point of our limit curve λ. Continue inductively, defining

$$x_{ij} \in J^+(p, \mathscr{U}_1) \cap \dot{\mathscr{B}}(p, i^{-1}jb)$$

as a limit point of the subsequence $\lambda(i-1, i-1)_n$ for $j = 0$, $\lambda(i, j-1)_n$ for $i \geqslant j \geqslant 1$, and defining $\lambda(i, j)_n$ as a subsequence of the above subsequence which converges to x_{ij}. In other words we are dividing the interval $[0, b]$ into smaller and smaller sections and getting points on our limit curve on the corresponding spheres about p. As any two of the x_{ij} will have non-spacelike separation, the closure of the union of all the x_{ij} $(j \geqslant i)$ will give a non-spacelike curve λ from $p = x_{i0}$ to $x_{11} = x_{ii}$. It now remains to construct a subsequence λ'_n of the λ_n such that for each $q \in \lambda$, λ'_n converges to q. We do this by choosing λ'_m to be a member of the subsequence $\lambda(m, m)_n$ which intersects each of the balls $\mathscr{B}(x_{mj}, m^{-1}b)$ for $0 \leqslant j \leqslant m$. Thus λ will be a limit curve of the λ_n from p to x_{11}. Now let $\mathscr{U}_2$ be a convex normal neighbourhood about x_{11} and repeat the construction using this time the sequence λ'_n. Continuing in this fashion, one can extend λ indefinitely. □

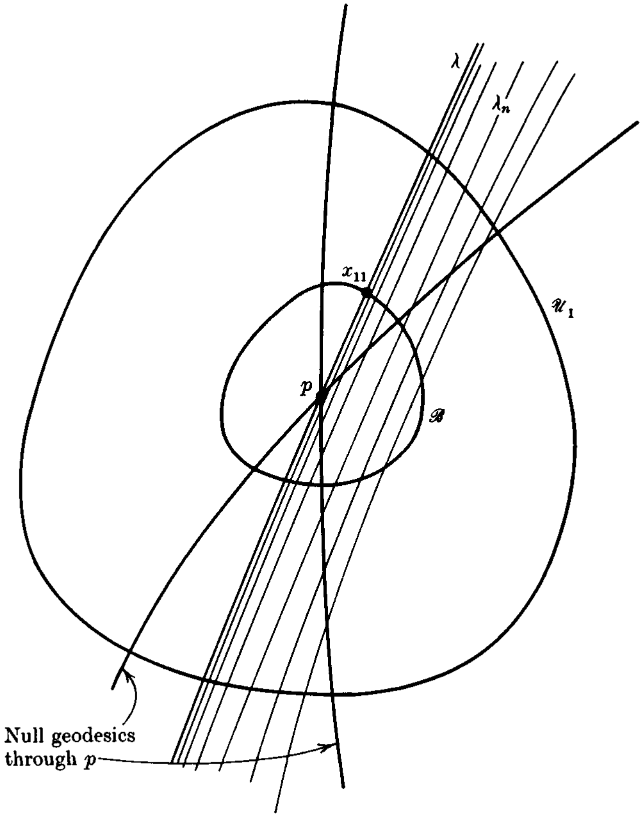

FIGURE 35. The non-spacelike limit curve λ through p of a family of non-spacelike curves λ_n for which p is a limit point.

6.3 Achronal boundaries

From proposition 4.5.1 it follows that in a convex normal neighbourhood $\mathscr{U}$, the boundary of $I^+(p, \mathscr{U})$ or $J^+(p, \mathscr{U})$ is formed by the future-directed null geodesics from p. To derive the properties of more general boundaries we introduce the concepts of achronal and future sets.

A set $\mathscr{S}$ is said to be *achronal* (sometimes referred to as 'semi-spacelike' in the literature) if $I^+(\mathscr{S}) \cap \mathscr{S}$ is empty, in other words if there are no two points of $\mathscr{S}$ with timelike separation. $\mathscr{S}$ is said to be a *future set* if $\mathscr{S} \supset I^+(\mathscr{S})$. Note that if $\mathscr{S}$ is a future set, $\mathscr{M} - \mathscr{S}$ is a past set. Examples of future sets include $I^+(\mathscr{N})$ and $J^+(\mathscr{N})$, where $\mathscr{N}$ is any set. Examples of achronal sets are given by the following fundamental result.

Proposition 6.3.1

If $\mathscr{S}$ is a future set then $\dot{\mathscr{S}}$, the boundary of $\mathscr{S}$, is a closed, imbedded, achronal three-dimensional C^{1-} submanifold.

If $q \in \dot{\mathscr{S}}$, any neighbourhood of q intersects $\mathscr{S}$ and $\mathscr{M}-\mathscr{S}$. If $p \in I^+(q)$, then there is a neighbourhood of q in $I^-(p)$. Thus $I^+(q) \subset \mathscr{S}$. Similarly $I^-(q) \subset (\mathscr{M}-\mathscr{S})$. If $r \in I^+(q)$, there is a neighbourhood $\mathscr{V}$ of r such that $\mathscr{V} \subset I^+(q) \subset \mathscr{S}$. Thus r cannot belong to $\dot{\mathscr{S}}$. One can introduce normal coordinates (x^1, x^2, x^3, x^4) in a neighbourhood $\mathscr{U}_\alpha$ about q with $\partial/\partial x^4$ timelike and such that the curves $\{x^i = \text{constant } (i = 1, 2, 3)\}$ intersect both $I^+(q, \mathscr{U}_\alpha)$ and $I^-(q, \mathscr{U}_\alpha)$. Then each of these curves must contain precisely one point of $\dot{\mathscr{S}}$. The x^4-coordinate of these points must be a Lipschitz function of the x^i $(i = 1, 2, 3)$ since no two points of $\dot{\mathscr{S}}$ have timelike separation. Therefore the one–one map $\phi_\alpha: \dot{\mathscr{S}} \cap \mathscr{U}_\alpha \rightarrow R^3$ defined by $\phi_\alpha(p) = x^i(p)$ $(i = 1, 2, 3)$ for $p \in \dot{\mathscr{S}} \cap \mathscr{U}_\alpha$ is a homeomorphism. Thus $(\dot{\mathscr{S}} \cap \mathscr{U}_\alpha, \phi_\alpha)$ is a C^{1-} atlas for $\dot{\mathscr{S}}$. □

We shall call a set with the properties of $\dot{\mathscr{S}}$ listed in proposition 6.3.1, an *achronal boundary*. Such a set can be divided into four disjoint subsets $\dot{\mathscr{S}}_N$, $\dot{\mathscr{S}}_+$, $\dot{\mathscr{S}}_-$, $\dot{\mathscr{S}}_0$ as follows: for a point $q \in \dot{\mathscr{S}}$ there may or may not exist points $p, r \in \dot{\mathscr{S}}$ with $p \in E^-(q) - q$, $r \in E^+(q) - q$. The different possibilities define the subsets of $\dot{\mathscr{S}}$ according to the scheme:

	$\exists p$	$\nexists p$	
$q \in$	$\dot{\mathscr{S}}_N$	$\dot{\mathscr{S}}_-$	$\exists r$
	$\dot{\mathscr{S}}_+$	$\dot{\mathscr{S}}_0$	$\nexists r$

If $q \in \dot{\mathscr{S}}_N$, then $r \in E^+(p)$ since $r \in J^+(p)$ and by proposition 6.3.1, $r \notin I^+(p)$. This means that there is a null geodesic segment in $\dot{\mathscr{S}}$ through q. If $q \in \dot{\mathscr{S}}_+$ (respectively $\dot{\mathscr{S}}_-$) then q is the future (respectively, past) endpoint of a null geodesic in $\dot{\mathscr{S}}$. The subset $\dot{\mathscr{S}}_0$ is spacelike (more strictly, acausal). These divisions are illustrated in figure 36.

A useful condition for a point to lie in $\dot{\mathscr{S}}_N$, $\dot{\mathscr{S}}_+$ or $\dot{\mathscr{S}}_-$ is given in the following lemma due to Penrose (Penrose (1968)):

Lemma 6.3.2

Let $\mathscr{W}$ be a neighbourhood of $q \in \dot{\mathscr{S}}$ where $\mathscr{S}$ is a future set. Then

(i) $I^+(q) \subset I^+(\mathscr{S} - \mathscr{W})$ implies $q \in \dot{\mathscr{S}}_N \cup \dot{\mathscr{S}}_+$,

(ii) $I^-(q) \subset I^-(\mathscr{M} - \mathscr{S} - \mathscr{W})$ implies $q \in \dot{\mathscr{S}}_N \cup \dot{\mathscr{S}}_-$.

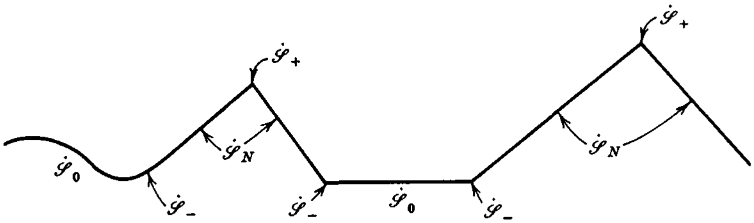

FIGURE 36. An achronal boundary $\dot{\mathscr{S}}$ can be divided into four sets: $\dot{\mathscr{S}}_0$ is spacelike, $\dot{\mathscr{S}}_N$ is null, and $\dot{\mathscr{S}}_+$ (respectively, $\dot{\mathscr{S}}_-$) is the future (respectively, past) endpoint of a null geodesic in $\dot{\mathscr{S}}$.

It is sufficient to prove (i) since $\dot{\mathscr{S}}$ can also be regarded as the boundary of the past set $(\mathscr{M}-\mathscr{S})$. Let $\{x_n\}$ be an infinite sequence of points in $I^+(q) \cap \mathscr{W}$ which converge on q. If $I^+(q) \subset I^+(\mathscr{S}-\mathscr{W})$, there will be a past-directed timelike curve λ_n to $\mathscr{S}-\mathscr{W}$ from each x_n. By lemma 6.2.1 there will be a past-directed limit curve λ from q to $(\overline{\mathscr{S}-\mathscr{W}})$. As $I^-(q)$ is open and contained in $\mathscr{M}-\mathscr{S}$, $I^-(q) \cap \mathscr{S}$ is empty. Thus λ must be a null geodesic and must lie in $\dot{\mathscr{S}}$. □

As an example of the above results, consider $\dot{J}^+(\mathscr{K}) = \dot{I}^+(\mathscr{K})$, the boundary of the future of a closed set $\mathscr{K}$. By proposition 6.3.1 it is an achronal manifold and by the above lemma, every point of $\dot{J}(\mathscr{K})-\mathscr{K}$ belongs to $[\dot{J}^+(\mathscr{K})]_N$ or $[\dot{J}^+(\mathscr{K})]_+$. This means that $\dot{J}(\mathscr{K})-\mathscr{K}$ is generated by null geodesic segments which may have future endpoints in $\dot{J}^+(\mathscr{K})-\mathscr{K}$ but which, if they do have past endpoints, can have them only on $\mathscr{K}$ itself. As figure 34 shows, there may be null geodesic generating segments which do not have past endpoints at all but which go out to infinity. This example is admittedly rather artificial but Penrose (1965*a*) has shown that similar behaviour occurs in something as simple as the plane wave solutions; the anti-de Sitter (§ 5.2) and Reissner–Nordström (§ 5.5) solutions provide other examples. We shall see in § 6.6 that this behaviour is connected with the absence of a Cauchy surface for these solutions.

We shall say that an open set $\mathscr{U}$ is *causally simple* if for every compact set $\mathscr{K} \subset \mathscr{U}$,

$$\dot{J}^+(\mathscr{K}) \cap \mathscr{U} = E^+(\mathscr{K}) \cap \mathscr{U} \quad \text{and} \quad \dot{J}^-(\mathscr{K}) \cap \mathscr{U} = E^-(\mathscr{K}) \cap \mathscr{U}.$$

This is equivalent to saying that $J^+(\mathscr{K})$ and $J^-(\mathscr{K})$ are closed in $\mathscr{U}$.

6.4 Causality conditions

Postulate (*a*) of § 3.2 required only that causality should hold locally; the global question was left open. Thus we did not rule out the possibility that on a large scale there might be closed timelike curves (i.e. timelike S^1's). However the existence of such curves would seem to lead to the possibility of logical paradoxes: for, one could imagine that with a suitable rocketship one could travel round such a curve and, arriving back before one's departure, one could prevent oneself from setting out in the first place. Of course there is a contradiction only if one assumes a simple notion of free will; but this is not something which can be dropped lightly since the whole of our philosophy of science is based on the assumption that one is free to perform any experiment. It might be possible to form a theory in which there were closed timelike curves and in which the concept of free will was modified (see, for example, Schmidt (1966)) but one would be much more ready to believe that space–time satisfies what we shall call the *chronology condition*: namely, that there are no closed timelike curves. One must however bear in mind the possibility that there might be points (maybe where the density or curvature was very high) of space–time at which this condition does not hold. The set of all such points will be called the *chronology violating* set of $\mathscr{M}$ and has the following character:

Proposition 6.4.1 (*Carter*)

The chronology violating set of $\mathscr{M}$ is the disjoint union of sets of the form $I^+(q) \cap I^-(q)$, $q \in \mathscr{M}$.

If q is in the chronology violating set of $\mathscr{M}$, there must be a future-directed timelike curve λ with past and future endpoints at q. If $r \in I^-(q) \cap I^+(q)$, there will be past- and future-directed timelike curves μ_1 and μ_2 from q to r. Then $(\mu_1)^{-1} \circ \lambda \circ \mu_2$ will be a future-directed timelike curve with past and future endpoints at r. Moreover if

$$r \in [I^-(q) \cap I^+(q)] \cap [I^-(p) \cap I^+(p)]$$

then
$$p \in I^-(q) \cap I^+(q) = I^-(p) \cap I^+(p).$$

To complete the proof, note that every point r at which chronology is violated is in the set $I^-(r) \cap I^+(r)$. □

Proposition 6.4.2

If $\mathscr{M}$ is compact, the chronology violating set of $\mathscr{M}$ is non-empty.

$\mathscr{M}$ can be covered by open sets of the form $I^+(q)$, $q \in \mathscr{M}$. If the chronology condition holds at q, then $q \notin I^+(q)$. Thus if the chronology condition held at every point, $\mathscr{M}$ could not be covered by a finite number of sets of the form $I^+(q)$. □

From this result it would seem reasonable to assume that space–time is non-compact. Another argument against compactness is that any compact, four-dimensional manifold on which there is a Lorentz metric cannot be simply connected. (The existence of a Lorentz metric implies that the Euler number $\chi(\mathscr{M})$ is zero (Steenrod (1951), p. 207). Now $\chi = \sum_{n=0}^{4}(-1)^n B_n$ where $B_n \geqslant 0$ is the nth Betti number of $\mathscr{M}$. By duality (Spanier (1966), p. 297) $B_n = B_{4-n}$. Since $B_0 = B_4 = 1$, this implies that $B_1 \neq 0$ which in turn implies $\pi_1(\mathscr{M}) \neq 0$ (Spanier (1966), p. 398).) Thus a compact space–time is really a non-compact manifold in which points have been identified. It would seem physically reasonable not to identify points but to regard the covering manifold as representing space–time.

We shall say that the *causality condition* holds if there are no closed non-spacelike curves. Similar to proposition 6.4.1, one has:

Proposition 6.4.3

The set of points at which the causality condition does not hold is the disjoint union of sets of the form $J^-(q) \cap J^+(q)$, $q \in \mathscr{M}$. □

In particular, if the causality condition is violated at $q \in \mathscr{M}$ but the chronology condition holds, there must be a closed null geodesic curve γ through q. Let v be an affine parameter on γ (regarded as a map of an open interval of R^1 to $\mathscr{M}$) and let $\ldots, v_{-1}, v_0, v_1, v_2, \ldots$ be successive values of v at q. Then we may compare at q the tangent vector $\partial/\partial v|_{v=v_0}$ and the tangent vector $\partial/\partial v|_{v=v_1}$, obtained by parallelly transporting $\partial/\partial v|_{v=v_0}$ round γ. Since they both point in the same direction, they must be proportional: $\partial/\partial v|_{v=v_1} = a\,\partial/\partial v|_{v=v_0}$. The factor a has the following significance: the affine distance covered in the nth circuit of γ, $(v_{n+1}-v_n)$, is equal to $a^{-n}(v_1-v_0)$. Thus if $a > 1$, v never attains the value $(v_1-v_0)(1-a^{-1})^{-1}$ and so γ is geodesically incomplete in the future direction even though one can go round an infinite number of times. Similarly if $a < 1$, γ is incomplete in the past direction, while if $a = 1$, it is complete in both directions. In the two-dimensional model of Taub–NUT space described in § 5.7, there is a closed null geodesic which is an example with $a > 1$. Since the factor a is a conformal invariant, this incompleteness is independent of the conformal factor. This kind of behaviour, however, can happen only if there is a violation of causality in some sense; if the strong causality condition (see below) holds, a suitable conformal transformation of the metric will make all null geodesics complete (Clarke (1971)).

The factor a has a further significance from the following result.

Proposition 6.4.4

If γ is a closed null geodesic curve which is incomplete in the future direction then there is a variation of γ which moves each point of γ towards the future and which yields a closed timelike curve.

By § 2.6, one can find on $\mathscr{M}$ a timelike line-element field $(\mathbf{V}, -\mathbf{V})$ normalized so that $g(\mathbf{V}, \mathbf{V}) = -1$. As we are assuming that $\mathscr{M}$ is time-orientable, one can consistently choose one direction of $(\mathbf{V}, -\mathbf{V})$ and so obtain a future-directed timelike unit vector field $\mathbf{V}$. One can then define a positive definite metric $\mathbf{g}'$ by

$$g'(\mathbf{X}, \mathbf{Y}) = g(\mathbf{X}, \mathbf{Y}) + 2g(\mathbf{X}, \mathbf{V})\, g(\mathbf{Y}, \mathbf{V}).$$

Let t be a (non-affine) parameter on γ which is zero at some point $q \in \gamma$ and which is such that $g(\mathbf{V}, \partial/\partial t) = -2^{-\frac{1}{2}}$. Then t measures proper distance along γ in the metric $\mathbf{g}'$ and has the range $-\infty < t < \infty$. Consider a variation of γ with variation vector $\partial/\partial u$ equal to $x\mathbf{V}$, where x is a function $x(t)$. By § 4.5,

$$\begin{aligned} \frac{1}{2}\frac{\partial}{\partial u} g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right) &= \frac{\mathrm{d}}{\mathrm{d}t} g\left(\frac{\partial}{\partial u}, \frac{\partial}{\partial t}\right) - g\left(\frac{\partial}{\partial u}, \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t}\right) \\ &= -2^{-\frac{1}{2}}\left(\frac{\mathrm{d}x}{\mathrm{d}t} - xf\right), \end{aligned}$$

where $f\,\partial/\partial t = (\mathrm{D}/\partial t)(\partial/\partial t)$. Now suppose v were an affine parameter on γ. Then $\partial/\partial v$ would be proportional to $\partial/\partial t$: $\partial/\partial v = h\,\partial/\partial t$, where $h^{-1}\,\mathrm{d}h/\mathrm{d}t = -f$. On going round one circuit of γ, $\partial/\partial v$ increases by a factor $a > 1$. Thus

$$\oint f\,\mathrm{d}t = -\log a \leqslant 0.$$

Therefore if we take $x(t)$ to be

$$\exp\left(\int_0^t f(t')\,\mathrm{d}t' + b^{-1}t\log a\right),$$

where $b = \oint \mathrm{d}t$, this will give a variation of γ to the future and gives a closed timelike curve. □

Proposition 6.4.5

If (*a*) $R_{ab}K^aK^b \geqslant 0$ for every null vector $\mathbf{K}$;

(*b*) the generic condition holds, i.e. every null geodesic contains a point at which $K_{[a}R_{b]cd[e}K_{f]}K^cK^d$ is non-zero, where $\mathbf{K}$ is the tangent vector;

(*c*) the chronology condition holds on $\mathscr{M}$,

then the causality condition holds on $\mathscr{M}$.

If there were closed null geodesic curves which were incomplete, then by the previous result they could be varied to give closed timelike curves. If they were complete, then by proposition 4.4.5 they would contain conjugate points and so by proposition 4.5.12 they could again be varied to give closed timelike curves. □

This shows that in physically realistic solutions, the causality and chronology conditions are equivalent.

As well as ruling out closed non-spacelike curves, it would seem reasonable to exclude situations in which there were non-spacelike curves which returned arbitrarily close to their point of origin or which passed arbitrarily close to other non-spacelike curves which then passed arbitrarily close to the origin of the first curve – and so on. In fact Carter (1971*a*) has pointed out that there is a more than countably infinite hierarchy of such higher degree causality conditions depending on the number and order of the limiting processes involved. We shall describe the first three of these conditions and shall then give the ultimate in causality conditions.

The *future* (respectively, *past*) *distinguishing condition* (Kronheimer and Penrose (1967)) is said to hold at $p \in \mathscr{M}$ if every neighbourhood of p contains a neighbourhood of p which no future (respectively, past) directed non-spacelike curve from p intersects more than once. An equivalent statement is that $I^+(q) = I^+(p)$ (respectively, $I^-(q) = I^-(p)$) implies that $q = p$. Figure 37 shows an example in which the causality and past distinguishing conditions hold everywhere but the future distinguishing condition does not hold at p.

The *strong causality condition* is said to hold at p if every neighbourhood of p contains a neighbourhood of p which no non-spacelike curve intersects more than once. Figure 38 shows an example of violation of this condition.

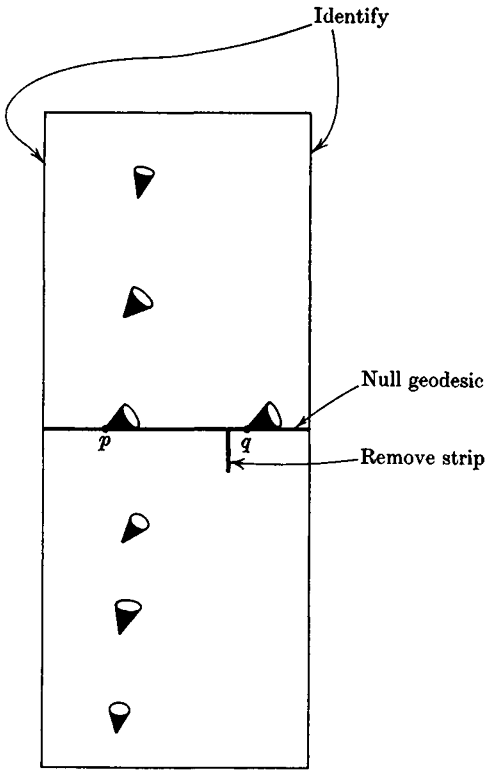

FIGURE 37. A space in which the causality and past distinguishing conditions hold everywhere, but the future distinguishing condition does not hold at p or q (in fact, $I^+(p) = I^+(q)$). The light cones on the cylinder tip over until one null direction is horizontal, and then tip back up; a strip has been removed, thus breaking the closed null geodesic that would otherwise occur.

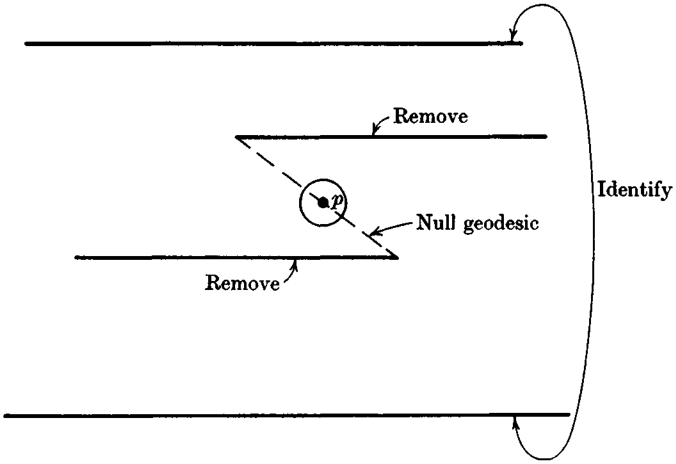

FIGURE 38. A space–time satisfying the causality, future and past distinguishing conditions, but not satisfying the strong causality condition at p. Two strips have been removed from a cylinder; light cones are at ± 45°.

Proposition 6.4.6

If conditions (*a*) to (*c*) of proposition 6.4.5 hold and if in addition, (*d*) $\mathscr{M}$ is null geodesically complete, then the strong causality condition holds on $\mathscr{M}$.

Suppose the strong causality condition did not hold at $p \in \mathscr{M}$. Let $\mathscr{U}$ be a convex normal neighbourhood of p and let $V_n \subset \mathscr{U}$ be an infinite sequence of neighbourhoods of p such that any neighbourhood of p contains all the V_n for n large enough. For each V_n there would be a future-directed non-spacelike curve λ_n which left $\mathscr{U}$ and then returned to V_n. By lemma 6.2.1, there would be an inextendible non-spacelike curve λ through p which was a limit curve of the λ_n. No two points of λ could have timelike separation as otherwise one could join up some λ_n to give a closed non-spacelike curve. Thus λ must be a null geodesic. But by (*a*), (*b*) and (*d*) λ would contain conjugate points and therefore points with timelike separation. □

Corollary

The past and future distinguishing conditions would also hold on $\mathscr{M}$ since they are implied by strong causality.

Closely related to these three higher degree causality conditions is the phenomenon of *imprisonment.*

A non-spacelike curve γ that is future-inextendible can do one of three things as one follows it to the future: it can

(i) enter and remain within a compact set $\mathscr{S}$,

(ii) not remain within any compact set but continually re-enter a compact set $\mathscr{S}$,

(iii) not remain within any compact set $\mathscr{S}$ and not re-enter any such set more than a finite number of times.

In the third case γ can be thought of as going off to the edge of space–time, that is either to infinity or a singularity. In the first and second cases we shall say that γ is *totally* and *partially future imprisoned* in $\mathscr{S}$, respectively. One might think that imprisonment could occur only if the causality condition was violated, but the example due to Carter which is illustrated in figure 39 shows that this is not the case. Nevertheless one does have the following result:

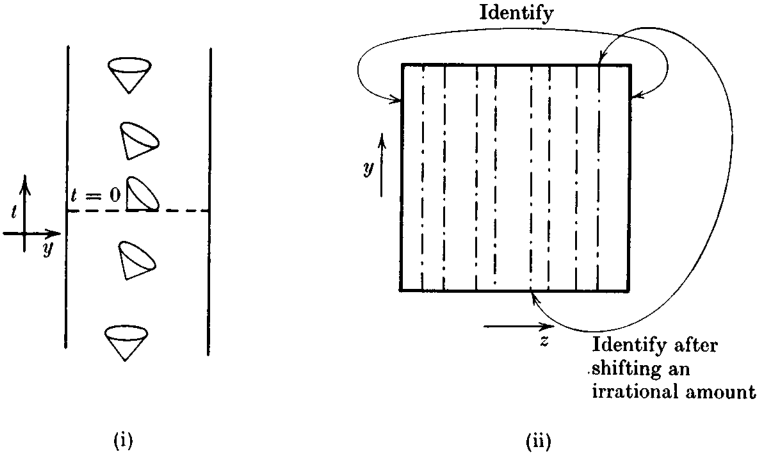

FIGURE 39. A space with imprisoned non-spacelike lines but no closed non-spacelike curves. The manifold is $R^1 \times S^1 \times S^1$ described by coordinates (t, y, z) where (t, y, z) and $(t, y, z+1)$ are identified, and (t, y, z) and $(y, y+1, z+a)$ are identified, where a is an irrational number. The Lorentz metric is given by

$$ds^2 = (\cosh t - 1)^2 (dt^2 - dy^2) + dt\,dy - dz^2.$$

(i) A section $\{z = \text{constant}\}$ showing the orientation of the null cones.
(ii) The section $t = 0$ showing part of a null geodesic.

Proposition 6.4.7

If the strong causality condition holds on a compact set $\mathscr{S}$, there can be no future-inextendible non-spacelike curve totally or partially future imprisoned in $\mathscr{S}$.

$\mathscr{S}$ can be covered by a finite number of convex normal coordinate neighbourhoods $\mathscr{U}_i$ with compact closure, such that no non-spacelike curve intersects any $\mathscr{U}_i$ more than once. (We shall call such neighbourhoods, *local causality neighbourhoods*.) Any future-inextendible non-spacelike curve which intersects one of these neighbourhoods must leave it again and not re-enter it. □

Proposition 6.4.8

If the future or past distinguishing condition holds on a compact set $\mathscr{S}$, there can be no future-inextendible non-spacelike curve totally future imprisoned in $\mathscr{S}$. (This result is included for its interest but is not needed for what follows.)

Let $\{\mathscr{V}_\alpha\}$, $(\alpha = 1, 2, 3, \ldots)$, be a countable basis of open sets for $\mathscr{M}$ (i.e. any open set in $\mathscr{M}$ can be represented as a union of the $\mathscr{V}_\alpha$). As the future or past distinguishing condition holds on $\mathscr{S}$, any point $p \in \mathscr{S}$ will have a convex normal coordinate neighbourhood $\mathscr{U}$ such that no future (respectively, past) directed non-spacelike curve from p intersects $\mathscr{U}$ more than once. We define $f(p)$ to be equal to the least value of α such that $\mathscr{V}_\alpha$ contains p and is contained in some such neighbourhood $\mathscr{U}$.

Suppose there were a future-inextendible non-spacelike curve λ which was totally future imprisoned in $\mathscr{S}$. Let $q \in \lambda$ be such that $\lambda' = \lambda \cap J^+(q)$ is contained in $\mathscr{S}$. Define $\mathscr{A}_0$ to be the closed, non-empty set consisting of all points of $\mathscr{S}$ which are limit points of λ. Let $p_0 \in \mathscr{A}_0$ be such that $f(p_0)$ is equal to the smallest value of $f(p)$ on $\mathscr{A}_0$. Through p_0 there would be an inextendible non-spacelike curve γ_0 every point of which was a limit point of λ'. No two points of γ_0 could have timelike separation since otherwise some segment of λ' could be deformed to give a closed non-spacelike curve. Thus γ_0 would be an inextendible null geodesic which was totally imprisoned in $\mathscr{S}$ in both the past and future directions. Let $\mathscr{A}_1$ be the closed set consisting of all limit points of $\gamma_0 \cap J^+(p_0)$ (or, in the case that the past distinguishing condition holds on $\mathscr{S}$, $\gamma_0 \cap J^-(p_0)$). As every such point would also be a limit point of λ', $\mathscr{A}_1 \subset \mathscr{A}_0$. Since $\mathscr{V}_{f(p_0)}$ could contain no limit point of $\gamma_0 \cap J^+(p_0)$ (respectively, $\gamma_0 \cap J^-(p_0)$), $\mathscr{A}_1$ would be strictly smaller than $\mathscr{A}_0$. We would thus obtain an infinite sequence of closed sets $\mathscr{A}_0 \supset \mathscr{A}_1 \supset \mathscr{A}_2 \supset \ldots \supset \mathscr{A}_\beta \supset \ldots$. Each $\mathscr{A}_\beta$ would be non-empty, being the set of all limit points of the totally future (respectively, past) imprisoned null geodesic $\gamma_{\beta-1} \cap J^+(p_{\beta-1})$ (respectively, $\gamma_{\beta-1} \cap J^-(p_{\beta-1})$). Let $\mathscr{K} = \bigcap\limits_{\beta} \mathscr{A}_\beta$. As $\mathscr{S}$ is compact, $\mathscr{K}$ would be non-empty since the intersection of any finite number of the $\mathscr{A}_\beta$ would be non-empty (Hocking and Young (1961), *p*. 19). Suppose $r \in \mathscr{K}$. Then $f(r) = f(p_\beta)$ for some β. But $\mathscr{V}_{f(p_\beta)} \cap \mathscr{A}_{\beta+1}$ would be empty so r could not be in $\mathscr{A}_{\beta+1}$ and so could not be in $\mathscr{K}$. This shows that there can be no future-inextendible non-spacelike curve totally future imprisoned in $\mathscr{S}$. □

The causal relations on $(\mathscr{M}, \mathbf{g})$ may be used to put a topology on $\mathscr{M}$ called the *Alexandrov topology* This is the topology in which a set is defined to be open if and only if it is the union of one or more sets of the form $I^+(p) \cap I^-(q)$, $p, q \in \mathscr{M}$. As $I^+(p) \cap I^-(q)$ is open in the manifold topology, any set which is open in the Alexandrov topology will be open in the manifold topology, though the converse is not necessarily true.

Suppose however that the strong causality condition holds on $\mathscr{M}$.

Then about any point $r \in \mathscr{M}$ one can find a local causality neighbourhood $\mathscr{U}$. The Alexandrov topology of $(\mathscr{U}, \mathbf{g}|_{\mathscr{U}})$ regarded as a space-time in its own right, is clearly the same as the manifold topology of $\mathscr{U}$. Thus the Alexandrov topology of $\mathscr{M}$ is the same as the manifold topology since $\mathscr{M}$ can be covered by local causality neighbourhoods. This means that if the strong causality condition holds, one can determine the topological structure of space–time by observation of causal relationships.

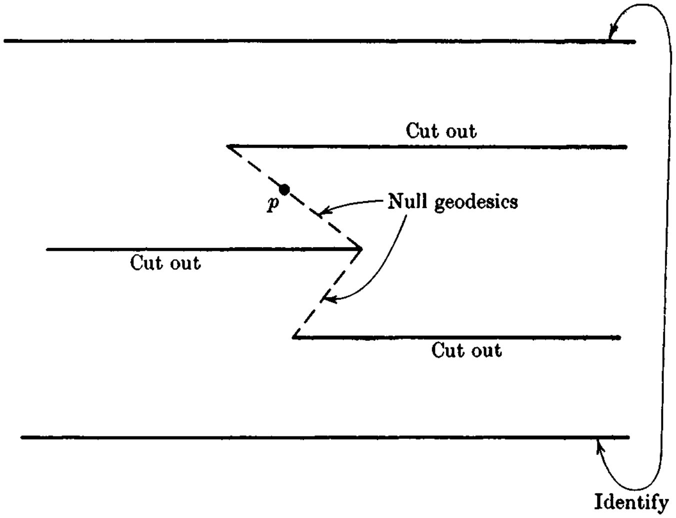

FIGURE 40. A space satisfying the strong causality condition, but in which the slightest variation of the metric would permit there to be closed timelike lines through p. Three strips have been removed from a cylinder; light cones are at ± 45°.

Even imposition of the strong casuality condition does not rule out all causal pathologies, as figure 40 shows one can still have a space–time which is on the verge of violating the chronology condition in that the slightest variation of the metric can lead to closed timelike curves. Such a situation would not seem to be physically realistic since General Relativity is presumably the classical limit of some, as yet unknown, quantum theory of space–time and in such a theory the Uncertainty Principle would prevent the metric from having an exact value at every point. Thus in order to be physically significant, a property of space–time ought to have some form of stability, that is to say, it should also be a property of 'nearby' space–times. In order to give a precise meaning to 'nearby' one has to define a topology on the set of all space-times, that is, all non-compact four-dimensional manifolds and all Lorentz metrics on them. We shall leave the problem of uniting in one connected topological space manifolds of different topologies (this can be done); and shall just consider putting a topology on the set of all C^r Lorentz metrics ($r \geqslant 1$) on a given manifold. There are various ways in which this can be done, depending on whether one requires a 'nearby' metric to be nearby in just its values (C^0 topology) or also in its derivatives up to the kth order (C^k topology) and whether one requires it to be nearby everywhere (open topology) or only on compact sets (compact open topology).

For our purposes here, we shall be interested in the C^0 *open topology*. This may be defined as follows: the symmetric tensor spaces $T_{S2}^{0}(p)$ of type $(0, 2)$ at every point $p \in \mathscr{M}$ form a manifold (with the natural structure) $T_{S2}^{0}(\mathscr{M})$, the bundle of symmetric tensors of type $(0, 2)$ over $\mathscr{M}$. A Lorentz metric $\mathbf{g}$ on $\mathscr{M}$ is an assignment of an element of $T_{S2}^{0}(\mathscr{M})$ at each point $p \in \mathscr{M}$ and so can be regarded as a map or cross-section $\hat{g}: \mathscr{M} \rightarrow T_{S2}^{0}(\mathscr{M})$ such that $\pi \circ \hat{g} = 1$ where π is the projection $T_{S2}^{0}(\mathscr{M}) \rightarrow \mathscr{M}$ which sends $x \in T_{S2}^{0}(p)$ to p. Let $\mathscr{U}$ be an open set in $T_{S2}^{0}(\mathscr{M})$ and let $O(\mathscr{U})$ be the set of all C^0 Lorentz metrics $\mathbf{g}$ such that $\hat{g}(\mathscr{M})$ is contained in $\mathscr{U}$ (figure 41). Then the open sets in the C^0 open topology of the C^r Lorentz metrics on $\mathscr{M}$ are defined to be the union of one or more sets of the form $O(\mathscr{U})$.

We say that the *stable causality condition* holds on $\mathscr{M}$ if the space-time metric $\mathbf{g}$ has an open neighbourhood in the C^0 open topology such that there are no closed timelike curves in any metric belonging to the neighbourhood. (It would not make any difference if one used the C^k topology here, but one could not use a compact open topology since in that topology each neighbourhood of any metric contains closed timelike curves.) In other words, what this condition means is that one can expand the light cones slightly at every point without introducing closed timelike curves.

Proposition 6.4.9

The stable causality condition holds everywhere on $\mathscr{M}$ if and only if there is a function f on $\mathscr{M}$ whose gradient is everywhere timelike.

Remark. The function f can be thought of as a sort of cosmic time in the sense that it increases along every future-directed non-spacelike curve.

Proof. The existence of a function f with an everywhere timelike gradient implies the stable causality condition since there can be no closed timelike curves in any metric $\mathbf{h}$ which is sufficiently close to $\mathbf{g}$ that for every point $p \in \mathcal{M}$, the null cone of p in the metric $\mathbf{h}$ intersects the surface $\{f = \text{constant}\}$ through p only at p. To show that the converse is true we introduce a volume measure μ (unrelated to the volume measure defined by the metric $\mathbf{g}$) on $\mathcal{M}$ such that the total volume of $\mathcal{M}$ is one. One way of doing this is as follows: choose a countable atlas $(\mathcal{U}_\alpha, \phi_\alpha)$ for $\mathcal{M}$ such that $\overline{\phi_\alpha(\mathcal{U}_\alpha)}$ is compact in R^4. Let μ_0 be the natural Euclidean measure on R^4 and let f_α be a partition of unity for the atlas $(\mathcal{U}_\alpha, \phi_\alpha)$. Then μ may be defined as $\sum_\alpha f_\alpha 2^{-\alpha}[\mu_0(\mathcal{U}_\alpha)]^{-1} \phi_\alpha{}^* \mu_0$.

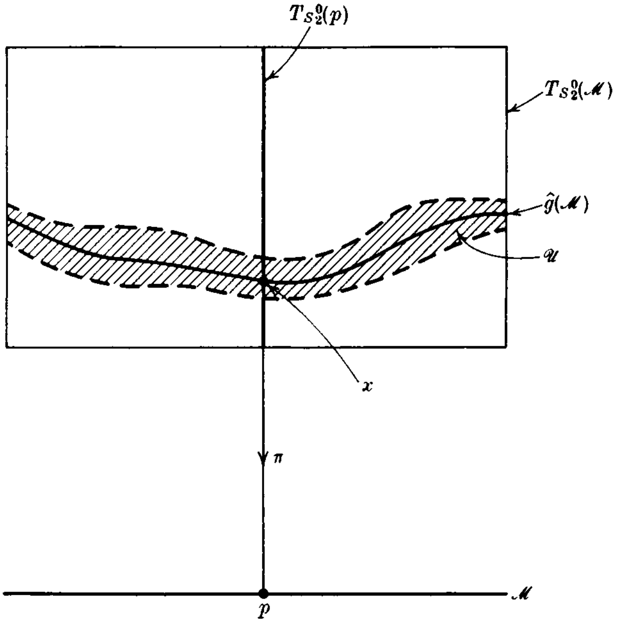

FIGURE 41. An open set $\mathcal{U}$ in the C^0 open topology on the space $T_{S_2^0}(\mathcal{M})$ of symmetric tensors of type $(0, 2)$ on $\mathcal{M}$.

Now if the stable causality condition holds one can find a family of C^r Lorentz metrics $\mathbf{h}(a)$, $a \in [0, 3]$, such that:

(1) $\mathbf{h}(0)$ is the space–time metric $\mathbf{g}$;

(2) there are no closed timelike curves in the metric $\mathbf{h}(a)$ for each $a \in [0, 3]$;

(3) if $a_1, a_2 \in [0, 3]$ with $a_1 < a_2$, then every non-spacelike vector in the metric $\mathbf{h}(a_1)$ is timelike in the metric $\mathbf{h}(a_2)$.

For $p \in \mathscr{M}$, let $\theta(p,a)$ be the volume of $I^-(p, \mathscr{M}, \mathbf{h}(a))$ in the measure μ where we use $I^-(\mathscr{S}, \mathscr{U}, \mathbf{h})$ to denote the past of $\mathscr{S}$ relative to $\mathscr{U}$ in the metric $\mathbf{h}$. For a given value of $a \in (0,3)$, $\theta(p,a)$ will be a bounded function which increases along every non-spacelike curve. It may not, however, be continuous: as figure 42 shows, it may be possible that a slight alteration of position may allow one to see past an obstruction and so greatly increase the volume of the past. One thus needs some way of smearing out $\theta(p,a)$ so as to obtain a continuous function which increases along every curve which is future-directed and non-spacelike in the metric $\mathbf{h}(0)$. One can do this by averaging over a range of a: let

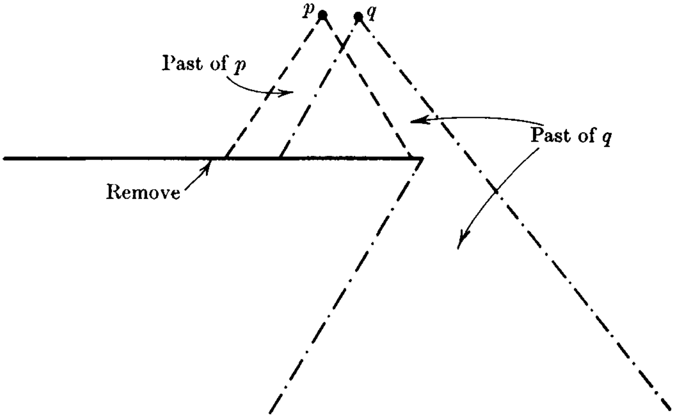

FIGURE 42. A small displacement of a point from p to q results in a large change in the volume of the past of the point. Light cones are at ± 45° and a strip has been removed as shown.

$$\bar{\theta}(p) = \int_1^2 \theta(p,a)\,\mathrm{d}a.$$

We shall show that $\bar{\theta}(p)$ is continuous on $\mathscr{M}$.

First to show that it is upper semi-continuous: given $\epsilon > 0$, let $\mathscr{B}$ be a ball about p such that the volume of $\mathscr{B}$ in the measure μ is less than $\frac{1}{2}\epsilon$. By property (3), for a_1, $a_2 \in [0,3]$ with $a_1 < a_2$ one can find a neighbourhood $\mathscr{F}(a_1, a_2)$ of p in $\mathscr{B}$ such that

$$[I^-(\mathscr{F}(a_1,a_2), \bar{\mathscr{B}}, \mathbf{h}(a_1)) \cap \dot{\mathscr{B}}] \subset [I^-(p, \bar{\mathscr{B}}, \mathbf{h}(a_2)) \cap \dot{\mathscr{B}}].$$

Let n be a positive integer greater than $2\epsilon^{-1}$. Then we define the set $\mathscr{G}$ to be $\mathscr{G} = \bigcap_i \mathscr{F}(1+\frac{1}{2}in^{-1}, 1+\frac{1}{2}(i+1)n^{-1})$, $i = 0, 1, \ldots, 2n$. $\mathscr{G}$ will be a neighbourhood of p and will be contained in $\mathscr{F}(a, a+n^{-1})$ for any $a \in [1,2]$. Therefore $I^-(q, \mathscr{M}, \mathbf{h}(a)) - \bar{\mathscr{B}}$ will be contained in

$$I^-(p, \mathscr{M}, \mathbf{h}(a+n^{-1})) - \bar{\mathscr{B}} \quad \text{for} \quad q \in \mathscr{G} \quad \text{and} \quad a \in [1,2].$$

Thus
$$\theta(q,a) \leqslant \theta(p, a+\tfrac{1}{2}\epsilon) + \tfrac{1}{2}\epsilon$$

and so $\bar{\theta}(q) \leqslant \bar{\theta}(p) + \epsilon$, showing that $\bar{\theta}$ is upper semi-continuous. The proof that it is lower semi-continuous is similar. To obtain a differentiable function one can average $\bar{\theta}$ over a neighbourhood of each point with a suitable smoothing function. By taking the neighbourhood small enough one can obtain a function f which has everywhere a timelike gradient in the metric $\mathbf{g}$. Details of this smoothing procedure are given in Seifert (1968). □

The spacelike surfaces $\{f = \text{constant}\}$ may be thought of as surfaces of simultaneity in space–time, though of course they are not unique. If they are all compact they are all diffeomorphic to each other, but this is not necessarily true if some of them are non-compact.

6.5 Cauchy developments

In Newtonian theory there is instantaneous action-at-a-distance and so in order to predict events at future points in space–time one has to know the state of the entire universe at the present time and also to assume some boundary conditions at infinity, such as that the potential goes to zero. In relativity theory, on the other hand, it follows from postulate (*a*) of § 3.2 that events at different points of space–time can be causally related only if they can be joined by a non-spacelike curve. Thus a knowledge of the appropriate data on a closed set $\mathscr{S}$ (if one knew data on an open set, that on its closure would follow by continuity) would determine events in a region $D^+(\mathscr{S})$ to the future of $\mathscr{S}$ called the *future Cauchy development* or *domain of dependence* of $\mathscr{S}$, and defined as the set of all points $p \in \mathscr{M}$ such that every past-inextendible non-spacelike curve through p intersects $\mathscr{S}$ (N.B. $D^+(\mathscr{S}) \supset \mathscr{S}$).

Penrose (1966, 1968) defines the Cauchy development of $\mathscr{S}$ slightly differently, as the set of all points $p \in \mathscr{M}$ such that every past-inextendible timelike curve through p intersects $\mathscr{S}$. We shall denote this set by $\tilde{D}^+(\mathscr{S})$. One has the following result:

Proposition 6.5.1
$\tilde{D}^+(\mathscr{S}) = \overline{D^+}(\mathscr{S})$.

Clearly $\tilde{D}^+(\mathscr{S}) \supset D^+(\mathscr{S})$. If $q \in \mathscr{M} - \tilde{D}^+(\mathscr{S})$ there is a neighbourhood $\mathscr{U}$ of q which does not intersect $\mathscr{S}$. From q there is a past-inextendible curve λ which does not intersect $\mathscr{S}$. If $r \in \lambda \cap I^-(q, \mathscr{U})$ then $I^+(r, \mathscr{U})$ is an open neighbourhood of q in $\mathscr{M} - \tilde{D}^+(\mathscr{S})$. Thus $\mathscr{M} - \tilde{D}^+(\mathscr{S})$ is open and the set $\tilde{D}^+(\mathscr{S})$ is closed. Suppose there were a point $p \in \tilde{D}^+(\mathscr{S})$ which had a neighbourhood $\mathscr{V}$ which did not intersect $D^+(\mathscr{S})$. Choose a point $x \in I^-(p, \mathscr{V})$. From x there would be a past-inextendible non-spacelike curve γ which did not intersect $\mathscr{S}$. Let y_n be a sequence of points on γ which did not converge to any point and which were such that y_{n+1} was to the past of y_n. Let $\mathscr{W}_n$ be convex normal neighbourhoods of the corresponding points y_n such that $\mathscr{W}_{n+1}$ did not intersect $\mathscr{W}_n$. Let z_n be a sequence of points such that

$$z_{n+1} \in I^+(y_{n+1}, \mathscr{W}_{n+1}) \cap I^-(z_n, \mathscr{M} - \mathscr{S}).$$

There would be an inextendible timelike curve from p which passed through each point z_n and which did not intersect $\mathscr{S}$. This would contradict $p \in \tilde{D}^+(\mathscr{S})$. Thus $\tilde{D}^+(\mathscr{S})$ is contained in the closure of $D^+(\mathscr{S})$, and so $\tilde{D}^+(\mathscr{S}) = \overline{D^+(\mathscr{S})}$. □

The future boundary of $D^+(\mathscr{S})$, that is $\overline{D^+(\mathscr{S})} - I^-(D^+(\mathscr{S}))$, marks the limit of the region that can be predicted from knowledge of data on $\mathscr{S}$. We call this closed achronal set the *future Cauchy horizon* of $\mathscr{S}$ and denote it by $H^+(\mathscr{S})$. As figure 43 shows, it will intersect $\mathscr{S}$ if $\mathscr{S}$ is null or if $\mathscr{S}$ has an 'edge'. To make this precise we define edge $(\mathscr{S})$ for an achronal set $\mathscr{S}$ as the set of all points $q \in \overline{\mathscr{S}}$ such that in every neighbourhood $\mathscr{U}$ of q there are points $p \in I^-(q, \mathscr{U})$ and $r \in I^+(q, \mathscr{U})$ which can be joined by a timelike curve in $\mathscr{U}$ which does not intersect $\mathscr{S}$. By an argument similar to that in proposition 6.3.1 it follows that if edge $(\mathscr{S})$ is empty for a non-empty achronal set $\mathscr{S}$, then $\mathscr{S}$ is a three-dimensional imbedded C^{1-} submanifold.

Proposition 6.5.2
For a closed achronal set $\mathscr{S}$,

$$\text{edge}\,(H^+(\mathscr{S})) = \text{edge}\,(\mathscr{S}).$$

Let $\mathscr{U}_n$ be a sequence of neighbourhoods of a point $q \in \text{edge}\,(H^+(\mathscr{S}))$

such that any neighbourhood of q encloses all the $\mathscr{U}_n$ for n sufficiently large. In each $\mathscr{U}_n$ there will be points $p_n \in I^-(q, \mathscr{U}_n)$ and $r_n \in I^+(q, \mathscr{U}_n)$ which can be joined by a timelike curve λ_n which does not intersect $H^+(\mathscr{S})$. This means that λ_n cannot intersect $\overline{D^+(\mathscr{S})}$. By proposition 6.5.1, $q \in \tilde{D}^+(\mathscr{S})$ and so $I^-(q) \subset I^-(\tilde{D}^+(\mathscr{S})) \subset I^-(\mathscr{S}) \cup \tilde{D}^+(\mathscr{S})$. Thus p_n must lie in $I^-(\mathscr{S})$. Also every timelike curve from q which is inextendible in the past direction must intersect $\mathscr{S}$. Therefore for each n, there

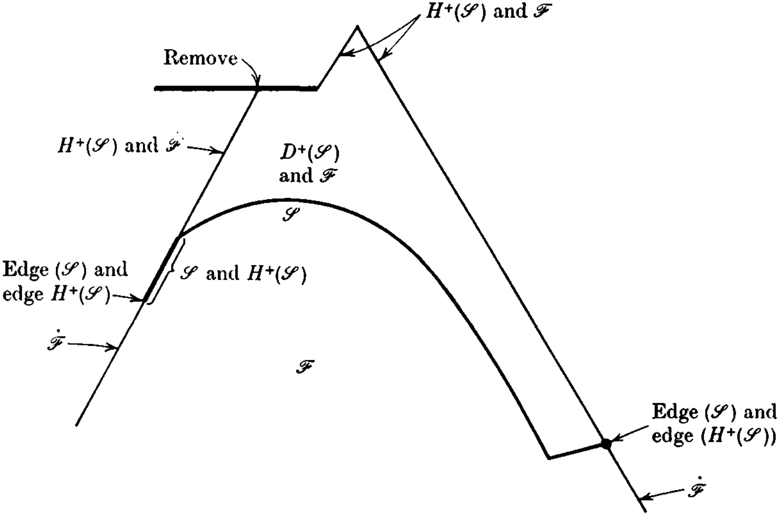
$H^+(\mathscr{S})$ and $\dot{\mathscr{F}}$
Remove
$H^+(\mathscr{S})$ and $\dot{\mathscr{F}}$
$D^+(\mathscr{S})$ and $\mathscr{F}$
$\mathscr{S}$
Edge ($\mathscr{S}$) and edge $H^+(\mathscr{S})$
$\mathscr{S}$ and $H^+(\mathscr{S})$
$\dot{\mathscr{F}}$
$\mathscr{F}$
Edge ($\mathscr{S}$) and edge ($H^+(\mathscr{S})$)
$\dot{\mathscr{F}}$

FIGURE 43. The future Cauchy development $D^+(\mathscr{S})$ and future Cauchy horizon $H^+(\mathscr{S})$ of a closed set $\mathscr{S}$ which is partly null and partly spacelike. Note that $H^+(\mathscr{S})$ is not necessarily connected. Null lines are at ± 45° and a strip has been removed.

must be a point of $\mathscr{S}$ on every timelike curve in $\mathscr{U}_n$ between q and p_n and so q must lie in $\bar{\mathscr{F}}$. As the curves λ_n do not intersect $\mathscr{S}$, q lies in edge ($\mathscr{S}$). The proof the other way round is similar. □

Proposition 6.5.3

Let $\mathscr{S}$ be a closed achronal set. Then $H^+(\mathscr{S})$ is generated by null geodesic segments which either have no past endpoints or have past endpoints at edge ($\mathscr{S}$).

The set $\mathscr{F} \equiv \tilde{D}^+(\mathscr{S}) \cup I^-(\mathscr{S})$ is a past set. Thus by proposition 6.3.1 $\dot{\mathscr{F}}$ is an achronal C^{1-} manifold. $H^+(\mathscr{S})$ is a closed subset of $\dot{\mathscr{F}}$. Let q be a point of $H^+(\mathscr{S})-\text{edge}(\mathscr{S})$. If q is not in $\mathscr{S}$ then $q \in I^+(\mathscr{S})$ since $q \in \tilde{D}^+(\mathscr{S})$. As $\mathscr{S}$ is achronal one can find a convex normal neighbourhood $\mathscr{W}$ of q which does not intersect $I^-(\mathscr{S})$. Alternatively if q is in $\mathscr{S}$, let $\mathscr{W}$ be a convex normal neighbourhood of q such that no point of $I^+(q, \mathscr{W})$ can be joined to any point in $I^-(q, \mathscr{W})$ by a timelike curve in $\mathscr{W}$ which does not intersect $\mathscr{S}$. In either case, if p is any point in $I^+(q)$ there must be a past-directed timelike curve from p to some point of $\mathscr{M}-\mathscr{F}-\mathscr{W}$ since otherwise p would be in $D^+(\mathscr{S})$. Therefore by condition (i) of lemma 6.3.2, applied to the future set $\mathscr{M}-\mathscr{F}$, $q \in \dot{\mathscr{F}}_N \cup \mathscr{F}_+$. □

Corollary

If edge $(\mathscr{S})$ vanishes, then $H^+(\mathscr{S})$, if non-empty, is an achronal three-dimensional imbedded C^{1-} manifold which is generated by null geodesic segments which have no past endpoint.

We shall call an acausal set $\mathscr{S}$ with no edge, a *partial Cauchy surface*. That is, a partial Cauchy surface is a spacelike hypersurface which no non-spacelike curve intersects more than once. Suppose there were a connected spacelike hypersurface $\mathscr{S}$ (with no edge) which some non-spacelike curve λ intersected at points p_1 and p_2. Then one could join p_1 and p_2 by a curve μ in $\mathscr{S}$ and $\mu \cup \lambda$ would be a closed curve which crossed $\mathscr{S}$ once only. This curve could not be continuously deformed to zero since such a deformation could change the number of times it crossed $\mathscr{S}$ by an even number only. Thus $\mathscr{M}$ could not be simply connected. This means we could 'unwrap' $\mathscr{M}$ by going to the simply connected universal covering manifold $\tilde{\mathscr{M}}$ in which each connected component of the image of $\mathscr{S}$ is a spacelike hypersurface (with no edge) and is therefore a partial Cauchy surface in $\tilde{\mathscr{M}}$ (figure 44). However going to the universal covering manifold may unwrap $\mathscr{M}$ more than is required to obtain a partial Cauchy surface and may result in

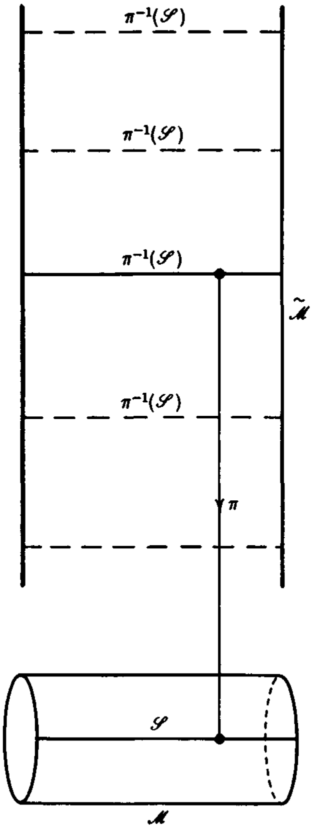

FIGURE 44. $\mathscr{S}$ is a connected spacelike hypersurface without edge in $\mathscr{M}$. It is not a partial Cauchy surface; however each image $\pi^{-1}(\mathscr{S})$ of $\mathscr{S}$ in the universal covering manifold $\tilde{\mathscr{M}}$ of $\mathscr{M}$, is a partial Cauchy surface in $\tilde{\mathscr{M}}$.

the partial Cauchy surface being non-compact even though $\mathscr{S}$ was compact. For the purposes of the following chapters we would like a covering manifold which unwrapped $\mathscr{M}$ sufficiently so that each connected component of the image of $\mathscr{S}$ was a partial Cauchy surface but so that each such component remained homeomorphic to $\mathscr{S}$. Such a covering manifold may be obtained in at least two different ways.

Recall that the universal covering manifold may be defined as the set of all pairs of the form $(p, [\lambda])$ where $p \in \mathscr{M}$ and where $[\lambda]$ is an equivalence class of curves in $\mathscr{M}$ from some fixed point $q \in \mathscr{M}$ to p, which are homotopic modulo q and p. The covering manifold $\mathscr{M}_H$ is defined as the set of all pairs $(p, [\lambda])$ where now $[\lambda]$ is an equivalence class of curves from $\mathscr{S}$ to p homotopic modulo $\mathscr{S}$ and p (i.e. the end-points on $\mathscr{S}$ can be slid around). $\mathscr{M}_H$ may be characterized as the largest covering manifold such that each connected component of the image of $\mathscr{S}$ is homeomorphic to $\mathscr{S}$. The covering manifold $\mathscr{M}_G$ (Geroch (1967*b*)) is defined as the set of all pairs $(p, [\lambda])$ where this time $[\lambda]$ is an equivalence class of curves from a fixed point q to p which cross $\mathscr{S}$ the same number of times, crossings in the future direction being counted positive and those in the past direction, negative. $\mathscr{M}_G$ may be characterized as the smallest covering manifold in which each connected component of the image of $\mathscr{S}$ divides the manifold into two parts. In each case the topological and differential structure of the covering manifold is fixed by requiring that the projection which maps $(p, [\lambda])$ to p is locally a diffeomorphism.

Define $D(\mathscr{S}) = D^+(\mathscr{S}) \cup D^-(\mathscr{S})$. A partial Cauchy surface $\mathscr{S}$ is said to be a global Cauchy surface (or simply, a *Cauchy surface*) if $D(\mathscr{S})$ equals $\mathscr{M}$. That is, a Cauchy surface is a spacelike hypersurface which every non-spacelike curve intersects exactly once. The surfaces $\{x^4 = \text{constant}\}$ are examples of Cauchy surfaces in Minkowski space, but the hyperboloids

$$\{(x^4)^2-(x^3)^2-(x^2)^2-(x^1)^2 = \text{constant}\}$$

are only partial Cauchy surfaces since the past or future null cones of the origin are Cauchy horizons for these surfaces (see § 5.1 and figure 13). Being a Cauchy surface is a property not only of the surface itself but also of the whole space–time in which it is imbedded. For example, if one cuts a single point out of Minkowski space, the resultant space–time admits no Cauchy surface at all.

If there were a Cauchy surface for $\mathscr{M}$, one could predict the state of the universe at any time in the past or future if one knew the relevant data on the surface. However one could not know the data unless one was to the future of every point in the surface, which would be impossible in most cases. There does not seem to be any physically compelling reason for believing that the universe admits a Cauchy surface; in fact there are a number of known exact solutions of the Einstein field equations which do not, among them the anti-de Sitter space, plane waves, Taub–NUT space and Reissner–Nordström solution, all described in chapter 5. The Reissner–Nordström solution (figure 25) is a specially interesting case: the surface $\mathscr{S}$ shown is adequate for predicting events in the exterior regions I where $r > r_+$ and in the neighbouring region II where $r_- < r < r_+$, but then there is a Cauchy horizon at $r = r_-$. Points in the neighbouring region III are not in $D^+(\mathscr{S})$ since there are non-spacelike curves which are inextendible in the past direction and which do not cross $r = r_-$ but approach the points i^+ (which may be considered to be at infinity) or the singularity at $r = 0$ (which cannot be considered to be in the space–time; see § 8.1). There could be extra information coming in from infinity or from the singularity which would upset any predictions made simply on the basis of data on $\mathscr{S}$. Thus in General Relativity one's ability to predict the future is limited both by the difficulty of knowing data on the whole of a spacelike surface and by the possibility that even if one did it would still be insufficient. Nevertheless despite these limitations one can still predict the occurrence of singularities under certain conditions.

6.6 Global hyperbolicity

Closely related to Cauchy developments is the property of global hyperbolicity (Leray (1952)). A set $\mathscr{N}$ is said to be *globally hyperbolic* if the strong causality assumption holds on $\mathscr{N}$ and if for any two points $p, q \in \mathscr{N}$, $J^+(p) \cap J^-(q)$ is compact and contained in $\mathscr{N}$. In a sense this can be thought of as saying that $J^+(p) \cap J^-(q)$ does not contain any points on the edge of space–time, i.e. at infinity or at a singularity. The reason for the name 'global hyperbolicity' is that on $\mathscr{N}$, the wave equation for a δ-function source at $p \in \mathscr{N}$ has a unique solution which vanishes outside $\mathscr{N} - J^+(p, \mathscr{N})$ (see chapter 7).

Recall that $\mathscr{N}$ is said to be causally simple if for every compact set $\mathscr{K}$ contained in $\mathscr{N}$, $J^+(\mathscr{K}) \cap \mathscr{N}$ and $J^-(\mathscr{K}) \cap \mathscr{N}$ are closed in $\mathscr{N}$.

Proposition 6.6.1

An open globally hyperbolic set $\mathscr{N}$ is causally simple.

Let p be any point of $\mathscr{N}$. Suppose there were a point

$$q \in (\overline{J^+(p)} - J^+(p)) \cap \mathscr{N}.$$

As $\mathscr{N}$ is open, there would be a point $r \in (I^+(q) \cap \mathscr{N})$. But then $q \in \overline{J^+(p) \cap J^-(r)}$, which is impossible as $J^+(p) \cap J^-(r)$ would be compact and therefore closed. Thus $J^+(p) \cap \mathscr{N}$ and $J^-(p) \cap \mathscr{N}$ are closed in $\mathscr{N}$.

Now suppose there exists a point $q \in (\overline{J}^+(\mathscr{K}) - J^+(\mathscr{K})) \cap \mathscr{N}$. Let q_n be an infinite sequence of points in $I^+(q) \cap \mathscr{N}$ converging to q, with $q_{n+1} \in I^-(q_n)$. For each n, $J^-(q_n) \cap \mathscr{K}$ would be a compact non-empty set. Therefore $\bigcap_n \{J^-(q_n) \cap \mathscr{K}\}$ would be a non-empty set. Let p be a point of this set. Then $J^+(p)$ would contain q_n for all n. But $J^+(p)$ is closed. Therefore $J^+(p)$ contains q. □

Corollary

If $\mathscr{K}_1$ and $\mathscr{K}_2$ are compact sets in $\mathscr{N}$, $J^+(\mathscr{K}_1) \cap J^-(\mathscr{K}_2)$ is compact.

One can find a finite number of points $p_i \in \mathscr{N}$ such that

$$\{\bigcup_i J^+(p_i)\} \supset \mathscr{K}_1.$$

Similarly, there will be a finite number of points q_j with $\mathscr{K}_2$ contained in

$$\bigcup_j J^-(q_j).$$

Then $J^+(\mathscr{K}_1) \cap J^-(\mathscr{K}_2)$ will be contained in

$$\bigcup_{i,j} \{J^+(p_i) \cap J^-(q_j)\}$$

and will be closed. □

Leray (1952) did not, in fact, give the above definition of global hyperbolicity but an equivalent one which we shall present: for points $p, q \in \mathscr{M}$ such that strong causality holds on $J^+(p) \cap J^-(q)$, we define $C(p,q)$ to be the space of all (continuous) non-space-like curves from p to q, regarding two curves $\gamma(t)$ and $\lambda(u)$ as representing the same point of $C(p,q)$ if one is a reparametrization of the other, i.e. if there is a continuous monotonic function $f(u)$ such that $\gamma(f(u)) = \lambda(u)$. ($C(p,q)$ can be defined even when the strong causality condition does not hold on $J^+(p) \cap J^-(q)$, but we shall only be interested in the case in which its does hold.) The topology of $C(p,q)$ is defined by saying that a neighbourhood of γ in $C(p,q)$ consists of all the curves in $C(p,q)$ whose points in $\mathscr{M}$ lie in a neighbourhood $\mathscr{W}$ of the points of γ in $\mathscr{M}$ (figure 45). Leray's definition is that an open set $\mathscr{N}$ is globally hyperbolic if $C(p,q)$ is compact for all $p,q \in \mathscr{N}$. These definitions are equivalent, as is shown by the following result.

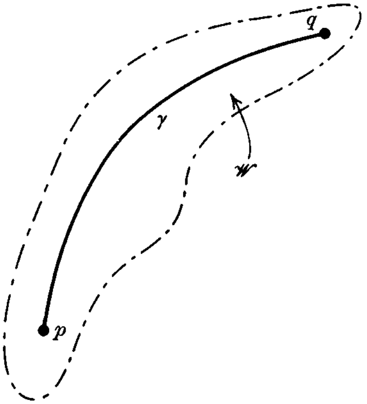

FIGURE 45. A neighbourhood $\mathscr{W}$ of the points of γ in $\mathscr{M}$. A neighbourhood of γ in $C(p, q)$ consists of all non-spacelike curves from p to q whose points lie in $\mathscr{W}$.

Proposition 6.6.2 (*Seifert* (1967), *Geroch* (1970*b*)).

Let strong causality hold on an open set $\mathscr{N}$ such that

$$\mathscr{N} = J^-(\mathscr{N}) \cap J^+(\mathscr{N}).$$

Then $\mathscr{N}$ is globally hyperbolic if and only if $C(p,q)$ is compact for all $p,q \in \mathscr{N}$.

Suppose first that $C(p,q)$ is compact. Let r_n be an infinite sequence of points in $J^+(p) \cap J^-(q)$ and let λ_n be a sequence of non-spacelike curves from p to q through the corresponding r_n. As $C(p,q)$ is compact, there will be a curve λ to which some subsequence λ'_n converges in the topology on $C(p,q)$. Let $\mathscr{U}$ be a neighbourhood of λ in $\mathscr{M}$ such that $\overline{\mathscr{U}}$ is compact. Then $\mathscr{U}$ will contain all λ'_n and hence all r'_n for n sufficiently large, and so there will be a point $r \in \mathscr{U}$ which is a limit point of the r'_n. Clearly r lies on λ. Thus every infinite sequence in $J^+(p) \cap J^-(q)$ has a limit point in $J^+(p) \cap J^-(q)$. Hence $J^+(p) \cap J^-(q)$ is compact.

Conversely, suppose $J^+(p) \cap J^-(q)$ is compact. Let λ_n be an infinite sequence of non-spacelike curves from p to q. By lemma 6.2.1 applied to the open set $\mathscr{M} - q$, there will be a future-directed non-spacelike curve λ from p which is inextendible in $\mathscr{M} - q$, and is such that there is a subsequence λ'_n which converges to r for every $r \in \lambda$. The curve λ must have a future endpoint at q since by proposition 6.4.7 it cannot be totally future imprisoned in the compact set $J^+(p) \cap J^-(q)$, and it cannot leave the set except at q.

Let $\mathscr{U}$ be any neighbourhood of λ in $\mathscr{M}$ and let r_i $(1 \leqslant i \leqslant k)$ be a finite set of points on λ such that $r_1 = p$, $r_k = q$ and each r_i has a neighbourhood $\mathscr{V}_i$ with $J^+(\mathscr{V}_i) \cap J^-(\mathscr{V}_{i+1})$ contained in $\mathscr{U}$. Then for sufficiently large n, λ'_n will be contained in $\mathscr{U}$. Thus λ'_n converge to λ in the topology on $C(p, q)$ and so $C(p, q)$ is compact. □

The relation between global hyperbolicity and Cauchy developments is given by the following results.

Proposition 6.6.3

If $\mathscr{S}$ is a closed achronal set, then $\operatorname{int}(D(\mathscr{S})) \equiv D(\mathscr{S}) - \dot{D}(\mathscr{S})$, if non-empty, is globally hyperbolic.

We first establish a number of lemmas.

Lemma 6.6.4

If $p \in D^+(\mathscr{S}) - H^+(\mathscr{S})$, then every past-inextendible non-spacelike curve through p intersects $I^-(\mathscr{S})$.

Let p be in $D^+(\mathscr{S}) - H^+(\mathscr{S})$ and let γ be a past-inextendible non-spacelike curve through p. Then one can find a point $q \in D^+(\mathscr{S}) \cap I^+(p)$ and a past-inextendible non-spacelike curve λ through q such that for each point $x \in \lambda$ there is a point $y \in \gamma$ with $y \in I^-(x)$. As λ will intersect $\mathscr{S}$ at some point x_1 there will be a $y_1 \in \gamma \cap I^-(\mathscr{S})$. □

Corollary

If $p \in \operatorname{int}(D(\mathscr{S}))$ then every inextendible non-spacelike curve through p intersects $I^-(\mathscr{S})$ and $I^+(\mathscr{S})$.

$\operatorname{int}(D(\mathscr{S})) = D(\mathscr{S}) - \{H^+(\mathscr{S}) \cup H^-(\mathscr{S})\}$. If $p \in I^+(\mathscr{S})$ or $I^-(\mathscr{S})$ the result follows immediately. If $p \in D^+(\mathscr{S}) - I^+(\mathscr{S})$ then $p \in \mathscr{S} \subset D^-(\mathscr{S})$ and the result again follows. □

Lemma 6.6.5

The strong causality condition holds on $\operatorname{int} D(\mathscr{S})$.

Suppose there were a closed non-spacelike curve λ through $p \in \operatorname{int}(D(\mathscr{S}))$. By the previous result there would be points $q \in \lambda \cap I^-(\mathscr{S})$ and $r \in \lambda \cap I^+(\mathscr{S})$. As $r \in J^-(q)$, it would also be in $I^-(\mathscr{S})$

which would contradict the fact that $\mathscr{S}$ is achronal. Thus the causality condition holds on int $(D(\mathscr{S}))$. Now suppose that the strong causality condition did not hold at p. Then as in lemma 6.4.6 there would be an infinite sequence of future-directed non-spacelike curves λ_n which converged to an inextendible null geodesic γ through p. There would be points $q \in \gamma \cap I^-(\mathscr{S})$ and $r \in \gamma \cap I^+(\mathscr{S})$ and so there would be some λ_n which intersected $I^+(\mathscr{S})$ and then $I^-(\mathscr{S})$, which would contradict the fact that $\mathscr{S}$ was achronal. □

Proof of proposition 6.6.3. We wish to show that $C(p,q)$ is compact for $p, q \in \text{int}(D(\mathscr{S}))$. Consider first the case that $p, q \in I^-(\mathscr{S})$ and suppose $p \in J^-(q)$. Let λ_n be an infinite sequence of non-spacelike curves from q to p. By lemma 6.2.1 there will be a future-directed non-spacelike limit curve from p which is inextendible in $\mathscr{M} - q$. This must have a future endpoint at q since otherwise it would intersect $\mathscr{S}$ which would be impossible as $q \in I^-(\mathscr{S})$. Consider now the case that $p \in J^-(\mathscr{S})$, $q \in J^+(\mathscr{S}) \cap J^+(p)$. If the limit curve λ has an endpoint at q, it is the desired limit point in $C(p,q)$. If it does not have an endpoint at q, it would contain a point $y \in I^+(\mathscr{S})$ since it is inextendible in $\mathscr{M} - q$. Let λ'_n be a subsequence which converges to r for every point r on λ between p and y. Let $\hat{\lambda}$ be a past-directed limit curve from q of the λ'_n. If $\hat{\lambda}$ has a past endpoint at p, it would be the desired limit point in $C(p,q)$. If $\hat{\lambda}$ passed through y, it could be joined up with λ to provide a non-spacelike curve from p to q which would be the desired limit point in $C(p,q)$. Suppose $\hat{\lambda}$ does not have endpoint at p and does not pass through y. Then it would contain some point $z \in I^-(\mathscr{S})$. Let λ''_n be a subsequence of the λ'_n which converges to r for every point r on $\hat{\lambda}$ between q and z. Let $\mathscr{V}$ be an open neighbourhood of $\hat{\lambda}$ which does not contain y. Then for sufficiently large n, all $\lambda''_n \cap J^+(\mathscr{S})$ would be contained in $\mathscr{V}$. This would be impossible as y is a limit point of the λ''_n. Thus there will be a non-spacelike curve from p to q which is a limit point of the λ_n in $C(p,q)$.

The cases $p, q \in I^-(\mathscr{S})$ and $p \in J^-(\mathscr{S})$, $q \in J^+(\mathscr{S})$ together with their duals cover all possible combinations. Thus in all cases we get a non-spacelike curve from p to q which is a limit point of the λ_n in the topology on $C(p,q)$. □

By a similar procedure one can prove:

Proposition 6.6.6

If $q \in \text{int}(D(\mathscr{S}))$, then $J^+(\mathscr{S}) \cap J^-(q)$ is compact or empty. □

To show that the whole of $D(\mathscr{S})$ and not merely its interior is globally hyperbolic, one has to impose some extra conditions.

Proposition 6.6.7

If $\mathscr{S}$ is a closed achronal set such that $J^+(\mathscr{S}) \cap J^-(\mathscr{S})$ is both strongly causal and either

(1) acausal (this is the case if and only if $\mathscr{S}$ is acausal), or

(2) compact,

then $D(\mathscr{S})$ is globally hyperbolic.

Suppose that strong causality did not hold at some point $q \in D(\mathscr{S})$. Then by an argument similar to lemma 6.6.5, there would be an inextendible null geodesic through q at each point of which strong causality did not hold. This is impossible, since it would intersect $\mathscr{S}$. Therefore strong causality holds on $D(\mathscr{S})$.

If $p, q \in I^-(\mathscr{S})$, the argument of proposition 6.6.3 holds. If $p \in J^-(\mathscr{S})$, $q \in J^+(\mathscr{S})$ one can as in proposition 6.6.3 construct a future-directed limit curve λ from p and a past-directed limit curve $\hat{\lambda}$ from q, and choose a subsequence λ''_n which converges to r for every point r on λ or $\hat{\lambda}$. In case (1), λ would intersect $\mathscr{S}$ in a single point x. Any neighbourhood of x would contain points of λ''_n for n sufficiently large, and so would contain x''_n, defined as $\lambda''_n \cap \mathscr{S}$, since $\mathscr{S}$ is achronal. Therefore x''_n would converge to x. Similarly x''_n would converge to $\hat{x} \equiv \hat{\lambda} \cap \mathscr{S}$. Thus $\hat{x} = x$ and so one could join λ and $\hat{\lambda}$ to give a non-spacelike limit curve in $C(p, q)$.

In case (2), suppose that λ did not have a future endpoint at q. Then λ would leave $J^-(\mathscr{S})$ since it would intersect $\mathscr{S}$ and by proposition 6.4.7 it would have to leave the compact set $J^+(\mathscr{S}) \cap J^-(\mathscr{S})$. Thus one could find a point x on λ which was not in $J^-(\mathscr{S})$. For each n, choose a point $x''_n \in \mathscr{S} \cap \lambda''_n$. Since $\mathscr{S}$ is compact, there will be some point $y \in \mathscr{S}$ and a subsequence λ'''_n such that the corresponding points x'''_n converge to y. Suppose that y does not lie on λ. Then for sufficiently large n each x'''_n would lie to the future of any neighbourhood $\mathscr{U}$ of x. This would imply $x \in \overline{J^-(\mathscr{S})}$. This is impossible as x is in $J^+(\mathscr{S})$ but is not in the compact set $J^+(\mathscr{S}) \cap J^-(\mathscr{S})$. Therefore λ would pass through y. Similarly $\hat{\lambda}$ would pass through y. One could then join them to obtain a limit curve. □

Proposition 6.6.3 shows that the existence of a Cauchy surface for an open set $\mathscr{N}$ implies global hyperbolicity of $\mathscr{N}$. The following result shows that the converse is also true:

Proposition 6.6.8 (*Geroch* (1970*b*))

If an open set $\mathscr{N}$ is globally hyperbolic, then $\mathscr{N}$, regarded as a manifold, is homeomorphic to $R^1 \times \mathscr{S}$ where $\mathscr{S}$ is a three-dimensional manifold, and for each $a \in R^1$, $\{a\} \times \mathscr{S}$ is a Cauchy surface for $\mathscr{N}$.

As in proposition 6.4.9, put a measure μ on $\mathscr{N}$ such that the total volume of $\mathscr{N}$ in this measure is one. For $p \in \mathscr{N}$ define $f^+(p)$ to be the volume of $J^+(p, \mathscr{N})$ in the measure μ. Clearly $f^+(p)$ is a bounded function on $\mathscr{N}$ which decreases along every future-directed non-spacelike curve. We shall show that global hyperbolicity implies that $f^+(p)$ is continuous on $\mathscr{N}$ so that we do not need to 'average' the volume of the future as in proposition 6.4.9. To do this it will be sufficient to show that $f^+(p)$ is continuous on any non-spacelike curve λ.

Let $r \in \lambda$ and let x_n be an infinite sequence of points on λ strictly to the past of r. Let $\mathscr{F}$ be $\bigcap_n J^+(x_n, \mathscr{N})$. Suppose that $f^+(p)$ was not upper semi-continuous on λ at r. There would be a point $q \in \mathscr{F} - J^+(r, \mathscr{N})$. Then $r \notin J^-(q, \mathscr{N})$; but each $x_n \in J^-(q, \mathscr{N})$ and so $r \in \overline{J^-}(q, \mathscr{N})$, which is impossible as $J^-(q, \mathscr{N})$ is closed in $\mathscr{N}$ by proposition 6.6.1. The proof that it is lower semi-continuous is similar

As p is moved to the future along an inextendible non-spacelike curve λ in $\mathscr{N}$ the value of $f^+(p)$ must tend to zero. For suppose there were some point q which lay to the future of every point of λ. Then the future-directed curve λ would enter and remain within the compact set $J^+(r) \cap J^-(q)$ for any $r \in \lambda$ which would be impossible by proposition 6.4.7 as the strong causality condition holds on $\mathscr{N}$.

Now consider the function $f(p)$ defined on $\mathscr{N}$ by $f(p) = f^-(p)/f^+(p)$. Any surface of constant f will be an acausal set and, by proposition 6.3.1, will be a three-dimensional C^{1-} manifold imbedded in $\mathscr{N}$. It will also be a Cauchy surface for $\mathscr{N}$ since along any non-spacelike curve, f^- will tend to zero in the past and f^+ will tend to zero in the future. One can put a timelike vector field $\mathbf{V}$ on $\mathscr{N}$ and define a continuous map β which takes points of $\mathscr{N}$ along the integral curves of $\mathbf{V}$ to where they intersect the surface $\mathscr{S}$ $(f = 1)$. Then $(\log f(p), \beta(p))$ is a homeomorphism of $\mathscr{N}$ onto $R \times \mathscr{S}$. If one smoothed f as in proposition 6.4.9, one could improve this to a diffeomorphism. □

Thus if the whole of space–time were globally hyperbolic, i.e. if there were a global Cauchy surface, its topology would be very dull.

6.7 The existence of geodesics

The importance of global hyperbolicity for chapter 8 lies in the following result:

Proposition 6.7.1

Let p and q lie in a globally hyperbolic set $\mathscr{N}$ with $q \in J^+(p)$. Then there is a non-spacelike geodesic from p to q whose length is greater than or equal to that of any other non-spacelike curve from p to q.

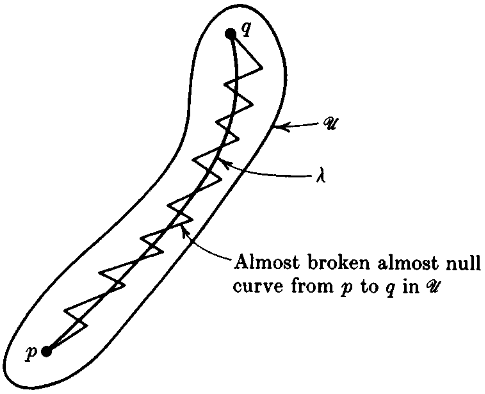
q
$\mathscr{U}$
λ
Almost broken almost null curve from p to q in $\mathscr{U}$
p

FIGURE 46. $\mathscr{U}$ is an open neighbourhood of the timelike curve λ from p to q. There exist in $\mathscr{U}$ timelike curves from p to q which approximate broken null curves and are of arbitrarily small length.

We shall present two proofs of this result: the first, due to Avez (1963) and Seifert (1967), is an argument from the compactness of $C(p,q)$, and the second (applicable only when $\mathscr{N}$ is open) is a procedure whereby the actual geodesic is constructed.

The space $C(p,q)$ contains a dense subset $C'(p,q)$ consisting of all the timelike C^1 curves from p to q. The length of one of these curves λ is defined (cf. § 4.5) as

$$L[\lambda] = \int_p^q (-g(\partial/\partial t, \partial/\partial t))^{\frac{1}{2}}\, dt,$$

where t is a C^1 parameter on λ. The function L is not continuous on $C'(p,q)$ since any neighbourhood of λ contains a zig–zag piecewise almost null curve of arbitrarily small length (figure 46). This lack of continuity arises because we have used the C^0 topology which says that two curves are close if their points in $\mathscr{M}$, but not necessarily their tangent vectors, are close. We could put a C^1 topology on $C'(p,q)$ and so make L continuous but we do not do this because $C'(p,q)$ is not compact; one gets a compact space only when one includes all the continuous non-spacelike curves. Instead, we use the C^0 topology and extend the definition of L to $C(p,q)$.

Because of the signature of the metric, putting wiggles in a timelike curve reduces its length. Thus L is not lower semi-continuous. However one has:

Lemma 6.7.2

L is upper semi-continuous in the C^0 topology on $C'(p,q)$.

Consider a C^1 timelike curve $\lambda(t)$ from p to q, where the parameter t is chosen to be the arc-length from p. In a sufficiently small neighbourhood $\mathscr{U}$ of λ, one can find a function f which is equal to t on λ and is such that the surfaces $\{f = \text{constant}\}$ are spacelike and orthogonal to $\partial/\partial t$ (i.e. $g^{ab}f_{;b}|_\lambda = (\partial/\partial t)^a$). One way to define such an f would be to construct the spacelike geodesics orthogonal to λ. For a sufficiently small neighbourhood $\mathscr{U}$ of λ, they will give a unique mapping of $\mathscr{U}$ to λ, and the value of f at a point in $\mathscr{U}$ can be defined as the value of t at the point on λ into which it is mapped. Any curve μ in $\mathscr{U}$ can be parametrized by f. The tangent vector $(\partial/\partial f)_\mu$ to μ can be expressed as

$$\left(\left(\frac{\partial}{\partial f}\right)_\mu\right)^a = g^{ab}f_{;b}+k^a,$$

where $\mathbf{k}$ is a spacelike vector lying in the surface $\{f = \text{constant}\}$, i.e. $k^a f_{;a} = 0$. Then

$$\begin{aligned} g\left(\left(\frac{\partial}{\partial f}\right)_\mu, \left(\frac{\partial}{\partial f}\right)_\mu\right) &= g^{ab}f_{;a}f_{;b}+g_{ab}k^a k^b \\ &\geqslant g^{ab}f_{;a}f_{;b}. \end{aligned}$$

However on λ, $g^{ab}f_{;a}f_{;b} = -1$. Thus given any $\epsilon > 0$, one can choose $\mathscr{U}' \subset \mathscr{U}$ sufficiently small that on $\mathscr{U}'$, $g^{ab}f_{;a}f_{;b} > -1+\epsilon$. Therefore for any curve μ in $\mathscr{U}'$,

$$L[\mu] \leqslant (1+\epsilon)^{\frac{1}{2}} L[\lambda]. \qquad \square$$

We now define the length of a continuous non-spacelike curve λ from p to q as follows: let $\mathscr{U}$ be a neighbourhood of λ in $\mathscr{M}$ and let $l(\mathscr{U})$ be the least upper bound of the lengths of timelike curves in $\mathscr{U}$ from p to q. Then we define $L[\lambda]$ as the greatest lower bound of $l(\mathscr{U})$ for all neighbourhoods $\mathscr{U}$ of λ in $\mathscr{M}$. This definition of length will work for all curves λ from p to q which have a C^1 timelike curve in every neighbour-

https://doi.org/10.1017/CBO9780511524646.007 Published online by Cambridge University Press

hood, i.e. it will work for all points in $C(p,q)$ which lie in the closure of $C'(p,q)$. By § 4.5, a non-spacelike curve from p to q which is not an unbroken null geodesic curve can be varied to give a piecewise C^1 timelike curve from p to q, and the corners of this curve can be rounded off to give a C^1 timelike curve from p to q. Thus points in $C(p,q)-\overline{C'(p,q)}$ are unbroken null geodesics (containing no conjugate points), and we define their length to be zero.

This definition of L makes it an upper semi-continuous function on the compact space $\overline{C'(p,q)}$. (Actually, as a continuous non-spacelike curve satisfies a local Lipschitz condition, it is differentiable almost everywhere. Thus the length could still be defined as

$$\int(-g(\partial/\partial t,\ \partial/\partial t))^{\frac{1}{2}}\,\mathrm{d}t,$$

and this would agree with the definition above.) If $\overline{C'(p,q)}$ is empty but $C(p,q)$ is non-empty, p and q are joined by an unbroken null geodesic and there are no non-spacelike curves from p to q which are not unbroken null geodesics. If $\overline{C'(p,q)}$ is non-empty, it will contain some point at which L attains its maximum value, i.e. there will be a non-spacelike curve γ from p to q whose length is greater than or equal to that of any other such curve. By proposition 4.5.3, γ must be a geodesic curve as otherwise one could find points $x, y \in \gamma$ which lay in a convex normal coordinate neighbourhood and which could be joined by a geodesic segment of greater length than the portion of γ between x and y. □

For the other, constructive, proof, we first define $d(p,q)$ for $p,q\in\mathscr{M}$ to be zero if $q\notin J^+(p)$ and otherwise to be the least upper bound of the lengths of future-directed piecewise non-spacelike curves from p to q. (Note that $d(p,q)$ may be infinite.) For sets $\mathscr{S}$ and $\mathscr{U}$, we define $d(\mathscr{S},\mathscr{U})$ to be the least upper bound of $d(p,q)$, $p\in\mathscr{S}$, $q\in\mathscr{U}$.

Suppose $q\in I^+(p)$ and that $d(p,q)$ is finite. Then for any $\delta>0$ one can find a timelike curve λ of length $d(p,q)-\frac{1}{2}\delta$ from p to q and a neighbourhood $\mathscr{U}$ of q such that λ can be deformed to give a timelike curve of length $d(p,q)-\delta$ from p to any point $r\in\mathscr{U}$. Thus $d(p,q)$, where finite, is lower semi-continuous. In general $d(p,q)$ is not upper semi-continuous but:

Lemma 6.7.3

$d(p,q)$ is finite and continuous in p and q when p and q are contained in a globally hyperbolic set $\mathscr{N}$.

We shall first prove $d(p,q)$ is finite. Since strong causality holds on the compact set $J^+(p) \cap J^-(q)$, one can cover it with a finite number of local causality sets such that each set contains no non-spacelike curve longer than some bound ϵ. Since any non-spacelike curve from p to q can enter each neighbourhood at most once, it must have finite length.

Now suppose that for $p, q \in \mathscr{N}$, there is a $\delta > 0$ such that every neighbourhood of q contains a point $r \in \mathscr{N}$ such that

$$d(p,r) > d(p,q) + \delta.$$

Let x_n be an infinite sequence of points in $\mathscr{N}$ converging to q such that $d(p, x_n) > d(p,q) + \delta$. Then from each x_n one can find a non-spacelike curve λ_n to p of length $> d(p,q) + \delta$. By lemma 6.2.1 there will be a past-directed non-spacelike curve λ through q which is a limit curve of the λ_n. Let $\mathscr{U}$ be a local causality neighbourhood of q. Then λ cannot intersect $I^-(q) \cap \mathscr{U}$ since if it did one of the λ_n could be deformed to give a non-spacelike curve from p to q of length $> d(p,q)$. Thus $\lambda \cap \mathscr{U}$ must be a null geodesic from q and at each point x of $\lambda \cap \mathscr{U}$, $d(p,x)$ will have a discontinuity greater than δ. This argument can be repeated to show that λ is a null geodesic and at each point $x \in \lambda$, $d(p,x)$ has a discontinuity greater than δ. This shows that λ cannot have an endpoint at p, since by proposition 4.5.3, $d(p,x)$ is continuous on a local causality neighbourhood of p. On the other hand, λ would be inextendible in $\mathscr{M} - p$ and so if it did not have an endpoint at p, it would have to leave the compact set $J^+(p) \cap J^-(q)$ by proposition 6.4.7. This shows that $d(p,q)$ is upper semi-continuous on $\mathscr{N}$. □

In the case that $\mathscr{N}$ is open, one can easily construct the geodesic of maximum length from p to q by using the distance function. Let $\mathscr{U} \subset \mathscr{N}$ be a local causality neighbourhood of p which does not contain q and let $x \in J^+(p) \cap J^-(q)$ be such that $d(p,r) + d(r,q)$, $r \in \dot{\mathscr{U}}$, is maximized for $r = x$. Construct the future-directed geodesic γ from p through x. The relation $d(p,r) + d(r,q) = d(p,q)$ will hold for all points r on γ between p and x. Suppose there were a point $y \in J^-(q) - q$ which was the last point on γ at which this relation held. Let $\mathscr{V} \subset \mathscr{N}$ be a local causality neighbourhood of y which does not contain q and let $z \in J^+(y) \cap J^-(q) \cap \dot{\mathscr{V}}$ be such that $d(y,r) + d(r,q)$, $r \in \dot{\mathscr{V}}$, attains its maximum value $d(y,q)$ for $r = z$. If z did not lie on γ, then

$$d(p,z) > d(p,y) + d(y,z) \quad \text{and} \quad d(p,z) + d(z,q) > d(p,q)$$

which is impossible. This shows that the relation

$$d(p,r) + d(r,q) = d(p,q)$$

https://doi.org/10.1017/CBO9780511524646.007 Published online by Cambridge University Press

must hold for all $r \in \gamma \cap J^-(q)$. As $J^+(p) \cap J^-(q)$ is compact, γ must leave $J^-(q)$ at some point y. Suppose $y \neq q$; then y would lie on a past-directed null geodesic λ from q. Joining γ to λ would give a non-spacelike curve from p to q which could be varied to give a curve longer than $d(p, q)$, which is impossible. Thus γ is a geodesic curve from p to q of length $d(p, q)$. □

Corollary

If $\mathscr{S}$ is a C^2 partial Cauchy surface, then to each point $q \in D^+(\mathscr{S})$ there is a future-directed timelike geodesic curve orthogonal to $\mathscr{S}$ of length $d(\mathscr{S}, q)$, which does not contain any point conjugate to $\mathscr{S}$ between $\mathscr{S}$ and q.

By proposition 6.5.2, $H^+(\mathscr{S})$ and $H^-(\mathscr{S})$ do not intersect $\mathscr{S}$ and so are not in $D(\mathscr{S})$. Thus $D(\mathscr{S}) = \operatorname{int} D(\mathscr{S})$ is globally hyperbolic by proposition 6.6.3. By proposition 6.6.6, $\mathscr{S} \cap J^-(q)$ is compact and so $d(p, q)$, $p \in \mathscr{S}$, will attain its maximum value of $d(\mathscr{S}, q)$ at some point $r \in \mathscr{S}$. There will be a geodesic curve γ from r to q of length $d(\mathscr{S}, q)$ which by lemma 4.5.5 and proposition 4.5.9 must be orthogonal to $\mathscr{S}$ and not contain a point conjugate to $\mathscr{S}$ between $\mathscr{S}$ and q. □

6.8 The causal boundary of space–time

In this section we shall give a brief outline of the method of Geroch, Kronheimer and Penrose (1972) for attaching a boundary to space–time. The construction depends only on the causal structure of $(\mathscr{M}, \mathbf{g})$. This means that it does not distinguish between boundary points at a finite distance (singular points) and boundary points at infinity. In § 8.3 we shall describe a different construction which attaches a boundary which represents only singular points. Unfortunately there does not seem to be any obvious relation between the two constructions.

We shall assume that $(\mathscr{M}, \mathbf{g})$ satisfies the strong causality condition. Then any point p in $(\mathscr{M}, \mathbf{g})$ is uniquely determined by its chronological past $I^-(p)$ or its future $I^+(p)$, i.e.

$$I^-(p) = I^-(q) \Leftrightarrow I^+(p) = I^+(q) \Leftrightarrow p = q.$$

The chronological past $\mathscr{W} \equiv I^-(p)$ of any point $p \in \mathscr{M}$ has the properties:

(1) $\mathscr{W}$ is open;

(2) $\mathscr{W}$ is a past set, i.e. $I^-(\mathscr{W}) \subset \mathscr{W}$;

(3) $\mathscr{W}$ cannot be expressed as the union of two proper subsets which have properties (1) and (2).

We shall call a set with properties (1), (2) and (3) an *indecomposable past set*, abbreviated as IP. (The definition given by Geroch, Kronheimer and Penrose does not include property (1). However it is equivalent to the definition given here, since by 'a past set' they mean a set which equals its chronological past, rather than merely containing it.) One can define an IF, or *indecomposable future set*, similarly.

One can divide IPs into two classes: *proper IPs* (*PIPs*) which are the pasts of points in $\mathscr{M}$, and *terminal IPs* (*TIPs*) which are not the past of any point in $\mathscr{M}$. The idea is to regard these TIPs and the similarly defined TIFs as representing points of the causal boundary (*c-boundary*) of $(\mathscr{M}, \mathbf{g})$. For instance, in Minkowski space one would regard the shaded region in figure 47 (i) as representing the point p on $\mathscr{I}^+$. Note that in this example, the whole of $\mathscr{M}$ is itself a TIP and also a TIF. These can be thought of as representing the points i^+ and i^- respectively. In fact all the points of the conformal boundary of Minkowski space, except i^0, can be represented as TIPs or TIFs. In some cases, such as anti-de Sitter space, where the conformal boundary is timelike, points of the boundary will be represented by both a TIP and a TIF (see figure 47 (ii)).

One can also characterize TIPs as the pasts of future-inextendible timelike curves. This means that one can regard the past $I^-(\gamma)$ of a future-inextendible curve γ as representing the future endpoint of γ on the c-boundary. Another curve γ' has the same endpoint if and only if $I^-(\gamma) = I^-(\gamma')$.

Proposition 6.8.1 (*Geroch, Kronheimer and Penrose*)

A set $\mathscr{W}$ is a TIP if and only if there is a future-inextendible timelike curve γ such that $I^-(\gamma) = \mathscr{W}$.

Suppose first that there is a curve γ such that $I^-(\gamma) = \mathscr{W}$. Let $\mathscr{W} = \mathscr{U} \cup \mathscr{V}$ where $\mathscr{U}$ and $\mathscr{V}$ are open past sets. One wants to show that either $\mathscr{U}$ is contained in $\mathscr{V}$, or $\mathscr{V}$ contained in $\mathscr{U}$. Suppose that, on the contrary, $\mathscr{U}$ is not contained in $\mathscr{V}$ and $\mathscr{V}$ not contained in $\mathscr{U}$. Then one could find a point q in $\mathscr{U} - \mathscr{V}$ and a point r in $\mathscr{V} - \mathscr{U}$. Now $q, r \in I^-(\gamma)$, so there would be points $q', r' \in \gamma$ such that $q \in I^-(q')$ and $r \in I^-(r')$. But whichever of $\mathscr{U}$ or $\mathscr{V}$ contained the futuremost of q', r' would also contain both q and r, which contradicts the original definitions of q and r.

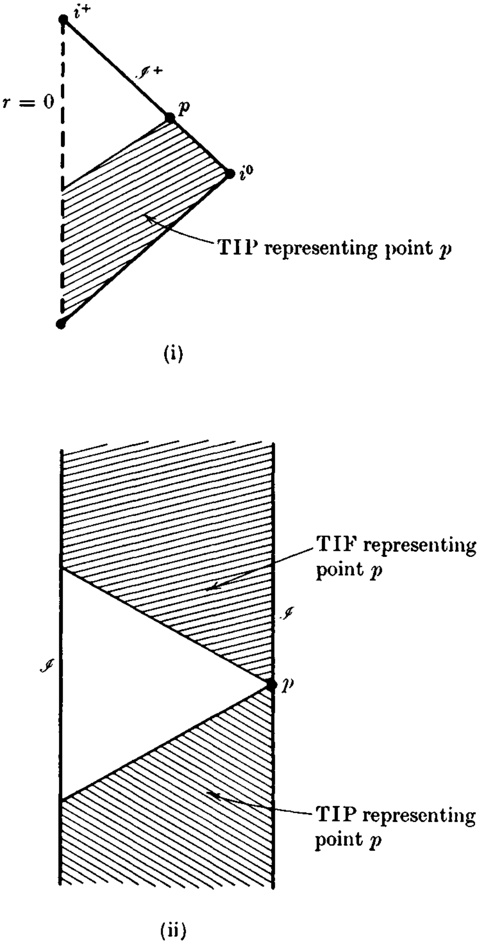
i^+
$\mathscr{I}^+$
$r = 0$
p
i^0
TIP representing point p
(i)
TIF representing point p
$\mathscr{I}$
$\mathscr{I}$
p
TIP representing point p
(ii)

FIGURE 47. Penrose diagrams of Minkowski space and anti-de Sitter space (cf. figures 15 and 20), showing (i) the TIP representing a point p on $\mathscr{I}^+$ in Minkowski space, and (ii) the TIP and the TIF representing a point p on $\mathscr{I}$ in anti-de Sitter space.

Conversely, suppose $\mathscr{W}$ is a TIP. Then one must construct a timelike curve γ such that $\mathscr{W} = I^-(\gamma)$. Now if p is any point of $\mathscr{W}$, then $\mathscr{W} = I^-(\mathscr{W} \cap I^+(p)) \cup I^-(\mathscr{W} - I^+(p))$. However $\mathscr{W}$ is indecomposable, so either $\mathscr{W} = I^-(\mathscr{W} \cap I^+(p))$ or $\mathscr{W} = I^-(\mathscr{W} - I^+(p))$. The point p is not contained in $I^-(\mathscr{W} - I^+(p))$, so the second possibility is eliminated. The conclusion may be restated in the following form: given any pair of points of $\mathscr{W}$, then $\mathscr{W}$ contains a point to the future of both of them. Now choose a countable dense family p_n of points of $\mathscr{W}$. Choose a point

q_0 in $\mathscr{W}$ to the future of p_0. Since q_0 and p_1 are in $\mathscr{W}$, one can choose a point q_1 in $\mathscr{W}$ to the future of both of them. Since q_1 and p_2 are in $\mathscr{W}$, one can choose q_2 in $\mathscr{W}$ to the future of both of them, and so on. Since each point q_n obtained in this way lies in the past of its successor, one can find a timelike curve γ in $\mathscr{W}$ through all the points of the sequence. Now for each point $p \in \mathscr{W}$, the set $\mathscr{W} \cap I^+(p)$ is open and non-empty, and so it must contain at least one of the p_n, since these are dense. But for each k, p_k lies in the past of q_k, whence p itself lies in the past of γ. This shows that every point of $\mathscr{W}$ lies to the past of γ, and so since γ is contained in the open past set $\mathscr{W}$, one must have $\mathscr{W} = I^-(\gamma)$. □

We shall denote by $\hat{\mathscr{M}}$ the set of all IPs of the space $(\mathscr{M}, \mathbf{g})$. Then $\hat{\mathscr{M}}$ represents the points of $\mathscr{M}$ plus a future c-boundary; similarly, $\check{\mathscr{M}}$, the set of all IFs of $(\mathscr{M}, \mathbf{g})$, represents $\mathscr{M}$ plus a past c-boundary. One can extend the causal relations I, J and E to $\hat{\mathscr{M}}$ and $\check{\mathscr{M}}$ in the following way. For each $\mathscr{U}, \mathscr{V} \subset \hat{\mathscr{M}}$, we shall say

$$\mathscr{U} \in J^-(\mathscr{V}, \hat{\mathscr{M}}) \quad \text{if} \quad \mathscr{U} \subset \mathscr{V},$$

$$\mathscr{U} \in I^-(\mathscr{V}, \hat{\mathscr{M}}) \quad \text{if} \quad \mathscr{U} \subset I^-(q) \text{ for some point } q \in \mathscr{V},$$

$$\mathscr{U} \in E^-(\mathscr{V}, \hat{\mathscr{M}}) \quad \text{if} \quad \mathscr{U} \in J^-(\mathscr{V}, \hat{\mathscr{M}}) \quad \text{but not} \quad \mathscr{U} \in I^-(\mathscr{V}, \hat{\mathscr{M}}).$$

With these relations, the IP-space $\hat{\mathscr{M}}$ is a causal space (Kronheimer and Penrose (1967)). There is a natural injective map $I^-: \mathscr{M} \to \hat{\mathscr{M}}$ which sends the point $p \in \mathscr{M}$ into $I^-(p) \in \hat{\mathscr{M}}$. This map is an isomorphism of the causality relation J^- as $p \in J^-(q)$ if and only if $I^-(p) \in J^-(I^-(q), \hat{\mathscr{M}})$. The causality relation is preserved by I^- but not by its inverse, i.e. $p \in I^-(q) \Rightarrow I^-(p) \in I^-(I^-(q), \mathscr{M})$. One can define causal relations on $\check{\mathscr{M}}$ similarly.

The idea now is to write $\hat{\mathscr{M}}$ and $\check{\mathscr{M}}$ in some way to form a space $\mathscr{M}^*$ which has the form $\mathscr{M} \cup \Delta$ where Δ will be called the *c-boundary* of $(\mathscr{M}, \mathbf{g})$. To do so, one needs a method of identifying appropriate IPs and IFs. One starts by forming the space $\mathscr{M}^\#$ which is the union of $\hat{\mathscr{M}}$ and $\check{\mathscr{M}}$, with each PIF identified with the corresponding PIP. In other words, $\mathscr{M}^\#$ corresponds to the points of $\mathscr{M}$ together with the TIPs and TIFs. However as the example of anti-de Sitter space shows, one also wants to identify some TIPs with some TIFs. One way of doing this is to define a topology on $\mathscr{M}^\#$, and then to identify some points of $\mathscr{M}^\#$ to make this topology Hausdorff.

As was mentioned in § 6.4, a basis for the topology of the topological space $\mathscr{M}$ is provided by sets of the form $I^+(p) \cap I^-(q)$. Unfortunately one cannot use a similar method to define a basis for the topology of $\mathscr{M}^\#$ as there may be some points of $\mathscr{M}^\#$ which are not in the chronological past of any points of $\mathscr{M}^\#$. However one can also obtain a topology of $\mathscr{M}$ from a sub-basis consisting of sets of the form $I^+(p)$, $I^-(p)$, $\mathscr{M}-\overline{I^+}(p)$ and $\mathscr{M}-\overline{I^-}(p)$. Following this analogy, Geroch, Kromheimer and Penrose have shown how one can define a topology on $\mathscr{M}^\#$. For an IF $\mathscr{A}\in\check{\mathscr{M}}$, one defines the sets

$$\mathscr{A}^{\text{int}} \equiv \{\mathscr{V} : \mathscr{V}\in\hat{\mathscr{M}} \text{ and } \mathscr{V}\cap\mathscr{A}\neq\varnothing\},$$

and $$\mathscr{A}^{\text{ext}} \equiv \{\mathscr{V} : \mathscr{V}\in\hat{\mathscr{M}} \text{ and } \mathscr{V}=I^-(\mathscr{W}) \Rightarrow I^+(\mathscr{W})\not\subset\mathscr{A}\}.$$

For an IP $\mathscr{B}\in\hat{\mathscr{M}}$, the sets $\mathscr{B}^{\text{int}}$ and $\mathscr{B}^{\text{ext}}$ are defined similarly. The open sets of $\mathscr{M}^\#$ are then defined to be the unions and finite intersections of sets of the form $\mathscr{A}^{\text{int}}$, $\mathscr{A}^{\text{ext}}$, $\mathscr{B}^{\text{int}}$ and $\mathscr{B}^{\text{ext}}$. The sets $\mathscr{A}^{\text{int}}$ and $\mathscr{B}^{\text{int}}$ are the analogues in $\mathscr{M}^\#$ of the sets $I^+(p)$ and $I^-(q)$. If in particular $\mathscr{A} = I^+(p)$ and $\mathscr{V} = I^-(q)$ then $\mathscr{V}\in\mathscr{A}^{\text{int}}$ if and only if $q\in I^+(p)$. However the definitions enable one also to incorporate TIPS into $\mathscr{A}^{\text{int}}$. The sets $\mathscr{A}^{\text{ext}}$ and $\mathscr{B}^{\text{ext}}$ are the analogues of $\mathscr{M}-\overline{I^+(p)}$ and $\mathscr{M}-\overline{I^-(q)}$.

Finally one obtains $\mathscr{M}^*$ by identifying the smallest number of points in the space $\mathscr{M}^\#$ necessary to make it a Hausdorff space. More precisely $\mathscr{M}^*$ is the quotient space $\mathscr{M}^\#/R_h$ where R_h is the intersection of all equivalence relations $R\subset\mathscr{M}^\#\times\mathscr{M}^\#$ for which $\mathscr{M}^\#/R$ is Hausdorff. The space $\mathscr{M}^*$ has a topology induced from $\mathscr{M}^\#$ which agrees with the topology of $\mathscr{M}$ on the subset $\mathscr{M}$ of $\mathscr{M}^*$. In general one cannot extend the differentiable structure of $\mathscr{M}$ to Δ, though one can on part of Δ in a special case which will be described in the next section.

6.9 Asymptotically simple spaces

In order to study bounded physical systems such as stars, one wants to investigate spaces which are asymptotically flat, i.e. whose metrics approach that of Minkowski space at large distances from the system. The Schwarzschild, Reissner–Nordström and Kerr solutions are examples of spaces which have asymptotically flat regions. As we saw in chapter 5, the conformal structure of null infinity in these spaces is similar to that of Minkowski space. This led Penrose (1964, 1965*b*, 1968) to adopt this as a definition of a kind of asymptotic flatness. We shall only consider strongly causal spaces. Penrose does not make the requirement of strong causality. However it simplifies matters and implies no loss of generality in the kind of situation we wish to consider.

A time- and space-orientable space $(\mathscr{M}, \mathbf{g})$ is said to be *asymptotically simple* if there exists a strongly causal space $(\tilde{\mathscr{M}}, \tilde{\mathbf{g}})$ and an imbedding $\theta: \mathscr{M} \to \tilde{\mathscr{M}}$ which imbeds $\mathscr{M}$ as a manifold with smooth boundary $\partial\mathscr{M}$ in $\tilde{\mathscr{M}}$, such that:

(1) there is a smooth (say C^3 at least) function Ω on $\tilde{\mathscr{M}}$ such that on $\theta(\mathscr{M})$, Ω is positive and $\Omega^2\mathbf{g} = \theta_*(\tilde{\mathbf{g}})$ (i.e. $\tilde{\mathbf{g}}$ is conformal to $\mathbf{g}$ on $\theta(\mathscr{M})$);

(2) on $\partial\mathscr{M}$, $\Omega = 0$ and $\mathrm{d}\Omega \neq 0$;

(3) every null geodesic in $\mathscr{M}$ has two endpoints on $\partial\mathscr{M}$.

We shall write $\mathscr{M} \cup \partial\mathscr{M} \equiv \bar{\mathscr{M}}$.

In fact this definition is rather more general than one wants since it includes cosmological models, such as de Sitter space. In order to restrict it to spaces which are asymptotically flat spaces, we will say that a space $(\mathscr{M}, \mathbf{g})$ is *asymptotically empty and simple* if it satisfies conditions (1), (2), and (3), and

(4) $R_{ab} = 0$ on an open neighbourhood of $\partial\mathscr{M}$ in $\bar{\mathscr{M}}$. (This condition can be modified to allow the existence of electromagnetic radiation near $\partial\mathscr{M}$).

The boundary $\partial\mathscr{M}$ can be thought of as being at infinity, in the sense that any affine parameter in the metric $\mathbf{g}$ on a null geodesic in $\mathscr{M}$ attains unboundedly large values near $\partial\mathscr{M}$. This is because an affine parameter v in the metric g is related to an affine parameter $\tilde{v}$ in the metric $\tilde{\mathbf{g}}$ by $\mathrm{d}v/\mathrm{d}\tilde{v} = \Omega^{-2}$. Since $\Omega = 0$ at $\partial\mathscr{M}$, $\int \mathrm{d}v$ diverges.

From conditions (2) and (4) it follows that the boundary $\partial\mathscr{M}$ is a null hypersurface. This is because the Ricci tensor $\tilde{R}_{ab}$ of the metric $\tilde{g}_{ab}$ is related to the Ricci tensor R_{ab} of g_{ab} by

$$\tilde{R}_a{}^b = \Omega^{-2}R_a{}^b - 2\Omega^{-1}(\Omega)_{|ac}\tilde{g}^{bc} + \{-\Omega^{-1}\Omega_{|cd} + 3\Omega^{-2}\Omega_{|c}\Omega_{|d}\}\tilde{g}^{cd}\delta_a{}^b$$

where $|$ denotes covariant differentiation with respect to $\tilde{g}_{ab}$. Thus

$$\tilde{R} = \Omega^{-2}R - 6\Omega^{-1}\Omega_{|cd}\tilde{g}^{cd} + 3\Omega^{-2}\Omega_{|c}\Omega_{|d}\tilde{g}^{cd}.$$

Since the metric $\tilde{g}_{ab}$ is C^3, $\tilde{R}$ is C^1 at $\partial\mathscr{M}$ where $\Omega = 0$. This implies that $\Omega_{|c}\Omega_{|d}\tilde{g}^{cd} = 0$. However by condition (2), $\Omega_{|c} \neq 0$. Thus $\Omega_{|c}\tilde{g}^{cd}$ is a null vector, and the surface $\partial\mathscr{M}$ $(\Omega = 0)$ is a null hypersurface.

In the case of Minkowski space, $\partial\mathscr{M}$ consists of the two null surfaces $\mathscr{I}^+$ and $\mathscr{I}^-$, each of which has the topology $R^1 \times S^2$. (Note that it does not include the points i^0, i^+ and i^- since the conformal boundary is not a smooth manifold at these points.) We shall show that in fact $\partial\mathscr{M}$ has this structure for any asymptotically simple and empty space.

Since $\partial\mathscr{M}$ is a null surface, $\mathscr{M}$ lies locally to the past or future of it. This shows that $\partial\mathscr{M}$ must consist of two disconnected components: $\mathscr{I}^+$ on which null geodesics in $\mathscr{M}$ have their future endpoints, and $\mathscr{I}^-$

on which they have their past endpoints. There cannot be more than two components of $\partial\mathscr{M}$, since there would then be some point $p \in \mathscr{M}$ for which some future-directed null geodesics would go to one component and others to another component. The set of null directions at p going to each component would be open, which is impossible, since the set of future null directions at p is connected.

We next establish an important property.

Lemma 6.9.1

An asymptotically simple and empty space $(\mathscr{M}, \mathbf{g})$ is causally simple.

Let $\mathscr{W}$ be a compact set of $\mathscr{M}$. One wants to show that every null geodesic generator of $\dot{J}^+(\mathscr{W})$ has past endpoint at $\mathscr{W}$. Suppose there were a generator that did not have endpoint there. Then it could not have any endpoint in $\mathscr{M}$, so it would intersect $\mathscr{I}^-$, which is impossible. □

Proposition 6.9.2

An asymptotically simple and empty space $(\mathscr{M}, \mathbf{g})$ is globally hyperbolic.

The proof is similar to that of proposition 6.6.7. One puts a volume element on $\mathscr{M}$ such that the total volume of $\mathscr{M}$ in this measure is unity. Since $(\mathscr{M}, \mathbf{g})$ is causally simple, the functions $f^+(p)$, $f^-(p)$ which are the volumes of $I^+(p)$, $I^-(p)$ are continuous on $\mathscr{M}$. Since strong causality holds on $\mathscr{M}$, $f^+(p)$ will decrease along every future-directed non-spacelike curve. Let λ be a future-inextendible timelike curve. Suppose that $\mathscr{F} = \bigcap_{p \in \lambda} I^+(p)$ was non-empty. Then $\mathscr{F}$ would be a future set and the null generators of the boundary of $\mathscr{F}$ in $\mathscr{M}$ would have no past endpoint in $\mathscr{M}$. Thus they would intersect $\mathscr{I}^-$, which again leads to a contradiction. This shows that $f^+(p)$ goes to zero as p tends to the future on λ. From this it follows that every inextendible non-spacelike curve intersects the surface $\mathscr{H} \equiv \{p : f^+(p) = f^-(p)\}$, which is therefore a Cauchy surface for $\mathscr{M}$. □

Lemma 6.9.3

Let $\mathscr{W}$ be a compact set of an asymptotically empty and simple space $(\mathscr{M}, \mathbf{g})$. Then every null geodesic generator of $\mathscr{I}^+$ intersects $\dot{J}^+(\mathscr{W}, \bar{\mathscr{M}})$ once, where $\dot{\ }$ indicates the boundary in $\bar{\mathscr{M}}$.

Let $p \in \lambda$, where λ is a null geodesic generator of $\mathscr{I}^+$. Then the past set (in $\mathscr{M}$) $J^-(p, \bar{\mathscr{M}}) \cap \mathscr{M}$ must be closed in $\mathscr{M}$, since every null geodesic generator of its boundary must have future endpoint on $\mathscr{I}^+$ at p. Since strong causality holds on $\tilde{\mathscr{M}}$, $\mathscr{M} - J^-(p, \bar{\mathscr{M}})$ will be non-empty. Now suppose that λ were contained in $J^+(\mathscr{W}, \bar{\mathscr{M}})$. Then the past set $\bigcap_{p \in \lambda} (J^-(p, \bar{\mathscr{M}}) \cap \mathscr{M})$ would be non-empty. This would be impossible, since the null generators of the boundary of the set would intersect $\mathscr{I}^+$. Suppose on the other hand that λ did not intersect $J^+(\mathscr{W}, \bar{\mathscr{M}})$. Then $\mathscr{M} - \bigcup_{p \in \lambda} (J^-(p, \bar{\mathscr{M}}) \cap \mathscr{M})$ would be non-empty. This would again lead to a contradiction, as the generators of the boundary of the past set $\bigcup_{p \in \lambda} (J^-(p, \bar{\mathscr{M}}) \cap \mathscr{M})$ would intersect $\mathscr{I}^+$. □

Corollary

$\mathscr{I}^+$ is topologically $R^1 \times (\dot{J}^+(\mathscr{W}, \bar{\mathscr{M}}) \cap \partial\mathscr{M})$.

We shall now show that $\mathscr{I}^+$ (and $\mathscr{I}^-$) and $\mathscr{M}$ are the same topologically as they are for Minkowski space.

Proposition 6.9.4 (*Geroch* (1971))

In an asymptotically simple and empty space $(\mathscr{M}, \mathbf{g})$, $\mathscr{I}^+$ and $\mathscr{I}^-$ are topologically $R^1 \times S^2$, and $\mathscr{M}$ is R^4.

Consider the set N of all null geodesics in $\mathscr{M}$. Since these all intersect the Cauchy surface $\mathscr{H}$, one can define local coordinates on N by the local coordinates and directions of their intersections with $\mathscr{H}$. This makes N into a fibre bundle of directions over $\mathscr{H}$ with fibre S^2. However every null geodesic also intersects $\mathscr{I}^+$. Thus N is also a fibre bundle over $\mathscr{I}^+$. In this case, the fibre is S^2 minus one point which corresponds to the null geodesic generator of $\mathscr{I}^+$ which does not enter $\mathscr{M}$. In other words, the fibre is R^2. Therefore N is topologically $\mathscr{I}^+ \times R^2$. However $\mathscr{I}^+$ is $R^1 \times (\dot{J}^+(\mathscr{W}, \bar{\mathscr{M}}) \cap \partial\mathscr{M})$. This is consistent with $N \approx \mathscr{H} \underset{\sim}{\times} S^2$ only if $\mathscr{H} \approx R^3$ and $\mathscr{I}^+ \approx R^1 \times S^2$. □

Penrose (1965*b*) has shown that this result implies that the Weyl tensor of the metric $\mathbf{g}$ vanishes on $\mathscr{I}^+$ and $\mathscr{I}^-$. This can be interpreted as saying that the various components of the Weyl tensor of the metric $\mathbf{g}$ 'peel off', that is, they go as different powers of the affine parameter on a null geodesic near $\mathscr{I}^+$ or $\mathscr{I}^-$. Further Penrose (1963), Newman and Penrose (1968) have given conservation laws for the energy–momentum as measured from $\mathscr{I}^+$, in terms of integrals on $\mathscr{I}^+$.

The null surfaces $\mathscr{I}^+$ and $\mathscr{I}^-$ form nearly all the *c*-boundary Δ of $(\mathscr{M}, \mathbf{g})$ defined in the previous section. To see this, note first that any point $p \in \mathscr{I}^+$ defines a TIP $I^-(p, \bar{\mathscr{M}}) \cap \mathscr{M}$. Suppose λ is a future-inextendible curve in $\mathscr{M}$. If λ has a future endpoint at $p \in \mathscr{I}^+$, then the TIP $I^-(\lambda)$ is the same as the TIP defined by p. If λ does not have a future endpoint on $\mathscr{I}^+$, then $\mathscr{M} - I^-(\lambda)$ must be empty, since if it were not, the null geodesic generators of $\dot{I}^-(\lambda)$ would intersect $\mathscr{I}^+$ which is impossible as λ does not intersect $\mathscr{I}^+$. The TIPs therefore consist of one for each point of $\mathscr{I}^+$, and one extra TIP, denoted by i^+, which is $\mathscr{M}$ itself. Similarly, the TIFs consist of one for each point of $\mathscr{I}^-$, and one, denoted by i^-, which again is $\mathscr{M}$ itself.

One now wants to verify that one does not have to identify any TIPs or TIFs, i.e. that $\mathscr{M}^\#$ is Hausdorff. It is clear that no two TIPs or TIFs corresponding to $\mathscr{I}^+$ or $\mathscr{I}^-$ are non-Hausdorff separated. If $p \in \mathscr{I}^+$ then one can find $q \in \mathscr{M}$ such that $p \notin I^+(q, \bar{\mathscr{M}})$. Then $(I^+(q, \bar{\mathscr{M}}))^{\text{ext}}$ is a neighbourhood in $\mathscr{M}^\#$ of the TIP $I^-(p, \bar{\mathscr{M}}) \cap \mathscr{M}$, and $(I^+(q, \bar{\mathscr{M}}))^{\text{int}}$ is a disjoint neighbourhood of the TIP i^+. Thus i^+ is Hausdorff separated from every point of $\mathscr{I}^+$. Similarly it is Hausdorff separated from every point of $\mathscr{I}^-$. Thus the c-boundary of any asymptotically simple and empty space $(\mathscr{M}, \mathbf{g})$ is the same as that of Minkowski space–time, consisting of $\mathscr{I}^+$, $\mathscr{I}^-$ and the two points i^+, i^-.

Asymptotically simple and empty spaces include Minkowski space and the asymptotically flat spaces containing bounded objects such as stars which do not undergo gravitational collapse. However they do not include the Schwarzschild, Reissner–Nordström or Kerr solutions, because in these spaces there are null geodesics which do not have endpoints on $\mathscr{I}^+$ or $\mathscr{I}^-$. Nevertheless these spaces do have asymptotically flat regions which are similar to those of asymptotically empty and simple spaces. This suggests that one should define a space $(\mathscr{M}, \mathbf{g})$ to be *weakly asymptotically simple and empty* if there is an asymptotically simple and empty space $(\mathscr{M}', \mathbf{g}')$ and a neighbourhood $\mathscr{U}'$ of $\partial\mathscr{M}'$ in $\mathscr{M}'$ such that $\mathscr{U}' \cap \mathscr{M}'$ is isometric to an open set $\mathscr{U}$ of $\mathscr{M}$. This definition covers all the spaces mentioned above. In the Reissner–Nordström and Kerr solutions there is an infinite sequence of asymptotically flat regions $\mathscr{U}$ which are isometric to neighbourhoods $\mathscr{U}'$ of asymptotically simple spaces. There is thus an infinite sequence of null infinities $\mathscr{I}^+$ and $\mathscr{I}^-$. However we shall consider only one asymptotically flat region in these spaces. One can then regard $(\mathscr{M}, \mathbf{g})$ as being conformally imbedded in a space $(\tilde{\mathscr{M}}, \tilde{\mathbf{g}})$ such that a neighbourhood $\mathscr{U}$ of $\partial\mathscr{M}$ in $\tilde{\mathscr{M}}$ is isometric to $\mathscr{U}'$. The boundary $\partial\mathscr{M}$ consists of a single pair of null surfaces $\mathscr{I}^+$ and $\mathscr{I}^-$.

We shall discuss weakly asymptotically simple and empty spaces in § 9.2 and § 9.3.