8
Space-time singularities

In this chapter, we use the results of chapters 4 and 6 to establish some basic results about space–time singularities. The astrophysical and cosmological implications of these results are considered in the next chapters.

In §8.1, we discuss the problem of defining singularities in space–time. We adopt b-incompleteness, a generalization of the idea of geodesic incompleteness, as an indication that singular points have been cut out of space–time, and characterize two possible ways in which b-incompleteness can be associated with some form of curvature singularity. In §8.2, four theorems are given which prove the existence of incompleteness under a wide variety of situations. In §8.3 we give Schmidt's construction of the b-boundary which represents the singular points of space–time. In §8.4 we prove that the singularities predicted by at least one of the the theorems cannot be just a discontinuity in the curvature tensor. We also show that there is not only one incomplete geodesic, but a three-parameter family of them. In §8.5 we discuss the situation in which the incomplete curves are totally or partially imprisoned in a compact region of space–time. This is shown to be related to non-Hausdorff behaviour of the b-boundary. We show that in a generic space–time, an observer travelling on one of these incomplete curves would experience infinite curvature forces. We also show that the kind of behaviour which occurs in Taub–NUT space cannot happen if there is some matter present.

8·1 The definition of singularities

By analogy with electrodynamics one might think it reasonable to define a space–time singularity as a point where the metric tensor was undefined or was not suitably differentiable. However the trouble with this is that one could simply cut out such points and say that the remaining manifold represented the whole of space–time, which would then be non-singular according to this definition. Indeed, it would seem

inappropriate to regard such singular points as being part of space–time, for the normal equations of physics would not hold at them and it would be impossible to make any measurements. We therefore defined space–time in §3.1 as a pair $(\mathscr{M}, \mathbf{g})$ where the metric $\mathbf{g}$ is Lorentzian and suitably differentiable and we ensured that no regular points were omitted from the manifold $\mathscr{M}$ along with the singular points by requiring that $(\mathscr{M}, \mathbf{g})$ could not be extended with the required differentiability.

The problem of defining whether space–time has a singularity now becomes one of determining whether any singular points have been cut out. One would hope to recognize this by the fact that space–time was incomplete in some sense.

In the case of a manifold $\mathscr{M}$ with a positive definite metric $\mathbf{g}$, one can define a distance function $\rho(x, y)$ which is the greatest lower bound of the length of curves from x to y. The distance function $\rho(x, y)$ is a metric in the topological sense; that is, a basis for the open sets of $\mathscr{M}$ is provided by the sets $\mathscr{B}(x, r)$ consisting of all points $y \in \mathscr{M}$ such that $\rho(x, y) < r$. The pair $(\mathscr{M}, \mathbf{g})$ is said to be *metrically complete* (*m-complete*) if every Cauchy sequence with respect to the distance function ρ converges to a point in $\mathscr{M}$. (A *Cauchy sequence* is an infinite sequence of points x_n such that for any $\epsilon > 0$ there is a number N such that $\rho(x_n, x_m) < \epsilon$ whenever n and m are greater than N.) An alternative formulation is that $(\mathscr{M}, \mathbf{g})$ is m-complete if every C^1 curve of finite length has an endpoint in the sense of §6.2 (note that the curve need not be C^1 at the endpoint). It therefore follows that m-completeness implies *geodesic completeness* (*g-completeness*), that is every geodesic can be extended to arbitrary values of its affine parameter. In fact it can be shown (see Kobayashi and Nomizu (1963)) that g-completeness and m-completeness are equivalent for a positive definite metric.

A Lorentz metric, on the other hand, does not define a topological metric and so one is left only with g-completeness. One can distinguish three kinds of g-incompleteness: that of timelike, null and spacelike geodesics. If one cuts a regular point out of space–time, the resulting manifold is incomplete in all three ways and so one might hope that a space–time which was complete in one of the above senses would also be complete in the other two. Unfortunately this is not necessarily so (Kundt (1963)), as is shown by the following example given by Geroch (1968*b*). Consider two-dimensional Minkowski space with coordinates x and t and metric g_{ab}. Define a new metric $\hat{g}_{ab} = \Omega^2 g_{ab}$ where the positive function Ω has the properties:

(1) $\Omega = 1$ outside the region between the vertical lines $x = -1$ and $x = +1$;

(2) Ω is symmetric about the t-axis, that is, $\Omega(t, x) = \Omega(t, -x)$;

(3) on the t-axis, $t^2\Omega \to 0$ as $t \to \infty$.

By (2) the t-axis is a timelike geodesic which by (3) is incomplete as $t \to \infty$. However every null and spacelike geodesic must leave and not re-enter the region between $x = -1$ and $x = +1$. Therefore by (1) the space is null and spacelike complete. In fact one can construct examples which are incomplete in any of the three possible ways and complete in the remaining two.

Timelike geodesic incompleteness has an immediate physical significance in that it presents the possibility that there could be freely moving observers or particles whose histories did not exist after (or before) a finite interval of proper time. This would appear to be an even more objectionable feature than infinite curvature and so it seems appropriate to regard such a space as singular. Although the affine parameter on a null geodesic does not have quite the same physical significance as proper time does on timelike geodesics, one should probably also regard a null geodesically incomplete space–time as singular both because null geodesics are the histories of zero rest-mass particles and because there are some examples (such as the Reissner–Nordström solution, §5.5) which one would think of as singular but which are timelike but not null geodesically complete. As nothing moves on spacelike curves, the significance of spacelike geodesic incompleteness is not so clear. We shall therefore adopt the view that *timelike and null geodesic completeness are minimum conditions for space–time to be considered singularity-free.* Therefore if a space–time is timelike or null geodesically incomplete, we shall say that it has a singularity.

The advantage of taking timelike and/or null incompleteness as being indicative of the presence of a singularity is that on this basis one can establish a number of theorems about their occurrence. However, the class of timelike and/or null incomplete space–times does not include all those one might wish to consider as singular in some sense. For example Geroch (1968*b*) has constructed a space–time which is geodesically complete but which contains an inextendible timelike curve of bounded acceleration and finite length. An observer with a suitable rocketship and a finite amount of fuel could traverse this curve. After a finite interval of time he would no longer be represented by a point of the space–time manifold. If one is going to say that there

is a singularity in a space–time in which a freely falling observer comes to an untimely end, one should presumably do the same for an observer in a rocketship. What one needs is some generalization of the concept of an affine parameter to all C^1 curves, geodesic or non-geodesic. One could then define a notion of completeness by requiring that every C^1 curve of finite length as measured by such a parameter had an endpoint. The idea we are going to use seems to have been first suggested by Ehresman (1957), and has been reformulated in an elegant manner by Schmidt (1971).

Let $\lambda(t)$ be a C^1 curve through $p \in \mathscr{M}$ and let $\{\mathbf{E}_i\}$ $(i = 1, 2, 3, 4)$ be a basis for T_p. One can parallelly propagate $\{\mathbf{E}_i\}$ along $\lambda(t)$ to obtain a basis for $T_{\lambda(t)}$ for each value of t. Then the tangent vector $\mathbf{V} = (\partial/\partial t)_{\lambda(t)}$ can be expressed in terms of the basis as $\mathbf{V} = V^i(t)\,\mathbf{E}_i$, and one can define a *generalized affine parameter* u on λ by

$$u = \int_p (\sum_i V^i V^i)^{\frac{1}{2}}\,dt.$$

The parameter u depends on the point p and the basis $\{\mathbf{E}_i\}$ at p. If $\{\mathbf{E}_{i'}\}$ is another basis at p, then there is some non-singular matrix $A_i{}^{j'}$ such that

$$\mathbf{E}_i = \sum_{j'} A_i{}^{j'}\mathbf{E}_{j'}.$$

As $\{\mathbf{E}_{i'}\}$ and $\{\mathbf{E}_i\}$ are parallelly transported along $\lambda(t)$, this relation is maintained with constant $A_i{}^{j'}$. Thus

$$V^{i'}(t) = \sum_j A_j{}^{i'} V^j(t).$$

Since $A_i{}^{j'}$ is a non-singular matrix, there is some constant $C > 0$ such that

$$C \sum_i V^i V^i \leqslant \sum_{i'} V^{i'} V^{i'} \leqslant C^{-1} \sum_i V^i V^i.$$

Thus the length of a curve λ is finite in the parameter u if and only if it is finite in the parameter u'. If λ is a geodesic curve then u is an affine parameter on λ, but the beauty of the definition is that u can be defined on any C^1 curve. We shall say that $(\mathscr{M}, \mathbf{g})$ is *b-complete* (short for bundle complete, see §8.3) if there is an endpoint for every C^1 curve of finite length as measured by a generalized affine parameter. If the length is finite in one such parameter it will be finite in all such parameters, so one loses nothing by restricting the bases to be orthonormal bases. If the metric $\mathbf{g}$ is positive definite, the generalized affine parameter defined by an orthonormal basis is arc-length and so b-completeness coincides with m-completeness. However b-completeness can be defined even if the metric is not positive definite; in fact it

can be defined providing there is a connection on $\mathscr{M}$. Clearly b-completeness implies g-completeness, but the example quoted shows that the converse is not true.

We shall therefore define a space–time to be *singularity-free* if it is b-complete. This definition conforms with the requirement made above, that timelike and null geodesic completeness are minimum conditions for a space–time to be considered singularity-free. One might possibly wish to weaken this condition slightly, to say that space–time is singularity-free it it is only *non-spacelike b-complete*, i.e. if there is an endpoint for all non-spacelike C^1 curves with finite length as measured by a generalized affine parameter. However this definition would appear rather awkward in the bundle formulation of b-completeness which we shall give in §8.3. In fact each of the theorems we give in §8.2 implies that $(\mathscr{M}, \mathbf{g})$ is timelike or null g-incomplete and hence has a singularity by both the above definitions.

One feels intuitively that a singularity ought to involve the curvature becoming unboundedly large near a singular point. However since we have excluded singular points from our definition of space–time, difficulty arises in defining both 'near' and 'unboundedly large'. One can say that points on a b-incomplete curve are near the singularity if they correspond to values of a generalized affine parameter which is near the upper bound of that parameter. 'Unboundedly large' is more difficult, since the size of components of the curvature tensor depend on the basis in which it is measured. One possibility is to look at scalar polynomials in g_{ab}, η_{abcd}, and R_{abcd}. We shall say that a b-incomplete curve corresponds to a scalar polynomial curvature singularity (*s.p. curvature singularity*) if any of these scalar polynomials is unbounded on the incomplete curve. However, with a Lorentz metric these polynomials do not fully characterize the Riemann tensor since, as Penrose has pointed out, in plane-wave solutions the scalar polynomials are all zero but the Riemann tensor does not vanish. (This is similar to the fact that a non-zero vector may have zero length.) Thus the curvature might become very large in some sense even though the scalar polynomials remained small. Alternatively one might measure the components of the curvature tensor in a basis that was parallelly propagated along a curve. We shall say that a b-incomplete curve corresponds to a curvature singularity with respect to a parallelly propagated basis (a *p.p. curvature singularity*) if any of these components is unbounded on the curve. Clearly an s.p. curvature singularity implies a p.p. curvature singularity.

One might expect that in any physically realistic solution, a b-incomplete curve would correspond both to an s.p. and a p.p. curvature singularity. However an example of a solution where this does not seem to be true is provided by Taub–NUT space (§5.8). Here the incomplete geodesics are totally imprisoned in a compact neighbourhood of the horizon. As the metric is perfectly regular on this compact neighbourhood, the scalar polynomials in the curvature remain finite. Because of the special nature of this solution, the components of the curvature in a parallelly propagated basis along the imprisoned geodesics remains bounded. Since the imprisoned geodesics are contained in a compact set, one could not extend the manifold $\mathscr{M}$ to a larger four-dimensional Hausdorff paracompact manifold $\mathscr{M}'$, in which the incomplete geodesics could be continued. Thus there is no possibility of the incompleteness having arisen from the cutting out of singular points. Nevertheless it would be unpleasant to be moving on one of the incomplete timelike geodesics for although one's world-line never comes to an end and would continue to wind round and round inside the compact set, one would never get beyond a certain time in one's life. It would, therefore, seem reasonable to say that such a space–time was singular even though there is no p.p. or s.p. curvature singularity. By lemma 6.4.8, such totally imprisoned incompleteness can only occur if strong causality is violated. In §8.5 we shall show that in a generic space–time, a partially or totally imprisoned b-incomplete curve will correspond to a p.p. curvature singularity. We shall also show that the Taub–NUT kind of totally imprisoned incompleteness cannot occur if there is some matter present.

8.2 Singularity theorems

In §5.4 it was shown that there would be singularities in spatially homogeneous solutions under certain reasonable conditions. Similar theorems can be obtained for a number of other types of exact symmetry. Such results, although suggestive, do not necessarily have any physical significance because they depend on the symmetry being exact and clearly in any physical situation this will not be the case. It was therefore suggested by a number of authors that singularities were simply the result of symmetries and that they would not occur in general solutions. This view was supported by Lifshitz, Khalatnikov and co-workers who showed that certain classes of solutions with space-

like singularities did not have the full number of arbitrary functions expected in a general solution of the field equations (see Lifshitz and Khalatnikov (1963) for an account of this work). This presumably indicates that the Cauchy data which gave rise to such singularities is of measure zero in the set of all possible Cauchy data and so should not occur in the real universe. However more recently Belinskii, Khalatnikov and Lifshitz (1970) have found other classes of solutions which seem to have the full number of arbitrary functions and to contain singularities. They have therefore withdrawn the claim that singularities do not occur in general solutions. Their methods are interesting for the light they shed on the possible structure of singularities but it is not clear whether the power series which are used will converge. Neither does one obtain general conditions which imply that a singularity is inevitable. Nevertheless we may take their results as supporting our view that the singularities implied by the theorems of this section involve infinite curvature in general.

The first theorem about singularities which did not involve any assumption of symmetry was given by Penrose (1965*c*). It was designed to prove the occurrence of a singularity in a star which collapsed inside its Schwarzschild radius. If the collapse were exactly spherical, the solution could be integrated explicitly and a singularity would always occur. However it is not obvious that this would be the case if there were irregularities or a small amount of angular momentum. Indeed in Newtonian theory the smallest amount of angular momentum could prevent the occurrence of infinite density and cause the star to re-expand. However Penrose showed that the situation was very different in General Relativity: once the star had passed inside the Schwarzschild surface (the surface $r = 2m$) it could not come out again. In fact the Schwarzschild surface is defined only for an exactly spherically symmetric solution but the more general criterion used by Penrose is equivalent for such a solution and is applicable also to solutions without exact symmetry. It is that there should exist a *closed trapped surface* $\mathscr{T}$. By this is meant a C^2 closed (i.e. compact, without boundary) spacelike two-surface (normally, S^2) such that the two families of null geodesics orthogonal to $\mathscr{T}$ are converging at $\mathscr{T}$ (i.e. ${}_1\hat{\chi}_{ab}g^{ab}$ and ${}_2\hat{\chi}_{ab}g^{ab}$ are negative, where ${}_1\hat{\chi}_{ab}$ and ${}_2\hat{\chi}_{ab}$ are the two null second fundamental forms of $\mathscr{T}$. In the following chapters we shall discuss the circumstances under which such a surface would arise.) One may think of $\mathscr{T}$ as being in such a strong gravitational field that even the 'outgoing' light rays are dragged back and

are, in fact, converging. Since nothing can travel faster than light, the matter within $\mathscr{T}$ is trapped inside a succession of two-surfaces of smaller and smaller area and so it seems that something must go wrong. That this is so is shown rigorously by Penrose's theorem:

Theorem 1

Space–time $(\mathscr{M}, \mathbf{g})$ cannot be null geodesically complete if:

(1) $R_{ab}K^aK^b \geqslant 0$ for all null vectors K^a (cf. §4.3);
(2) there is a non-compact Cauchy surface $\mathscr{H}$ in $\mathscr{M}$;
(3) there is a closed trapped surface $\mathscr{T}$ in $\mathscr{M}$.

Note: the method of proof is to show that the boundary of the future of $\mathscr{T}$ would be compact if $\mathscr{M}$ were null geodesically complete. This is then shown to be incompatible with $\mathscr{H}$ being non-compact.

Proof. The existence of a Cauchy surface implies that $\mathscr{M}$ is globally hyperbolic (proposition 6.6.3) and therefore causally simple (proposition 6.6.1). This means that the boundary of $J^+(\mathscr{T})$ will be $E^+(\mathscr{T})$ and will be generated by null geodesic segments which have past endpoints on $\mathscr{T}$ and which are orthogonal to $\mathscr{T}$. Suppose $\mathscr{M}$ were null geodesically complete. Then by conditions (1) and (3) and proposition 4.4.6 there would be a point conjugate to $\mathscr{T}$ along every future-directed null geodesic orthogonal to $\mathscr{T}$ within an affine distance $2c^{-1}$ where c is the value of ${}_n\hat{\chi}_{ab}g^{ab}$ at the point where the null geodesic intersects $\mathscr{T}$. By proposition 4.5.14, points on such a null geodesic beyond the point conjugate to $\mathscr{T}$ would lie in $I^+(\mathscr{T})$. Thus each generating segment of $\dot{J}^+(\mathscr{T})$ would have a future endpoint at or before the point conjugate to $\mathscr{T}$. At $\mathscr{T}$ one could assign, in a continuous manner, an affine parameter on each null geodesic orthogonal to $\mathscr{T}$. Consider the continuous map $\beta: \mathscr{T} \times [0, b] \times Q \to \mathscr{M}$ (Q is the discrete set 1, 2) defined by taking a point $p \in \mathscr{T}$ an affine distance $v \in [0, b]$ along one or other of the two future-directed null geodesics through p orthogonal to $\mathscr{T}$. Since $\mathscr{T}$ is compact, there will be some minimum value c_0 of $(-{}_1\hat{\chi}_{ab}g^{ab})$ and $(-{}_2\hat{\chi}_{ab}g^{ab})$. Then if $b_0 = 2c_0^{-1}$, $\beta(\mathscr{T} \times [0, b_0] \times Q)$ would contain $\dot{J}^+(\mathscr{T})$. Thus $\dot{J}^+(\mathscr{T})$ would be compact being a closed subset of a compact set. This would be possible if the Cauchy surface $\mathscr{H}$ were compact because then $\dot{J}^+(\mathscr{T})$ could meet up round the back and form a compact Cauchy surface homeomorphic to $\mathscr{H}$ (figure 49). However there is clearly going to be trouble if one demands that $\mathscr{H}$ is non-compact. To show this rigorously one can use the fact (see §2.6) that $\mathscr{M}$ admits a past-directed C^1 timelike

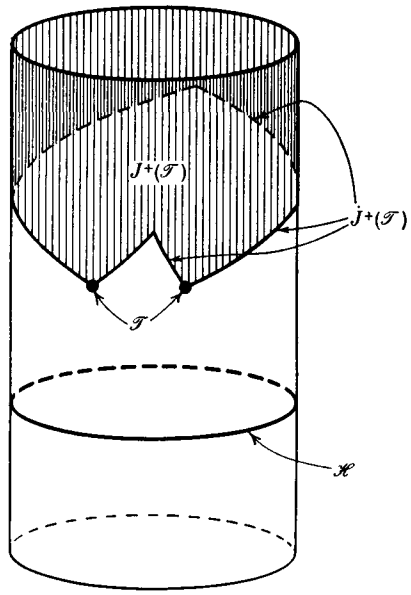

FIGURE 49. A two-dimensional section of a geodesically complete space with a compact Cauchy surface $\mathscr{H}$. The two-sphere $\mathscr{T}$ has a compact boundary $\dot{J}^+(\mathscr{T})$ to its future $J^+(\mathscr{T})$, as the outgoing null geodesics from $\mathscr{T}$ meet up round the back of the cylinder.

vector field. Each integral curve of this field will intersect $\mathscr{H}$ (as it is a Cauchy surface) and will intersect $\dot{J}^+(\mathscr{T})$ at most once. Thus they will define a continuous one-to-one map $\alpha: \dot{J}^+(\mathscr{T}) \to \mathscr{H}$. If $\dot{J}^+(\mathscr{T})$ were compact, its image $\alpha(\dot{J}^+(\mathscr{T}))$ would also be compact and would be homeomorphic to $\dot{J}^+(\mathscr{T})$. However as $\mathscr{H}$ is non-compact, $\alpha(\dot{J}^+(\mathscr{T}))$ could not contain the whole of $\mathscr{H}$ and would therefore have to have a boundary in $\mathscr{H}$. This would be impossible since by proposition 6.3.1, $\dot{J}^+(\mathscr{T})$, and therefore $\alpha(\dot{J}^+(\mathscr{T}))$, would be a three-dimensional manifold (without boundary). This shows that the assumption that $\mathscr{M}$ is null geodesically complete (which we made in order to prove $\dot{J}^+(\mathscr{T})$ compact) is incorrect. □

Condition (1) of this theorem (that $R_{ab}K^aK^b \geqslant 0$ for any null vector $\mathbf{K}$) was discussed in §4.3. It will hold no matter what value the value of the constant Λ, provided that the energy density is positive for every observer. It will be shown in chapter 9 that condition (3) (that there is a closed trapped surface) should be satisfied in at least some region of space–time. This leaves condition (2) (that there is a non-compact

spacelike surface $\mathscr{H}$ which is a Cauchy surface) to be discussed. By proposition 6.4.9, the existence of spacelike surfaces is guaranteed provided one assumes stable causality. That the spacelike surface $\mathscr{H}$ be non-compact is not too serious a restriction since the only place it was used was to show that $\alpha(\dot{J}^+(\mathscr{T}))$ could not be the whole of $\mathscr{H}$. This could also be shown if, instead of taking $\mathscr{H}$ to be non-compact, one required that there exist a future-directed inextendible curve from $\mathscr{H}$ which did not intersect $\dot{J}^+(\mathscr{T})$. In other words, the theorem would still hold even if $\mathscr{H}$ were compact, provided there was some observer who could avoid falling into the collapsing star. This might not be possible if the whole universe were collapsing also, but in such a case one would expect singularities anyway as will be shown presently. The real weakness of the theorem is the requirement that $\mathscr{H}$ be a Cauchy surface. This was used in two places: first, to show that $\mathscr{M}$ was causally simple which implied that the generators of $\dot{J}^+(\mathscr{T})$ had past endpoints on $\mathscr{T}$, and second, to ensure that under the map α every point of $\dot{J}^+(\mathscr{T})$ was mapped into a point of $\mathscr{H}$. That the Cauchy surface condition is necessary is shown by an example due to Bardeen. This has the same global structure as the Reissner–Nordström solution except that the real singularities at $r = 0$ have been smoothed out so that they are just the origins of polar coordinates. The space–time obeys the condition $R_{ab}K^aK^b \geqslant 0$ for any null but not timelike vector $\mathbf{K}$, and contains closed trapped surfaces. The only way in which it fails to satisfy the conditions of the theorem is that it does not have a Cauchy surface.

It therefore seems that what the theorem tells us is that in a collapsing star there will occur either a singularity or a Cauchy horizon. This is a very important result since in either case our ability to predict the future breaks down. However it does not answer the question of whether singularities occur in physically realistic solutions. To decide this we need a theorem which does not assume the existence of Cauchy surfaces. One of the conditions of such a theorem must be that $R_{ab}K^aK^b \geqslant 0$ for all *timelike* as well as null vectors, since failure to obey this condition is the only way in which Bardeen's example is unreasonable. The theorem we shall give below requires this condition and also the chronology condition that there be no closed timelike curves. On the other hand it is applicable to a wider class of situations since the existence of a closed trapped surface is now only one of three possible conditions. One of these alternative conditions is that there should be a compact partial Cauchy surface, and the other is that there

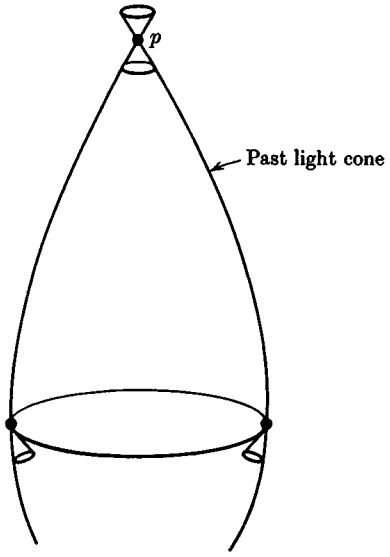

FIGURE 50. A point p whose past light cone starts reconverging.

should be a point whose past (or future) light cone starts converging again (figure 50). The first of these other conditions is satisfied in a spatially closed solution while the second is closely related to the existence of a closed trapped surface but is in a form which is more convenient for some purposes; for in the case in which the light cone is our own past light cone, one can directly determine whether this condition is satisfied. In the last chapter it will be shown that recent observations of the microwave background indicate that it is.

The precise statement is:

Theorem 2 (*Hawking and Penrose* (1970))

Space–time $(\mathscr{M}, \mathbf{g})$ is not timelike and null geodesically complete if:

(1) $R_{ab}K^aK^b \geqslant 0$ for every non-spacelike vector $\mathbf{K}$ (cf. §4.3).

(2) The generic condition is satisfied (§4.4), i.e. every non-spacelike geodesic contains a point at which $K_{[a}R_{b]cd[e}K_{f]}K^cK^d \neq 0$, where $\mathbf{K}$ is the tangent vector to the geodesic.

(3) The chronology condition holds on $\mathscr{M}$ (i.e. there are no closed timelike curves).

(4) There exists at least one of the following:

(i) a compact achronal set without edge,

(ii) a closed trapped surface,

(iii) a point p such that on every past (or every future) null geodesic from p the divergence θ of the null geodesics from p becomes negative (i.e. the null geodesics from p are focussed by the matter or curvature and start to reconverge).

Remark. An alternative version of the theorem is that the following three conditions cannot all hold:

(a) every inextendible non-spacelike geodesic contains a pair of conjugate points;

(b) the chronology condition holds on $\mathscr{M}$;

(c) there is an achronal set $\mathscr{S}$ such that $E^+(\mathscr{S})$ or $E^-(\mathscr{S})$ is compact. (We shall say that such a set is, respectively, *future trapped* or *past trapped*).

In fact it is this form of the theorem that we shall prove. The other version will then follow since if $\mathscr{M}$ were timelike and null geodesically complete, (1) and (2) would imply (a) by propositions 4.4.2 and 4.4.5, (3) is the same as (b), and (1) and (4) would imply (c), since in case (i) $\mathscr{S}$ would be the compact achronal set without edge and

$$E^+(\mathscr{S}) = E^-(\mathscr{S}) = \mathscr{S};$$

in cases (ii) and (iii) $\mathscr{S}$ would be the closed trapped surface and the point p respectively, and by propositions 4.4.4, 4.4.6, 4.5.12 and 4.5.14 $E^+(\mathscr{S})$ and $E^-(\mathscr{S})$ would be compact respectively, being the intersections of the closed sets $\dot{J}^+(\mathscr{S})$ and $\dot{J}^-(\mathscr{S})$ with compact sets consisting of all the null geodesics of some finite length from $\mathscr{S}$.

Proof. As the proof is rather long, we shall break it up by first establishing a lemma and corollary. We note that by an argument similar to that of proposition 6.4.6, (a) and (b) imply that strong causality holds on $\mathscr{M}$.

Lemma 8.2.1

If $\mathscr{S}$ is a closed set and if the strong causality condition holds on $\bar{J}^+(\mathscr{S})$ then $H^+(\bar{E}^+(\mathscr{S}))$ is non-compact or empty (figure 51).

By lemma 6.3.2, through every point $q \in \dot{J}^+(\mathscr{S}) - \mathscr{S}$ there is a past-directed null geodesic segment lying in $\dot{J}^+(\mathscr{S})$ which has a past endpoint if and only if $q \in E^+(\mathscr{S})$. (Note that as we no longer assume the existence of a Cauchy surface, $\mathscr{M}$ may not be causally simple and so $\dot{J}^+(\mathscr{S}) - E^+(\mathscr{S})$ may be non-empty.) Therefore if $q \in \dot{J}^+(\mathscr{S}) - E^+(\mathscr{S})$, there is a past-inextendible null geodesic through q which lies in $\dot{J}^+(\mathscr{S})$

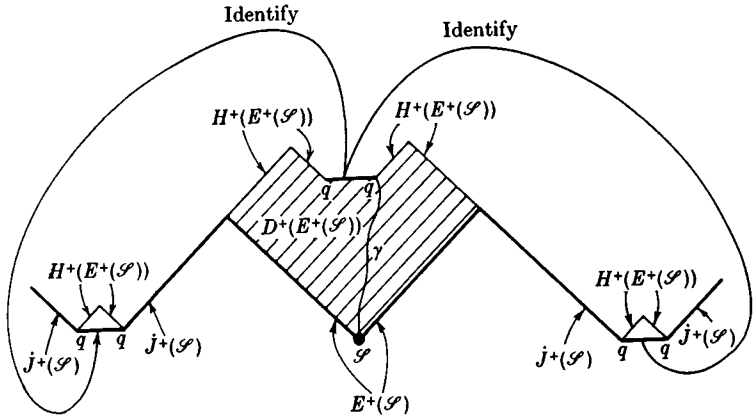

FIGURE 51. A future trapped set $\mathscr{S}$; null lines are at $\pm 45°$, three lines have been identified and the points q are at infinity. The achronal sets $E^+(\mathscr{S})$, $\dot{J}^+(\mathscr{S})$ and $H^+(E^+(\mathscr{S}))$ are shown. A future-inextendible timelike curve $\gamma \in D^+(E^+(\mathscr{S}))$ is shown.

and so does not intersect $I^-(\dot{J}^+(\mathscr{S}))$. From lemma 6.6.4 it then follows that q is not in $D^+(\dot{J}^+(\mathscr{S})) - H^+(\dot{J}^+(\mathscr{S}))$. Hence

$$D^+(\bar{E}^+(\mathscr{S})) - H^+(\bar{E}^+(\mathscr{S})) = D^+(\dot{J}^+(\mathscr{S})) - H^+(\dot{J}^+(\mathscr{S}))$$

and

$$H^+(\bar{E}^+(\mathscr{S})) \subset H^+(\dot{J}^+(\mathscr{S})).$$

Now suppose that $H^+(\bar{E}^+(\mathscr{S}))$ was non-empty and compact. Then it could be covered by a finite number of local causality neighbourhoods $\mathscr{U}_i$. Let p_1 be a point of $J^+(\mathscr{S}) \cap [\mathscr{U}_1 - D^+(\dot{J}^+(\mathscr{S}))]$. Then from p_1 there would be a past-inextendible non-spacelike curve λ_1 which did not intersect either $\dot{J}^+(\mathscr{S})$ or $D^+(\bar{E}^+(\mathscr{S}))$. Since the $\mathscr{U}_i$ have compact closure, λ_1 would leave $\mathscr{U}_1$. Let q_1 be a point on λ_1 not in $\mathscr{U}_1$. Then since $q_1 \in J^+(\mathscr{S})$ there would be a non-spacelike curve μ_1 from q_1 to $\mathscr{S}$. This curve would intersect $D^+(\bar{E}^+(\mathscr{S}))$ and hence would intersect some $\mathscr{U}_i$ other than $\mathscr{U}_1$ (say, $\mathscr{U}_2$). Then let p_2 be a point of $\mu_1 \cap [\mathscr{U}_2 - D^+(\dot{J}^+(\mathscr{S}))]$ and continue as before.

This leads to a contradiction since there were only a finite number of the local causality neighbourhoods $\mathscr{U}_i$, and one could not return to an earlier $\mathscr{U}_j$ because no non-spacelike curve can intersect a $\mathscr{U}_i$ more than once. Thus $H^+(\bar{E}^+(\mathscr{S}))$ must be non-compact or empty. □

Corollary

If $\mathscr{S}$ is a future trapped set, there is a future-inextendible timelike curve γ contained in $D^+(E^+(\mathscr{S}))$.

Put a timelike vector field on $\mathscr{M}$. If every integral curve of this field which intersected $E^+(\mathscr{S})$ also intersected $H^+(E^+(\mathscr{S}))$ they would define a continuous one–one mapping of $E^+(\mathscr{S})$ onto $H^+(E^+(\mathscr{S}))$ and hence $H^+(E^+(\mathscr{S}))$ would be compact. The intersection of $I^+(\mathscr{S})$ with a curve which does not intersect $H^+(E^+(\mathscr{S}))$ gives the desired curve γ (figure 51 indicates one possible situation). □

Now consider the compact set $\mathscr{F}$ defined as $E^+(\mathscr{S}) \cap \overline{J^-(\gamma)}$. Since γ was contained in $\operatorname{int} I^+(E^+(\mathscr{S}))$, $E^-(\mathscr{F})$ would consist of $\mathscr{F}$ and a portion of $\dot{J}^-(\gamma)$. Since γ was future inextendible, the null geodesic segments generating $\dot{J}^-(\gamma)$ could have no future endpoints. But by (a) every inextendible non-spacelike geodesic contains a pair of conjugate points. Thus by proposition 4.5.12, the past-inextendible extension ν' of each generating segment ν of $\dot{J}^-(\gamma)$ would enter $I^-(\gamma)$. There would be a past endpoint for ν at or before the first point p of $\overline{\nu' \cap I^-(\gamma)}$. As $I^-(\gamma)$ would be an open set, a neighbourhood of p would contain points in $I^-(\gamma)$ on neighbouring null geodesics. Thus the affine distance of the points p from $\mathscr{F}$ would be upper semi-continuous, and $E^-(\mathscr{F})$ would be compact being the intersection of the closed set $\dot{J}^-(\gamma)$ with a compact set generated by null geodesic segments from $\mathscr{F}$ of some bounded affine length. It would then follow from the lemma that there would be a past-inextendible timelike curve λ contained in $\operatorname{int} D^-(E^-(\mathscr{F}))$ (figure 52). Let a_n be an infinite sequence of points on λ such that:

(I) $a_{n+1} \in I^-(a_n)$,

(II) no compact segment of λ contains more than a finite number of the a_n.

Let b_n be a similar sequence on γ but with I^+ instead of I^- in (I) and with $b_1 \in I^+(a_1)$.

As γ and λ were contained in the globally hyperbolic set $\operatorname{int} D(E^-(\mathscr{F}))$ (proposition 6.6.3), there would be a non-spacelike geodesic μ_n of maximum length between each a_n and the corresponding b_n (proposition 6.7.1). Each would intersect the compact set $E^+(\mathscr{S})$. Thus there would be a $q \in E^+(\mathscr{S})$ which was a limit point of the $\mu_n \cap E^+(\mathscr{S})$ and a non-spacelike direction at q which is a limit of the directions of the μ_n. (The point q and the direction at q define a point of the bundle of directions over $\mathscr{M}$. Such a limit point exists because the portion of the bundle over $E^+(\mathscr{S})$ is compact.) Let μ'_n be a subsequence of the μ_n such that $\mu'_n \cap E^+(\mathscr{S})$ converges to q and such that the directions of the μ'_n at $E^+(\mathscr{S})$ converge to the limit direction.

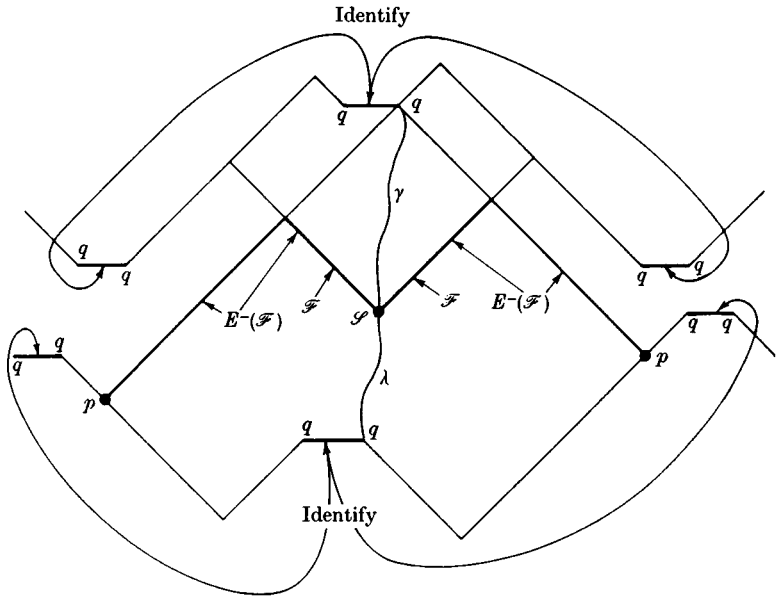

FIGURE 52. As figure 51, but with three further lines identified. $\mathscr{F}$ is the set $E^+(\mathscr{S}) \cap \bar{J}^-(\gamma)$; the points p are past endpoints of null geodesic generating segments of $E^-(\mathscr{F})$. The curve λ is a past-inextendible timelike curve contained in int $D^-(E^-(\mathscr{F}))$.

(More precisely, the points defined by the μ'_n in the bundle of directions over $E^+(\mathscr{S})$ converge to the limit point.) Let μ be the inextendible geodesic through q in the limit direction. By (a) there would be conjugate points x and y on μ with $y \in I^+(x)$. Let x' and y' be on μ to the past and future of x and y respectively. By proposition 4.5.8, there is some $\epsilon > 0$ and some timelike curve α from x' to y' whose length is ϵ plus the length of μ from x' to y'. Let $\mathscr{U}$ and $\mathscr{V}$ be convex normal coordinate neighbourhoods of x' and y' respectively, each of which contains no curve of length $\frac{1}{4}\epsilon$. Let x'' and y'' be $\dot{\mathscr{U}} \cap \alpha$ and $\dot{\mathscr{V}} \cap \alpha$ respectively. Let x'_n and y'_n be points on μ'_n converging to x' and y' respectively. For n sufficiently large, the length μ'_n from x'_n to y'_n will be less than $\frac{1}{4}\epsilon$ plus the length of μ from x' to y'. Also for n sufficiently large, x'_n and y'_n would be in $I^-(x'', \mathscr{U})$ and $I^+(y'', \mathscr{V})$ respectively. Then going from x'_n to x'', along α to y'', and from y'' to y'_n would give a longer non-spacelike curve than μ'_n from x'_n to y'_n. But by property (II), a'_n would lie to the past of x'_n on μ'_n and b'_n would lie to the future of y'_n on μ'_n, for n large enough. Therefore μ'_n ought to be the longest non-spacelike curve from x'_n to y'_n. This establishes the desired contradiction. □

While this theorem establishes the existence of singularities under very general conditions, it has the disadvantage of not showing whether the singularity is in the future or the past. In case (ii) of condition (4), when there is a compact spacelike surface, one has no reason to believe that it should be in the future rather than in the past, but in case (i) when there is a closed trapped surface, one would expect the singularity to be in the future, and in case (iii) when the past null cone starts reconverging, one would expect the singularity to be in the past. One can show that there is a singularity in the past if condition (iii) is strengthened somewhat to say that all past-directed timelike as well as null geodesics from p start to reconverge within a compact region in $J^-(p)$.

Theorem 3 (*Hawking* (1967))

If (1) $R_{ab}K^aK^b \geqslant 0$ for every non-spacelike vector $\mathbf{K}$ (cf. §4.3);

(2) the strong causality condition holds on $(\mathscr{M}, \mathfrak{g})$;

(3) there is some past-directed unit timelike vector $\mathbf{W}$ at a point p and a positive constant b such that if $\mathbf{V}$ is the unit tangent vector to the past-directed timelike geodesics through p, then on each such geodesic the expansion $\theta \equiv V^a{}_{;a}$ of these geodesics becomes less than $-3c/b$ within a distance b/c from p, where $c \equiv -W^aV_a$,

then there is a past incomplete non-spacelike geodesic through p.

Let K^a be the parallelly propagated tangent vector to the past-directed non-spacelike geodesics through p, normalized by $K^aW_a = -1$. Then for the timelike geodesics through p, $K^a = c^{-1}V^a$ and so $K^a{}_{;a} = c^{-1}V^a{}_{;a}$. Since $K^a{}_{;a}$ is continuous on the non-spacelike geodesics, it will become less than $-3/b$ on the null geodesics through p within an affine distance b. If $\mathbf{Y}_1$, $\mathbf{Y}_2$, $\mathbf{Y}_3$ and $\mathbf{Y}_4$ are a pseudo-orthonormal tetrad on these null geodesics with $\mathbf{Y}_1$ and $\mathbf{Y}_2$ spacelike unit vectors and $\mathbf{Y}_3$ and $\mathbf{Y}_4$ null with $Y_3{}^aY_{4a} = -1$ and $\mathbf{Y}_4 = \mathbf{K}$, the expansion $\hat{\theta}$ of the null geodesics through p is defined as

$$\begin{aligned}\hat{\theta} &= K_{a;b}(Y_1{}^aY_1{}^b + Y_2{}^aY_2{}^b)\\ &= K^a{}_{;a} + K_{a;b}(Y_3{}^aY_4{}^b + Y_4{}^aY_3{}^b).\end{aligned}$$

The second term is zero because K^a is parallelly propagated. The third term can be expressed as $\frac{1}{2}(K_aK^a)_{;b}Y_3{}^b$, which is less than zero as K_aK^a is zero on the null geodesics and negative for timelike geodesics. This shows that $\hat{\theta}$ will become less than $-3/b$ within an affine distance b along each null geodesic from p. Thus if all past-directed null geodesics

from p were complete, $E^-(p)$ would be compact. Any point $q \in J^-(E^-(p)) - E^-(p)$ would be in $I^-(p)$. Thus it could not be in $J^+(E^-(p))$ since $E^-(p)$ is achronal. Therefore

$$J^+(E^-(p)) \cap J^-(E^-(p)) = E^-(p)$$

and so would be compact. Then by proposition 6.6.7, $D^-(E^-(p))$ would be globally hyperbolic. By proposition 6.7.1, each point $r \in D^-(E^-(p))$ would be joined to p by a non-spacelike geodesic which did not contain any point conjugate to p between r and p. Thus by proposition 4.4.1, $D^-(E^-(p))$ would be contained in $\exp_p(F)$ where F is the compact region of T_p consisting of all past-directed non-spacelike vectors K^a such that $K^a W_a \leqslant -2b$. If all past non-spacelike geodesics from p were complete, $\exp_p(K^a)$ would be defined for every $K^a \in F$, and so $\exp_p(F)$ would be compact being the image of a compact set under a continuous map. However by the corollary to lemma 8.2.1, $D^-(E^-(p))$ contains a past-inextendible timelike curve. By proposition 6.4.7 this could not be totally imprisoned in the compact set $\exp_p(F)$, therefore the assumption that all past-directed non-spacelike geodesics from p are complete must be false. □

Theorems 2 and 3 are the most useful theorems on singularities since it can be shown that their conditions are satisfied in a number of physical situations (see next chapter). However it might be that what occurred was not a singularity but a closed timelike curve, violating the causality conditions. This would be much worse than the mere breakdown of prediction which was the alternative after theorem 1, and it is our personal opinion that it would be physically more objectionable than a singularity. Nevertheless one would like to know whether such causality violations would prevent the occurrence of singularities. The following theorem shows that they cannot in certain situations. This means that we have to take singularities seriously and it gives us confidence that, in general, causality breakdowns are not the way out.

Theorem 4 (*Hawking* (1967))

Space–time is not timelike geodesically complete if:

(1) $R_{ab} K^a K^b \geqslant 0$ for every non-spacelike vector $\mathbf{K}$ (cf. §4.3);

(2) there exists a compact spacelike three-surface $\mathscr{S}$ (without edge);

(3) the unit normals to $\mathscr{S}$ are everywhere converging (or everywhere diverging) on $\mathscr{S}$.

Remarks. Condition (2) may be interpreted as saying that the universe is spatially closed and condition (3) as saying that it is contracting (or expanding). As explained in § 6.5 one may take a covering manifold $\hat{\mathscr{M}}$ in which each connected component of the image of $\mathscr{S}$ is diffeomorphic to $\mathscr{S}$ and is a partial Cauchy surface in $\hat{\mathscr{M}}$. We shall work in $\hat{\mathscr{M}}$ and shall denote by $\hat{\mathscr{S}}$ one connected component of the image of $\mathscr{S}$. Considering the Cauchy evolution problem in $\hat{\mathscr{M}}$ one sees that the occurrence of singularities (though not necessarily their nature) is a stable property of the Cauchy data on $\hat{\mathscr{S}}$ since a sufficiently small variation of the data on $\hat{\mathscr{S}}$ will not violate condition (3). This is a counterexample to the conjecture by Lifshitz and Khalatnikov that singularities occur only for a set of Cauchy data of measure zero, though it must be remembered that the definition of a singularity adopted here is not that used by Lifshitz and Khalatnikov

Proof. By conditions (2) and (3) the contraction $\chi^a{}_a$ of the second fundamental form of $\hat{\mathscr{S}}$ has a negative upper bound on $\mathscr{S}$. Thus if $\mathscr{M}$ (and hence $\hat{\mathscr{M}}$) was timelike geodesically complete there would be a point conjugate to $\hat{\mathscr{S}}$ on every future-directed geodesic orthogonal to $\hat{\mathscr{S}}$ within a finite upper bound b of distance from $\hat{\mathscr{S}}$ (proposition 4.4.3). But by the corollary to proposition 6.7.1, to every point $q \in D^+(\hat{\mathscr{S}})$ there is a future-directed geodesic orthogonal to $\hat{\mathscr{S}}$ which does not contain any point conjugate to $\hat{\mathscr{S}}$ between $\hat{\mathscr{S}}$ and q. Let $\beta: \hat{\mathscr{S}} \times [0, b] \to \hat{\mathscr{M}}$ be the differentiable map which takes a point $p \in \hat{\mathscr{S}}$ a distance $s \in [0, b]$ up the future-directed geodesic through p orthogonal to $\mathscr{S}$. Then $\beta(\hat{\mathscr{S}} \times [0, b])$ would be compact and would contain $D^+(\hat{\mathscr{S}})$. Thus $\bar{D}^+(\hat{\mathscr{S}})$ and hence $H^+(\hat{\mathscr{S}})$ would be compact. If one assumed the strong causality condition the desired contradiction would follow from lemma 8.2.1. However even without strong causality one can obtain a contradiction. Consider a point $q \in H^+(\hat{\mathscr{S}})$. Since every past-directed non-spacelike curve from q to $\hat{\mathscr{S}}$ would consist of a (possibly zero) null geodesic segment in $H^+(\hat{\mathscr{S}})$ and then a non-spacelike curve in $D^+(\hat{\mathscr{S}})$, it follows that $d(\hat{\mathscr{S}}, q)$ would be less than or equal to b. Thus, as d is lower semi-continuous, one could find an infinite sequence of points $r_n \in D^+(\hat{\mathscr{S}})$ converging to q such that $d(\hat{\mathscr{S}}, r_n)$ converged to $d(\hat{\mathscr{S}}, q)$. To each r_n there would correspond at least one element $\beta^{-1}(r_n)$ of $\hat{\mathscr{S}} \times [0, b]$ Since $\hat{\mathscr{S}} \times [0, b]$ is compact there would be an element (p, s) which was a limit point of the $\beta^{-1}(r_n)$. By continuity $s = d(\hat{\mathscr{S}}, q)$ and $\beta(p, s) = q$. Thus to every point $q \in H^+(\hat{\mathscr{S}})$ there would be a timelike geodesic of length $d(\hat{\mathscr{S}}, q)$ from $\hat{\mathscr{S}}$. Now let

$q_1 \in H^+(\hat{\mathscr{S}})$ lie to the past of q on the same null geodesic generator λ of $H^+(\hat{\mathscr{S}})$. Joining the geodesic of length $d(\hat{\mathscr{S}}, q_1)$ from $\hat{\mathscr{S}}$ to q_1 to the segment of λ between q_1 and q, one would obtain a non-spacelike curve of length $d(\hat{\mathscr{S}}, q_1)$ from $\hat{\mathscr{S}}$ to q which could be varied to give a longer curve between these endpoints (proposition 4.5.10). Thus $d(\hat{\mathscr{S}}, q)$, $q \in H^+(\hat{\mathscr{S}})$, would strictly decrease along every past-directed generator of $H^+(\hat{\mathscr{S}})$. But by proposition 6.5.2, such generators could have no past endpoints. This leads to a contradiction since as $d(\hat{\mathscr{S}}, q)$ is lower semi-continuous in q, it would have a minimum on the compact set $H^+(\hat{\mathscr{S}})$. □

Condition (2) that $\mathscr{S}$ is compact is necessary, since in Minkowski space $(\mathscr{M}, \eta)$ the non-compact surface $\mathscr{S}$: $(x^1)^2 + (x^2)^2 + (x^3)^2 - (x^4)^2 = -1$, $x^4 < 0$, is a partial Cauchy surface with $\chi^a{}_a = -3$ at all points. If one took the region of Minkowski space defined by

$$x^4 < 0, \quad (x^1)^2 + (x^2)^2 + (x^3)^2 - (x^4)^2 < 0,$$

one could identify points under a discrete group of isometries G such that $\mathscr{S}/G$ was compact (Löbell (1931)). As required by theorem 4, the space $(\mathscr{M}/G, \eta)$ would be timelike geodesically incomplete because one could not extend the identification under G to the whole of $\mathscr{M}$ (neither conditions (1) nor (2) of §5.8 would hold at the origin). In this case the incompleteness singularity arises from bad global properties and is not accompanied by a curvature singularity. This example was suggested by Penrose.

Conditions (2) and (3) can be replaced by:

(2′) $\hat{\mathscr{S}}$ is a Cauchy surface for $\hat{\mathscr{M}}$;

(3′) $\chi^a{}_a$ is bounded away from zero on $\mathscr{S}$;

since in this case there cannot be a Cauchy horizon, yet all the future-directed timelike curves from $\hat{\mathscr{S}}$ must have lengths less than some finite upper bound.

Geroch (1966) has shown that if condition (2) holds, and if conditions (1) and (3) are replaced by:

(1″) $R_{ab} K^a K^b \geqslant 0$ for every non-spacelike vector, equality holding only if $R_{ab} = 0$;

(3″) there is a point $p \in \hat{\mathscr{S}}$ such that any inextendible non-spacelike curve which intersects $\hat{\mathscr{S}}$ also intersects both $J^+(p)$ and $J^-(p)$;

then either the Cauchy development of $\hat{\mathscr{S}}$ is flat, or $\hat{\mathscr{M}}$ is timelike geodesically incomplete.

Condition (3″) requires that an observer at p can see, and be seen by, every particle that intersects $\hat{\mathscr{P}}$. The method of proof is to consider all spacelike surfaces without edge which contain p. One can form a topological space $S(p)$ out of all these surfaces, in a manner analogous to that in which one forms a topological space out of all the non-spacelike curves between two points. Conditions (2) and (3″) then imply that $S(p)$ is compact. One can show that the area of the surfaces is an upper semi-continuous function on $S(p)$ and so there will be some surface $\mathscr{S}'$ through p which has an area greater than or equal to that of any other surface. By a variation argument similar to that used for non-spacelike curves, one can show that $\chi^a{}_a$ vanishes everywhere on $\mathscr{S}'$ except possibly at p, where the surface may not be differentiable.

Consider a one-parameter family of spacelike surfaces $\mathscr{S}(u)$ where $\mathscr{S}(0) = \mathscr{S}'$. The variation vector $\mathbf{W} \equiv \partial/\partial u$ can be expressed as $f\mathbf{n}$ where $\mathbf{n}$ is the unit normal to the surfaces and f is some function. One can apply the Raychaudhuri equation to the congruence of integral curves of $\mathbf{W}$ to show

$$\partial\theta/\partial u = f\{-\tfrac{1}{3}\theta^2 - 2\sigma^2 - R_{ab}n^a n^b + f^{-1}f_{;ab}h^{ab}\},$$

where $$\theta \equiv \chi^a{}_a, \quad \sigma_{ab} \equiv \chi_{ab} - \tfrac{1}{3}\theta h_{ab}, \quad h_{ab} \equiv g_{ab} + n_a n_b,$$

and $$\sigma^2 = \tfrac{1}{2}\sigma_{ab}\sigma^{ab}.$$

If there is some point $q \in \mathscr{S}'$ at which $R_{ab}n^a n^b \neq 0$ or $\chi_{ab} \neq 0$ one can find an f such that $\partial\theta/\partial u$ is negative everywhere on S'. If $R_{ab}n^a n^b$ and χ_{ab} were zero everywhere on $\mathscr{S}'$, but there was some point q on $\mathscr{S}'$ at which $C_{abcd}n^b n^d$ was not equal to zero, then $\partial\sigma/\partial u \neq 0$ and one could find an f such that $\partial\theta/\partial u = 0$ and $\partial^2\theta/\partial u^2 < 0$ everywhere on $\mathscr{S}'$. In either case, one would obtain a surface $\mathscr{S}''$ on which $\chi^a{}_a < 0$ everywhere, and so $\hat{\mathscr{M}}$ would be timelike geodesically incomplete by theorem 4. If R_{ab}, χ_{ab} and $C_{abcd}n^b n^d$ were zero everywhere on $\mathscr{S}'$, then the Ricci identities for n^a show that $C_{abcd} = 0$ on $\mathscr{S}'$. Hence space–time is flat in $D(\hat{\mathscr{P}})$. An example in which conditions (1″), (2) and (3″) hold and in which $D(\mathscr{S})$ is flat is Minkowski space with $\{x^1, x^2, x^3, x^4\}$ identified with $\{x^1+1, x^2, x^3, x^4\}$, $\{x^1, x^2+1, x^3, x^4\}$, and $\{x^1, x^2, x^3+1, x^4\}$. This is geodesically complete. However the example given previously also satisfies these conditions and shows that $D(\mathscr{S})$ can be both geodesically incomplete and flat.

8.3 The description of singularities

The preceding theorems prove the occurrence of singularities in a large class of solutions but give little information as to their nature. To investigate this in more detail, one would need to define what one meant by the size, shape, location and so on of a singularity. This would be fairly easy if the singular points were included in the space–time manifold. However it would be impossible to determine the manifold structure at such points by physical measurements. In fact there would be many manifold structures which agreed for the non-singular regions but which differed for the singular points. For example, the manifold at the $t = 0$ singularity in the Robertson–Walker solutions could be that described by the coordinates

$$\{t, r\cos\theta, r\sin\theta\cos\phi, r\sin\theta\sin\phi\}$$

or that described by

$$\{t, Sr\cos\theta, Sr\sin\theta\cos\phi, Sr\sin\theta\sin\phi\}.$$

In the first case the singularity would be a three-surface, in the second case a single point.

What is needed is a prescription for attaching some sort of boundary ∂ to $\mathscr{M}$ which is uniquely determined by measurements at non-singular points, i.e. by the structure of $(\mathscr{M}, \mathbf{g})$. One would then like to define at least a topology, and possibly a differentiable structure and metric, on the space $\mathscr{M}^+ \equiv \mathscr{M} \cup \partial$. One possibility would be to use the method of indecomposable infinity sets described in §6.8. However since this depends only on the conformal metric, it does not distinguish between infinity and singular points at a finite distance. To make this distinction it would seem one should base one's construction for $\mathscr{M}^+$ on the criterion that has been adopted for the existence of a singularity: namely b-incompleteness. An elegant way of doing this has been developed by Schmidt. This supersedes earlier constructions by Hawking (1966*b*) and Geroch (1968*a*) which defined the singular points as equivalence classes of incomplete geodesics. These constructions did not necessarily provide endpoints for all b-incomplete curves, such as incomplete timelike curves of bounded acceleration. There was also a certain ambiguity in their definition of equivalence classes. Schmidt's construction does not suffer from these weaknesses.

Schmidt's procedure is to define a positive definite metric $\mathbf{e}$ on the bundle of orthonormal frames $\pi: O(\mathscr{M}) \to \mathscr{M}$. Here $O(\mathscr{M})$ is the set of all orthonormal four-tuples of vectors $\{\mathbf{E}_a\}$, $\mathbf{E}_a \in T_p$ for each $p \in \mathscr{M}$

(a ranges from 1 to 4), and π is the projection which maps a basis at a point p to the point p. It turns out that $O(\mathscr{M})$ is m-incomplete in the metric $\mathbf{g}$ if and only if $\mathscr{M}$ is b-incomplete. If $O(\mathscr{M})$ is m-incomplete, one can form the metric space completion $\overline{O(\mathscr{M})}$ of $O(\mathscr{M})$ by Cauchy sequences. The projection π can be extended to $\overline{O(\mathscr{M})}$, and the quotient of $\overline{O(\mathscr{M})}$ by π is defined to be $\mathscr{M}^+$ which is the union of $\mathscr{M}$ with a set of additional points ∂. The set ∂ consists of the singular points of $\mathscr{M}$ in the sense that it is the set of endpoints for every b-incomplete curve in $\mathscr{M}$.

To perform this construction, we recall (§ 2.9) that the connection on $\mathscr{M}$ given by the metric $\mathbf{g}$ defines a four-dimensional *horizontal subspace* H_u of the ten-dimensional tangent space T_u at the point $u \in O(\mathscr{M})$. Then T_u is the direct sum of H_u and the vertical subspace V_u consisting of all the vectors in T_u which are tangent to the fibre $\pi^{-1}(\pi(u))$. We now construct a basis $\{\mathbf{G}_A\} = \{\overline{\mathbf{E}}_a, \mathbf{F}_i\}$ for T_u where A runs from 1 to 10, a runs from 1 to 4 and i runs from 1 to 6; $\{\overline{\mathbf{E}}_a\}$ is a basis for H_u, and $\{\mathbf{F}_i\}$ is a basis for V_u.

Given any vector $\mathbf{X} \in T_{\pi(u)}(\mathscr{M})$ there is a unique vector $\overline{\mathbf{X}} \in H_u(O(\mathscr{M}))$ such that $\pi_* \overline{\mathbf{X}} = \mathbf{X}$. Thus on $O(\mathscr{M})$ there are four uniquely defined horizontal vector fields $\overline{\mathbf{E}}_a$ which are the horizontal lifts of the orthonormal basis vectors $\mathbf{E}_a$ for each point $u \in O(\mathscr{M})$. The integral curves of the field $\overline{\mathbf{E}}_a$ in $O(\mathscr{M})$ represent parallel propagation of the basis $\{\mathbf{E}_a\}$ along the geodesic in $\mathscr{M}$ in the direction of the vector $\mathbf{E}_a$.

The group $O(3, 1)$, the multiplicative group of all non-singular 4×4 real Lorentz matrices A_{ab}, acts in the fibres of $O(\mathscr{M})$ sending a point $u = \{p, \mathbf{E}_a\} \in O(\mathscr{M})$ to the point $A(u) = \{p, A_{ab}\mathbf{E}_b\} \in O(\mathscr{M})$. One can regard $O(3, 1)$ as a six-dimensional manifold and represent the tangent space $T_I(O(3, 1))$ to $O(3, 1)$ at the unit matrix I by the vector space of all 4×4 matrices a such that $a_{ab}G_{bc} = -a_{cb}G_{ba}$. Then if $a \in T_I(O(3, 1))$, one can define a curve in $O(3, 1)$ by $A_t = \exp(ta)$ where

$$\exp(b) = \sum_{n=0}^{\infty} \frac{b^n}{n!}.$$

Thus if $u \in O(\mathscr{M})$ one can define a curve through u in $\pi^{-1}(\pi(u))$ by $\lambda_{au}(t) = A_t(u)$. As the curve $\lambda_{au}(t)$ lies in the fibre, its tangent vector $(\partial/\partial t)_{\lambda_{au}}$ is vertical. For each $a \in T_I$, one can therefore define a vertical vector field $\mathbf{F}(a)$ by $\mathbf{F}(a)|_u = (\partial/\partial t)_{\lambda_{au}}|_u$ for each $u \in O(\mathscr{M})$. If $\{a_i\}$ $(i = 1, 2, \ldots, 6)$ are a basis for T_I, then $\mathbf{F}_i \equiv \mathbf{F}(a_i)$ will be six vertical vector fields on $O(\mathscr{M})$ which will provide a basis for V_u at each point $u \in O(\mathscr{M})$.

A matrix $B \in O(3,1)$ defines a mapping $O(\mathscr{M}) \to O(\mathscr{M})$ by $u \to B(u)$. Under the induced map $B_*: T_u \to T_{B(u)}$, the vertical and horizontal vector fields transform as follows:

$$B_*(\overline{\mathbf{E}}_a) = B^{-1}_{ab}\,\overline{\mathbf{E}}_b,$$

$$B_*(\mathbf{F}_i) = C_i{}^j\,\mathbf{F}_j,$$

where $C_i{}^j = B_{ab}\,a_{i\,bc}\,B^{-1}{}_{cd}\,a^j{}_{da}$ and $\{a^j\}$ are the basis for $T^*{}_I$ dual to the basis $\{a_i\}$ for T_I(thus $a^i{}_{ab}\,a_{jab} = \delta^i{}_j$, $a^j{}_{ab}\,a_{jcd} = \frac{1}{4}\delta_{ac}\delta_{bd}$). The property of these induced maps which will be important for what follows is not their actual form but the fact that they are constant over $O(\mathscr{M})$.

One now has a basis $\{\mathbf{G}_A\} = \{\overline{\mathbf{E}}_a, \mathbf{F}_i\}$ $(A = 1, \ldots, 10)$ for T_u at each point $u \in O(\mathscr{M})$. One can thus define a positive definite metric $\mathbf{e}$ on $O(\mathscr{M})$ by $e(\mathbf{X}, \mathbf{Y}) = \sum_A X^A Y^A$ where $\mathbf{X}, \mathbf{Y} \in T(u)$ and X^A, Y^A are the components of $\mathbf{X}$, $\mathbf{Y}$ respectively in the basis $\{\mathbf{G}_A\}$.

Using the metric $\mathbf{e}$, one can define a distance function $\rho(u, v)$, $u, v \in O(\mathscr{M})$, as the greatest lower bound of lengths (measured by $\mathbf{e}$) of curves from u to v. One can then ask whether $O(\mathscr{M})$ is m-complete with the distance function ρ.

Proposition 8.3.1

$(O(\mathscr{M}), \mathbf{e})$ is m-complete if and only if $(\mathscr{M}, \mathbf{g})$ is b-complete.

Suppose $\gamma(t)$ is a curve in $\mathscr{M}$. Then given a point $u \in \pi^{-1}(p)$ where $p \in \gamma$ one can construct a horizontal curve $\overline{\gamma}(t)$ through u such that $\pi(\overline{\gamma}(t)) = \gamma(t)$. From the definition of the positive definite metric $\mathbf{e}$, it follows that the arc-length of $\overline{\gamma}(t)$ as measured in this metric is equal to the generalized affine parameter of $\gamma(t)$, defined by the basis at p represented by the point u. If therefore $\gamma(t)$ has no endpoint but has finite length as measured by the generalized affine parameter, then $\overline{\gamma}(t)$ will also have no endpoint but will have finite length in the metric $\mathbf{e}$. Thus m-completeness in $O(\mathscr{M})$ implies b-completeness in $\mathscr{M}$.

To prove the converse, one needs to show that if $\lambda(t)$ is a C^1 curve in $O(\mathscr{M})$ of finite length without endpoint, then $\pi(\lambda(t))$ is a C^1 curve in $\mathscr{M}$ with

(1) finite affine length,

(2) no endpoint in $\mathscr{M}$.

To prove (1), one proceeds as follows. Let $u \in \lambda(t)$. Then one can construct a horizontal curve $\overline{\lambda}(t)$ through u such that $\pi(\overline{\lambda}(t)) = \pi(\lambda(t))$. For each value of t, $\lambda(t)$ and $\overline{\lambda}(t)$ will lie in the same fibre, so there will

be a unique curve $B(t)$ in $O(3,1)$ such that $\lambda(t) = B(t)\bar{\lambda}(t)$. This implies

$$\left(\frac{\partial}{\partial t}\right)_{\lambda} = B_* \left(\frac{\partial}{\partial t}\right)_{\bar{\lambda}} + F(B^{\cdot}B^{-1}),$$

where $B^{\cdot} \equiv \mathrm{d}B/\mathrm{d}t$. Therefore

$$e\left(\left(\frac{\partial}{\partial t}\right)_{\lambda}, \left(\frac{\partial}{\partial t}\right)_{\lambda}\right) = \sum_b \left(\left\langle \overline{\mathbf{E}}^a, \left(\frac{\partial}{\partial t}\right)_{\bar{\lambda}} \right\rangle B^{-1}{}_{ab}\right)^2 + \sum_i (B^{\cdot}{}_{ab} B^{-1}{}_{bc} a^i{}_{ca})^2,$$

where $\{\overline{\mathbf{E}}^a\}$ is the basis of $H^*{}_u$ dual to the basis $\{\overline{\mathbf{E}}_a\}$ (i.e. $\langle \overline{\mathbf{E}}^a, \overline{\mathbf{E}}_b \rangle = \delta^a{}_b$) and $a^i{}_{ab}$ is the basis of $T_I{}^*$ dual to the basis $a_{i\,ab}$ (i.e. $a_{i\,ab} a^j{}_{ab} = \delta_i{}^j$).

The matrix B_{ab} satisfies $B_{ab} G_{bc} B_{dc} = G_{ad}$. Therefore

$$B_{ab} G_{ac} B_{cd} = G_{bd}$$

as $G_{ab} = G^{-1}{}_{ab}$. Differentiating with respect to t, one has

$$B^{\cdot}{}_{ab} B^{-1}{}_{bc} G_{cd} = -G_{ac} B^{\cdot}{}_{db} B^{-1}{}_{bc}.$$

Thus $B^{\cdot}{}_{ab} B^{-1}{}_{bc} \in T_I(O(3,1))$. Since the $a^i{}_{ab}$ are a basis for $T^*{}_I$, there is some constant C such that

$$\sum_i (B^{\cdot}{}_{ab} B^{-1}{}_{bc} a^i{}_{ca})^2 \geqslant C(B^{\cdot}{}_{ab} B^{-1}{}_{bc} B^{\cdot}{}_{ad} B^{-1}{}_{dc}).$$

Any matrix $B \in O(3,1)$ can be expressed in the form $B = \bar{\Omega}\Delta\Omega$, where (i) $\bar{\Omega}$ and Ω are orthogonal matrices of the form

$$\left(\begin{array}{c|c} \bar{O} & \\ \hline & 1 \end{array}\right) \quad \text{and} \quad \left(\begin{array}{c|c} O & \\ \hline & 1 \end{array}\right)$$

where $\bar{O}$ and O are 3×3 orthogonal matrices, and the basis $\{\mathbf{E}_a\}$ has been numbered so that $\mathbf{E}_4$ is the timelike vector; these matrices represent rotations; and (ii) Δ is the matrix

$$\begin{pmatrix} \cosh\xi & 0 & 0 & \sinh\xi \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \sinh\xi & 0 & 0 & \cosh\xi \end{pmatrix}$$

which represents a change of velocity in the 1-direction. With this decomposition,

$$B^{\cdot}{}_{ab} B^{-1}{}_{bc} B^{\cdot}{}_{ad} B^{-1}{}_{dc} \geqslant 2(\xi^{\cdot})^2.$$

For any vector $\mathbf{X} \in T_u$,

$$\sum_b (\langle \overline{\mathbf{E}}^a, \mathbf{X} \rangle \Omega_{ab})^2 = \sum_a (\langle \overline{\mathbf{E}}^a, \mathbf{X} \rangle)^2.$$

Thus
$$\sum_b \left(\left\langle \mathbf{E}^a, \left(\frac{\partial}{\partial t}\right)_{\bar\lambda} \right\rangle B^{-1}{}_{ab} \right)^2 \geqslant \sum_a \left(\left\langle \mathbf{E}^a, \left(\frac{\partial}{\partial t}\right)_{\bar\lambda} \right\rangle \right)^2 e^{-2|\xi|}$$
$$= e\left(\left(\frac{\partial}{\partial t}\right)_{\bar\lambda}, \left(\frac{\partial}{\partial t}\right)_{\bar\lambda} \right) e^{-2|\xi|}.$$

Therefore
$$e\left(\left(\frac{\partial}{\partial t}\right)_{\lambda}, \left(\frac{\partial}{\partial t}\right)_{\lambda} \right) \geqslant e\left(\left(\frac{\partial}{\partial t}\right)_{\bar\lambda}, \left(\frac{\partial}{\partial t}\right)_{\bar\lambda} \right) e^{-2|\xi|} + 2C(\xi^{\cdot})^2,$$

and so
$$\left[e\left(\left(\frac{\partial}{\partial t}\right)_{\lambda}, \left(\frac{\partial}{\partial t}\right)_{\lambda} \right) \right]^{\frac{1}{2}} \geqslant \frac{1}{2} \left[e\left(\left(\frac{\partial}{\partial t}\right)_{\bar\lambda}, \left(\frac{\partial}{\partial t}\right)_{\bar\lambda} \right) \right]^{\frac{1}{2}} e^{-|\xi|} + C^{\frac{1}{2}} |\xi^{\cdot}|.$$

Let $\xi_0 \leqslant \infty$ be the least upper bound for $|\xi|$ on $\lambda(t)$. Then
$$L(\lambda) \geqslant \tfrac{1}{2} L(\bar\lambda)\, e^{-\xi_0} + C^{\frac{1}{2}} \xi_0,$$

where $L(\lambda)$ is the length of the curve λ in the metric $\mathbf{e}$. Since this is finite, ξ_0 and $L(\bar\lambda)$ must be finite. Thus the affine length of the curve $\pi(\lambda(t))$ in $\mathscr{M}$, which is equal to $L(\bar\lambda)$, will be finite.

To complete the proof of proposition 8.3.1, we have to show that the curve $\pi(\lambda(t))$ in $\mathscr{M}$ has no endpoint, that is, we have to show that there is no point $p \in \mathscr{M}$ such that $\pi(\lambda(t))$ enters and remains within every neighbourhood $\mathscr{U}$ of p. Because of the existence of normal neighbourhoods $\mathscr{U}$ of p, this is a consequence of the following result:

Proposition 8.3.2 (*Schmidt* (1972))
Let $\mathscr{N}$ be a compact subset of $\mathscr{M}$. Suppose there is a curve $\lambda(t)$ in $O(\mathscr{M})$ without endpoint and of finite length, which enters and remains within $\pi^{-1}(\mathscr{N})$. Then there is an inextendible null geodesic γ contained in $\mathscr{N}$.

Let $\bar\lambda(t)$ be the horizontal curve through some point $u \in \lambda(t)$ such that $\pi(\bar\lambda(t)) = \pi(\lambda(t))$. The curve $\lambda(t)$ has no endpoint. Suppose there were a point $v \in O(\mathscr{M})$ which was an endpoint of the horizontal curve $\bar\lambda(t)$. Then there would be an open neighbourhood $\mathscr{W}$ of v with compact closure such that $\bar\lambda(t)$ entered and remained within $\mathscr{W}$. Let $\mathscr{W}'$ be the set $\{x \in O(\mathscr{M})\colon Bx \in \mathscr{W}$ for all matrices B with $|\xi| \leqslant \xi_0\}$. Since $\overline{\mathscr{W}}$ was compact and ξ_0 is finite, $\overline{\mathscr{W}'}$ would be compact. The curve $\lambda(t)$ would enter and remain within $\overline{\mathscr{W}'}$. But any compact set is m-complete with respect to the positive definite metric $\mathbf{e}$. Thus $\lambda(t)$, having finite length, would have an endpoint in $\overline{\mathscr{W}'}$. This shows that $\bar\lambda(t)$ has no endpoint.

Let $\{x_n\}$ be a sequence of points on $\bar{\lambda}(t)$ without any limit point. Since $\mathscr{N}$ is compact, there will be a point $x \in \mathscr{N}$ which is a limit point of $\pi(x_n)$. Let $\mathscr{U}$ be a normal neighbourhood of x with compact closure, and let $\sigma: \mathscr{U} \to O(\mathscr{M})$ be a cross-section of $O(\mathscr{M})$ over $\mathscr{U}$, i.e. $\sigma(p)$, $p \in \mathscr{U}$, is an orthonormal basis at p. Let $\tilde{\lambda}(t) \equiv \sigma(\pi(\lambda(t)))$ for $\lambda(t) \in \pi^{-1}(\mathscr{U})$. Then as in the previous proposition, there will be a unique family of matrices $A(t) \in O(3,1)$ such that $\bar{\lambda}(t) = A(t)\tilde{\lambda}(t)$, and one can express the matrix A in the form $A = \bar{\Omega}\Delta\Omega$. Suppose that $|\xi(t_{n'})|$ had a finite upper bound ξ_1, where $x_{n'} = \bar{\lambda}(t_{n'})$ is a subsequence of the x_n which converges to x. Then the points $x_{n'}$ would be contained in the set $\mathscr{U}' = \{v \in O(\mathscr{M}): A^{-1}v \subset \sigma(\mathscr{U})$ for some $A \in O(3,1)$ with $|\xi| < \xi_1\}$. However $\overline{\mathscr{U}'}$ would be compact and so would contain a limit point of the $\{x_{n'}\}$, which is contrary to our choice of the $\{x_n\}$. Thus $|\xi(t_{n'})|$ has no finite upper bound. Since the orthogonal group is compact, one can choose a subsequence $\{x_{n''}\}$ such that $\bar{\Omega}_{n''}$ converges to some $\bar{\Omega}'$, $\Omega_{n''}$ converges to some Ω', $\xi_{n''} \to \infty$, and

$$\xi_{n''+1} - \xi_{n''} > a > 0 \tag{8.1}$$

for some constant a (here $\bar{\Omega}_{n''} = \bar{\Omega}(t_{n''})$, etc.).

Let $\lambda'(t) = (\bar{\Omega}')^{-1}\bar{\lambda}(t)$, and let $\hat{\lambda}_{n''}(t) \equiv \Delta_{n''}{}^{-1}(\bar{\Omega}')^{-1}\bar{\lambda}(t)$. Then $\hat{\lambda}_{n''}(t_{n''})$ tends to $\hat{x} \equiv \Omega'\sigma(x)$. Since the length of the curve $\bar{\lambda}(t)$ is finite, the curve $\lambda'(t)$ also has finite length. This means that

$$\int_{t_{n''}}^{t_{n''+1}} ((X^u)^2 + (X^v)^2 + (X^2)^2 + (X^3)^2)^{\frac{1}{2}}\,dt$$

tends to zero, where

$$X^A \equiv \langle \bar{\mathbf{E}}^A, (\partial/\partial t)_{\lambda'} \rangle, \quad A = u, v, 2, 3,$$

and

$$\mathbf{E}^u = \frac{1}{\sqrt{2}}(\mathbf{E}^4 + \mathbf{E}^1), \quad \mathbf{E}^v = \frac{1}{\sqrt{2}}(\mathbf{E}^4 - \mathbf{E}^1).$$

Thus

$$\int_{t_{n''}}^{t_{n''+1}} |X^A|\,dt$$

tends to zero, for each A. The components $Y_{n''}{}^A$ of the tangent vector of the horizontal curve $\hat{\lambda}_{n''}(t)$ are

$$Y_{n''}{}^u = e^{-\xi_{n''}} X^u, \quad Y_{n''}{}^v = e^{\xi_{n''}} X^v, \quad Y_{n''}{}^2 = X^2, \quad Y_{n''}{}^3 = X^3.$$

Thus

$$\int_{t_{n''}}^{t_{n''+1}} |Y_{n''}{}^A|\,dt \quad (A = u, 2, 3), \tag{8.2}$$

tend to zero.

Let μ be the integral curve of the horizontal vector field $\mathbf{E}^v$ through $\hat{x}$. Then $\pi(\mu)$ will be a null geodesic in $\mathscr{M}$. Suppose that $\pi(\mu)$ left $\mathscr{N}$ in both the past and future directions. Then there would be some neighbourhood $\mathscr{V}$ of $\hat{x}$ with compact closure and with the property that in each direction μ left and did not re-enter the set $\overline{\mathscr{V}'}$, where $\mathscr{V}' \equiv \{v \in O(\mathscr{M})$: there is a Δ with Δv contained in $\mathscr{V}\}$. One could choose $\mathscr{V}$ sufficiently small that it had this property for any integral curve of $\mathbf{E}^v$ which intersected $\overline{\mathscr{V}}$ and so that any such curve would leave $\pi^{-1}(\mathscr{N})$ in both directions. Let $\mathscr{Y}$ be the tube consisting of all points on integral curves of $\mathbf{E}^v$ which intersect $\overline{\mathscr{V}}$. Then $\mathscr{Y} \cap \pi^{-1}(\mathscr{N})$ would be compact. For sufficiently large n'', $\hat{\lambda}_{n''}(t_{n''})$ would be contained in $\mathscr{V}$. By (8.2) the components of the tangent vector to $\hat{\lambda}_{n''}$ transverse to the direction $\mathbf{E}^v$ are so small that for large n'' and $t > t_{n''}$, the curve $\hat{\lambda}_{n''}(t)$ could not leave the tube $\mathscr{Y} \cap \pi^{-1}(\mathscr{N})$ except at its ends where $\mathscr{Y}$ left $\pi^{-1}(\mathscr{N})$. However $\hat{\lambda}_{n''}(t)$ cannot leave $\pi^{-1}(\mathscr{N})$, as $\lambda(t)$ does not leave $\pi^{-1}(\mathscr{N})$. Thus $\hat{\lambda}_{n''}(t)$ would be contained in $\mathscr{Y} \cap \pi^{-1}(\mathscr{N})$ for $t \geqslant t_{n''}$. This leads to a contradiction as follows: $\hat{\lambda}_{n''+1}(t_{n''+1})$ is contained in $\mathscr{V}$. However by (8.1), $\mathscr{V}$ can be chosen sufficiently small that

$$\hat{\lambda}_{n''}(t_{n''+1}) = \Delta_{n''+1}\Delta_{n''}{}^{-1}\hat{\lambda}_{n''+1}(t_{n''+1})$$

is not contained in $\mathscr{V}$, though it is contained in $\mathscr{V}'$. This shows that our assumption that the null geodesic $\pi(\mu)$ left $\mathscr{N}$ in both directions is false. Thus there will be some point $p \in \mathscr{N}$ which is a limit point of $\pi(\mu)$. By lemma 6.2.1 there will be an inextendible null geodesic γ through p which is contained in $\mathscr{N}$ and which is a limit curve of $\pi(\mu)$. □

If $O(\mathscr{M})$ is m-incomplete, one can form the metric space completion $\overline{O(\mathscr{M})}$. This is defined to be the set of equivalence classes of Cauchy sequences of points in $O(\mathscr{M})$. If $x \equiv \{x_n\}$ and $y \equiv \{y_m\}$ are Cauchy sequences in $O(\mathscr{M})$, the distance $\bar{\rho}(x, y)$ between x and y is defined to be $\lim_{n\to\infty} \rho(x_n, y_n)$ where ρ is the distance function on $O(\mathscr{M})$ defined by the positive definite metric $\mathbf{e}$; x and y are said to be equivalent if $\bar{\rho}(x, y) = 0$. One can decompose $\overline{O(\mathscr{M})}$ into a part homeomorphic to $O(\mathscr{M})$ and a set of boundary points $\bar{\partial}$ (i.e. $\overline{O(\mathscr{M})} = O(\mathscr{M}) \cup \bar{\partial}$). The distance function $\bar{\rho}$ defines a topology on $\overline{O(\mathscr{M})}$. From (8.1), it follows that the topology on $\overline{O(\mathscr{M})}$ is independent of the choice of basis $\{a_i\}$ of T_I.

One can extend the action of $O(3,1)$ to $\overline{O(\mathscr{M})}$. For under the action of $A \in O(3,1)$, the transformation of the basis $\{\mathbf{G}_A\}$ is independent of position in $O(\mathscr{M})$. Thus there are positive constants C_1 and C_2 (depending only on A) such that $C_1\rho(u,v) \leqslant \rho(A(u),A(v)) \leqslant C_2\rho(u,v)$. This means that under the action of A, Cauchy sequences will map to Cauchy sequences and equivalence classes of Cauchy sequences are mapped to equivalence classes of Cauchy sequences. Therefore the action of $O(3,1)$ extends to $\overline{O(\mathscr{M})}$ in a unique way. One can then define $\mathscr{M}^+$ to be the quotient of $\overline{O(\mathscr{M})}$ by the action of $O(3,1)$. Since the quotient of $O(\mathscr{M})$ by $O(3,1)$ is $\mathscr{M}$, and since $O(3,1)$ maps incomplete Cauchy sequences to incomplete Cauchy sequences, one can express $\mathscr{M}^+$ as the union of $\mathscr{M}$ and a set ∂ of points called the *b-boundary* of $\mathscr{M}$. One can regard points of ∂ as representing the endpoint of equivalence classes of b-incomplete curves in $\mathscr{M}$.

The projection $\bar{\pi}\colon \overline{O(\mathscr{M})} \to \mathscr{M}^+$, which assigns a point in $\overline{O(\mathscr{M})}$ to its equivalence class under $O(3,1)$, induces a topology on $\mathscr{M}^+$ from the topology on $O(\mathscr{M})$. However $\bar{\pi}$ does not induce a distance function on $\mathscr{M}^+$ because $\bar{\rho}$ is not invariant under $O(3,1)$. Thus although the topology of $\overline{O(\mathscr{M})}$ is a metric topology, and so Hausdorff, that of $\mathscr{M}^+$ need not be Hausdorff. This means that there may be a point $p \in \mathscr{M}$ and a point $q \in \partial$ such that every neighbourhood of p in $\mathscr{M}^+$ intersects every neighbourhood of q. This happens when the point q corresponds to an incomplete curve which is totally or partially imprisoned in $\mathscr{M}$. We shall discuss imprisoned incompleteness further in §8.5.

If $\mathbf{g}$ is a positive definite metric on $\mathscr{M}$, then $\mathscr{M}^+$ is homeomorphic to the completion of $(\mathscr{M}, \mathbf{g})$ by Cauchy sequences. Schmidt's construction also has the desirable property that if one cuts a closed set $\mathscr{A}$ out of a space, then one gets at least one point of the b-boundary for every point of $\mathscr{A}^\cdot$ that is the endpoint of a curve in $\mathscr{M} - \mathscr{A}$. An example where one gets more than one b-boundary point for a point of $\mathscr{A}^\cdot$ is provided by two-dimensional Minkowski space in which the set $\mathscr{A}$ is taken to be the t-axis between -1 and $+1$. Then there will be two b-boundary points for each point $(0,t)$ where $-1 < t < 1$. An example where a point in $\mathscr{A}^\cdot$ cannot be reached by a curve in $\mathscr{M} - \mathscr{A}$ is given by the set

$$\mathscr{A} = \left\{t = \sin\frac{1}{x},\ t \neq 0\right\} \cup \{-1 \leqslant t \leqslant 1,\ x = 0\}.$$

There is no curve in $\mathscr{M} - \mathscr{A}$ which has an endpoint at the origin, and hence this point will not be in $(\mathscr{M} - \mathscr{A})^+$, although it is in $\mathscr{A}^\cdot$.

Although Schmidt's construction has an elegant formulation, it is unfortunately very difficult to apply in practice. The only solutions for which $\mathscr{M}^+$ has been found, apart from spaces of constant curvature, are the two-dimensional Robertson–Walker solutions with normal matter. In these ∂ turns out to be a spacelike one-surface as might be expected from the conformal picture. In this case, one can define a natural differential structure on ∂ and make $\mathscr{M}^+$ into a manifold with boundary. However there does not seem to be any general way of defining a manifold structure on ∂. Indeed one might expect that in generic situations ∂ would be highly irregular and could not be given a smooth structure.

8.4 The character of the singularities

In this and the following section we shall discuss the character of the singularities predicted by theorem 4. We consider this theorem rather than the others because more information about the singularity can be obtained. We expect however that the singularities predicted by the other theorems will have similar properties.

First there is the question of how bad the breakdown of differentiability of the metric must be. The theorems of the previous section showed that space–time must be geodesically incomplete if the metric was C^2. The C^2 condition was necessary in order that the conjugate points and variation of arc-length should be well-defined; in other words, in order that solutions of the geodesic equation should depend *differentiably* on their initial position and direction. However one can talk about geodesic incompleteness provided that solutions of the geodesic equation are defined. They will exist if the metric is C^1 and will be unique and depend *continuously* on initial position and direction if the metric is C^{2-} (i.e. if the connection is locally Lipschitz). In fact one can discuss b-incompleteness provided merely that the positive definite metric $\mathbf{e}$ on the bundle of frames $O(\mathscr{M})$ is defined almost every where and is locally bounded. This will be the case if the components $\Gamma^a{}_{bc}$ of the connection are defined almost everywhere and are locally bounded, i.e. if the metric is C^{1-}.

It thus might appear that what the theorems indicate is not that the curvature becomes unboundedly large but merely that it has a discontinuity (i.e. the metric is C^{2-} rather than C^2). We shall show that this is not the case: under the conditions of theorem 4 space–time must be timelike geodesically incomplete (and hence b-incomplete) even if

the metric is only required to be C^{2-}. The method of proof is to approximate the C^{2-} metric by a C^2 metric and to perform variation of arc-length in this metric.

Suppose that space–time is defined to be inextendible with a C^{2-} metric and that the conditions of theorem 4 are satisfied. The timelike convergence condition, $R_{ab}K^aK^b \geqslant 0$, is now required to hold 'almost everywhere' with the Ricci tensor defined by generalized derivatives. The only part of the proof of theorem 4 that does not hold in a C^{2-} metric is where variation of arc-length is used to show that there can be no point $p \in D^+(\mathscr{S})$ such that $d(\mathscr{S}, p) > -3/\theta_0$, where θ_0 is the maximum value of $\chi^a{}_a$ on $\mathscr{S}$. Thus if $\mathscr{M}$ were timelike geodesically complete there would be some such point p and a geodesic orthogonal to $\mathscr{S}$ of length $d(\mathscr{S}, p)$ from $\mathscr{S}$ to p. Let $\mathscr{U}$ be an open set with compact closure which contains $J^-(p) \cap J^+(\mathscr{S})$ and let $\mathbf{e}$ and $\hat{\mathbf{g}}$ be C^∞ positive definite and Lorentz metrics respectively. For any $\epsilon > 0$ one could find a C^∞ Lorentz metric $g_\epsilon{}^{ab}$ such that on $\overline{\mathscr{U}}$

(1) $|g_\epsilon{}^{ab} - g^{ab}| < \epsilon$,
(2) $|g_\epsilon{}^{ab}{}_{|c} - g^{ab}{}_{|c}| < \epsilon$,
(3) $|g_\epsilon{}^{ab}{}_{|cd}| < C$, where C is a constant depending on $\mathscr{U}$, $\mathbf{e}$, $\hat{\mathbf{g}}$ and $\mathbf{g}$,
(4) $R_{\epsilon ab}K^aK^b > -\epsilon|K^a|^2$ for any vector $\mathbf{K}$ such that $g_{\epsilon ab}K^aK^b \geqslant 0$.

(The $g_\epsilon{}^{ab}$ may be constructed by covering $\overline{\mathscr{U}}$ by a finite number of local coordinate neighbourhoods $(\mathscr{V}_\alpha, \phi_\alpha)$, integrating the coordinate components of g^{ab} with a suitable smoothing function $\rho_\epsilon(x)$ and summing with a partition of unity $\{\psi_\alpha\}$, i.e.

$$g_\epsilon{}^{ab}(q) = \sum_\alpha \psi_\alpha(q) \int_{\phi_\alpha(\mathscr{V}_\alpha)} g^{ab}(x)\, \rho_\epsilon(x - \phi_\alpha(q))\, \mathrm{d}^4x,$$

where $\int \rho_\epsilon(x)\, \mathrm{d}^4x = 1$.)

Property (1) implies that for sufficiently small values of ϵ, p would be in $D^+(\mathscr{S}, \mathbf{g}_\epsilon)$ and $J^-(p, \mathbf{g}_\epsilon) \cap J^+(\mathscr{S}, \mathbf{g}_\epsilon)$ would be contained in $\mathscr{U}$. There would therefore be a geodesic γ_ϵ in the metric $\mathbf{g}_\epsilon$ from $\mathscr{S}$ to p of length $d_\epsilon(\mathscr{S}, p)$. Also $|d_\epsilon(\mathscr{S}, p) - d(\mathscr{S}, p)|$ would tend to zero as $\epsilon \to 0$.

By properties (1), (2) and (3), and the standard theorems on ordinary differential equations, as $\epsilon \to 0$ the tangent vector to a geodesic in the metric $\mathbf{g}_\epsilon$ would tend to that of the geodesic in the metric $\mathbf{g}$ with the same initial position and direction. There would be some upper bound to $|V^a|$ on $\overline{\mathscr{U}} \cap \beta(\mathscr{S} \times [0, 2d(\mathscr{S}, p)])$, where V^a is the unit tangent vector to the geodesic orthogonal to $\mathscr{S}$ in the metric $\mathbf{g}$. Thus for any $\delta > 0$ there would be an $\epsilon_1 > 0$ such that for any $\epsilon < \epsilon_1$, $R_{\epsilon ab}V_\epsilon^a V_\epsilon^b > -\delta$. We can now establish a contradiction by showing that

a sufficiently small variation of the energy condition will not prevent the occurrence of conjugate points in the metric $\mathbf{g}_\epsilon$ within a distance less than $d_\epsilon(\mathscr{P}, p)$. For the expansion θ_ϵ of the geodesics in the metric $\mathbf{g}_\epsilon$ obeys the Raychaudhuri equation:

$$\mathrm{d}\theta_\epsilon/\mathrm{d}s = -\tfrac{1}{3}\theta_\epsilon{}^2 - 2\sigma_\epsilon{}^2 - R_{\epsilon ab}V_\epsilon^a V_\epsilon^b.$$

Thus $\mathrm{d}(\theta_\epsilon{}^{-1})/\mathrm{d}s \geqslant \frac{1}{3} + R_{\epsilon ab}V^a V^b \theta_\epsilon{}^{-2}$. Therefore if the initial value θ_{ϵ_0} were negative and $3\delta\theta_{\epsilon_0}{}^{-2}$ were less than one, $\theta_\epsilon{}^{-1}$ would become zero within a distance $3/\theta_0(1 - 3\delta\theta_0{}^{-2})$ from $\mathscr{S}$. But $\theta_{\epsilon_0} \to \theta_0$ as $\epsilon \to 0$. This shows that for sufficiently small values of ϵ there would be a conjugate point on every geodesic in the metric $\mathbf{g}_\epsilon$ orthogonal to $\mathscr{P}$ within a distance less than $d_\epsilon(\mathscr{P}, p)$. Therefore $\mathscr{M}$ must be timelike geodesically incomplete even if the metric is required only to be C^{2-}.

This result implies that if space–time is extended to try to continue the incomplete geodesics, the metric must fail to be Lorentzian or the curvature must be locally unbounded, i.e. there would be a curvature singularity. However even though the curvature were locally unbounded, the metric might still be able to be interpreted as a distributional solution of the Einstein equations provided that the volume integrals of the components of the curvature tensor over any compact region were finite. This would be the case if the metric were Lorentz, continuous and had square integrable first derivatives. In particular this would be true if the metric were Lorentz and C^{1-} (i.e. locally Lipschitz). Examples of such C^{1-} solutions include gravitational shock waves (where the curvature has a δ-function behaviour on a null three-surface, see, for example, Choquet-Bruhat (1968) and Penrose (1972*a*)); thin mass shells (where the curvature has a δ-function behaviour on a timelike three-surface, see, for example, Israel (1966)); and solutions containing pressure-free matter where the geodesic flow lines have two- or three-dimensional caustics (see Papapetrou and Hamoui (1967), Grischuk (1967)). Because of the non-linear dependence of the curvature on the metric one cannot necessarily approximate a C^{1-} distributional solution by a C^2 metric which obeys the convergence condition at every point, or at least does not violate it by more than a small amount as in the case above (property (4)). However in all the examples given above one can. Indeed this is their physical justification: they are regarded as mathematical idealizations of C^2 or C^∞ solutions which obey the convergence condition and in which the curvature is very large in a small region. One could apply the theorems of §8.2 to these C^2 solutions and prove the existence of

incomplete geodesics in them. This shows that the singularities predicted cannot be just gravitational impulse waves or caustics of flow lines but must be more serious breakdowns of the metric. (Ordinary hydrodynamic shock waves involve only discontinuities of density and pressure and so can exist with a C^{2-} metric.) Although we are not quite able to prove it we believe that the singularities must be such that the metric cannot be extended to be even a distributional solution of the Einstein equations, i.e. as well as the components of the curvature being unbounded at a singular point, their volume integral over any neighbourhood of such a point must also be unbounded. This is so in all known examples of singularities other than the exceptional case of the Taub–NUT solution, which will be dealt with in the next section. If this conjecture is correct for 'generic' singularities (i.e. except for those arising from a set of initial conditions of measure zero), then one can regard a singularity as a point where the Einstein equations (and presumably the other presently known laws of physics) break down.

Another question one would like to answer is: how many incomplete geodesics are there? If there were only one, one might be tempted to feel that the singularity could be ignored. From the proof of theorem 4 one can see that if there is no Cauchy horizon, i.e. if $\mathscr{P}$ is a Cauchy surface, then no timelike curve from $\mathscr{P}$ (geodesic or not) can be extended to a length greater than $-3/\theta_0$ where θ_0 is the maximum value of $\chi^a{}_a$ on $\mathscr{S}$. In fact this result is true even if $\mathscr{S}$ is non-compact provided that $\chi^a{}_a$ still has a negative upper bound. However this does not necessarily indicate that what happens is that every timelike curve hits the singularity. Rather it suggests that a singularity will be accompanied by a Cauchy horizon and so our ability to predict the future will break down. An example of this is shown in figure 53. Here the metric is singular at the point p and so this point has been removed from the space–time manifold. Spreading out from this hole there is a Cauchy horizon. This example shows that the most one can hope to prove is that there is a three-dimensional family of geodesics which are incomplete and which remain within the Cauchy development of $\mathscr{P}$ (in the example these are the geodesics which would pass through p). There may be other geodesics which leave the Cauchy development of $\mathscr{P}$ and which are incomplete but one cannot predict their behaviour from knowledge of conditions on $\mathscr{P}$.

It is clear that there must be more than one incomplete geodesic in $D^+(\mathscr{P})$. For from theorem 4 it follows that there must be a geodesic γ, orthogonal to $\mathscr{P}$, which remains in $D^+(\mathscr{P})$ but which is incomplete.

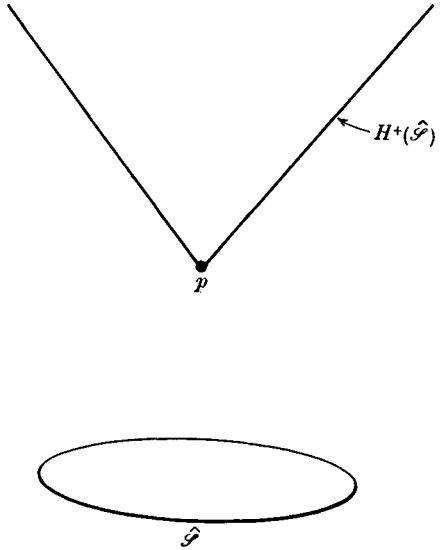

FIGURE 53. The point p has been removed from space–time because a singularity occurs there. Consequently there is a Cauchy horizon $H^+(\hat{\mathscr{S}})$ for the surface $\hat{\mathscr{S}}$.

Let p be the point where γ intersects $\hat{\mathscr{S}}$. Then one can make a small variation of $\hat{\mathscr{S}}$ in a neighbourhood of p to obtain a new surface $\hat{\mathscr{S}}'$ for which $\chi^a{}_a$ is still negative, but which is not orthogonal to γ. Then by theorem 4 there must be some other timelike geodesic γ' orthogonal to $\hat{\mathscr{S}}'$ which is incomplete and which does not cross $H^+(\hat{\mathscr{S}}')$, which is the same as $H^+(\hat{\mathscr{S}})$.

One can in fact prove that there is at least a three-dimensional family of timelike geodesics (one through each point of some achronal surface) which remain within $D^+(\hat{\mathscr{S}})$ and which are incomplete. These geodesics all correspond to the same boundary point in the sense of the indecomposable past sets of §6.8, that is, they all have the same past. They may not, however, all correspond to the same points as defined by the construction of the previous section. An outline of the proof is as follows: in theorem 4 it was shown that there must be a future-directed timelike geodesic orthogonal to $\mathscr{S}$ which cannot be extended to length $3/\theta_0$. One can say more than this: there must be such a geodesic γ which remains within $D^+(\hat{\mathscr{S}})$ and is at each point a curve of maximum length from $\hat{\mathscr{S}}$, i.e. for each $q \in \gamma$, the length of γ from $\hat{\mathscr{S}}$ to q equals $d(\hat{\mathscr{S}}, q)$. The idea is now to consider the function $d(r, \gamma)$ for $r \in J^-(\gamma)$. Clearly this is bounded on $J^+(\hat{\mathscr{S}}) \cap J^-(\gamma)$. From the fact that γ is a curve of maximum length from $\hat{\mathscr{S}}$, it follows that in a neighbour-

hood of γ, $d(r, \gamma)$ is continuous and the surfaces of constant $d(r, \gamma)$ are spacelike surfaces which intersect γ orthogonally. The timelike geodesics orthogonal to these surfaces will then remain within $J^-(\gamma)$ and so will be incomplete.

8.5 Imprisoned incompleteness

In §8.1 we proposed b-incompleteness as a definition of a singularity. The idea was that a b-incomplete curve corresponded to a singular point which had been left out of space–time. However suppose that there is a b-incomplete curve λ which has a limit point $p \in \mathscr{M}$, i.e. λ is partially or totally imprisoned in a compact neighbourhood of p. Then one cannot imbed $\mathscr{M}$ in a larger four-dimensional Hausdorff paracompact manifold $\mathscr{M}'$ such that λ can be continued in $\mathscr{M}'$. For if q were the point where λ intersected the boundary of $\mathscr{M}$ in $\mathscr{M}'$, then any neighbourhood of q would intersect any neighbourhood of p, which would be impossible as $\mathscr{M}'$ is Hausdorff and $q \neq p$. In fact, one can characterize imprisoned incompleteness of $\mathscr{M}$ by non-Hausdorff behaviour of the Schmidt completion $\mathscr{M}^+$.

Proposition 8.5.1

A point $p \in \mathscr{M}$ is not Hausdorff separated in $\mathscr{M}^+$ from a point $r \in \partial$ if there is an incomplete curve λ in $\mathscr{M}$ which has p as a limit point and which has r as an endpoint in $\mathscr{M}^+$.

Suppose that $p \in \mathscr{M}$ is a limit point of a b-incomplete curve λ. One can construct a horizontal lift $\bar{\lambda}$ of λ in the bundle of orthonormal frames $O(\mathscr{M})$. This will have an endpoint at some point

$$x \in \pi^{-1}(r) \subset \bar{\partial} \equiv \overline{O(\mathscr{M})} - O(\mathscr{M}).$$

If $\mathscr{V}$ is a neighbourhood of r in $\mathscr{M}^+$ then $\pi^{-1}(\mathscr{V})$ is an open neighbourhood of x in $\overline{O(\mathscr{M})}$. Thus it contains all points on $\bar{\lambda}$ beyond some point y. Therefore all points on λ beyond $\pi(y)$ will lie in $\mathscr{V}$ and hence $\mathscr{V}$ will intersect any neighbourhood of p since p is a limit point of λ. □

Taub–NUT space (§5.8) is an example where there are incomplete geodesics which are all totally imprisoned in compact neighbourhoods of the past and future horizons $U(t) = 0$. As the metric is perfectly regular on these compact neighbourhoods, the incomplete geodesis. do not correspond to s.p. (scalar polynomial) curvature singularities. Consider a future incomplete closed null geodesic $\lambda(v)$ in the future

horizon $U(t) = 0$. Let $p = \lambda(0)$ and let v_1 be the first positive value of v for which $\lambda(v) = p$. Then as in §6.4, the parallelly propagated tangent vector to λ will satisfy

$$(\partial/\partial v)|_{v=v_1} = a(\partial/\partial v)|_{v=0},$$

where $a > 1$. For each n, the point $\lambda(v_n) = p$, where

$$v_n = v_1 \sum_{r=1}^{n} a^{1-r} = v_1 \frac{1-a^{-n}}{1-a^{-1}},$$

and

$$(\partial/\partial v)|_{v=v_n} = a^n(\partial/\partial v)|_{v=0}.$$

Thus if one takes a pseudo-orthonormal parallelly propagated basis $\{\mathbf{E}_a\}$ on $\lambda(v)$, where $\mathbf{E}_4 = \partial/\partial v$, then the other null basis vector $\mathbf{E}_3$ obeys $\mathbf{E}_3|_{v=v_n} = a^{-n}\mathbf{E}_3|_{v=0}$. Each time one goes round the closed null geodesic λ, the vector $\mathbf{E}_4$ gets bigger and the vector $\mathbf{E}_3$ gets smaller. The vectors $\mathbf{E}_1$ and $\mathbf{E}_2$ remain the same. If therefore there were some non-zero component of the Riemann tensor which involved $\mathbf{E}_4$ and possibly $\mathbf{E}_1$ and $\mathbf{E}_2$, it would appear bigger and bigger each time one went round λ and so there would be a p.p. (parallelly propagated) curvature singularity. However in Taub–NUT space it turns out that the vector $\mathbf{E}_3$ can be chosen so that there is only one independent non-zero component of the Riemann tensor, which is $R(\mathbf{E}_3, \mathbf{E}_4, \mathbf{E}_3, \mathbf{E}_4)$. This involves $\mathbf{E}_3$ and $\mathbf{E}_4$ equally, and so has the same value each time round. Since a similar argument will probably hold for any imprisoned curve, it seems there is no p.p. curvature singularity in Taub–NUT space, although this space is singular by our definition. One would like to know whether this kind of behaviour would occur in physically realistic solutions containing matter, or whether Taub–NUT space is an isolated pathological example. This question is important because, as we shall argue in the next chapter, we interpret the preceding theorems as indicating not that geodesic incompleteness necessarily occurs, but that General Relativity breaks down in very strong gravitational fields. Such fields do not occur in the Taub–NUT kind of situation. This conclusion is a result of the very special nature of the Riemann tensor in Taub–NUT space. In general, one would expect some other components of the Riemann tensor to be non-zero on the imprisoned curve, and so there would be a p.p. curvature singularity even though there might be no s.p. curvature singularity. In fact one can prove:

Proposition 8.5.2

If $p \in \mathscr{M}$ is a limit point of a b-incomplete curve λ and if at p, $R_{ab}K^aK^b \neq 0$ for all non-spacelike vectors $\mathbf{K}$, then λ corresponds to

a p.p. curvature singularity. (This condition can be replaced by the condition that there do not exist any null directions K^a such that $K^aK^cC_{abcd}K_{e]} = 0$.)

Let $\mathscr{U}$ be a convex normal coordinate neighbourhood of p with compact closure, and let $\{\mathbf{Y}_i\}$, $\{\mathbf{Y}^i\}$ be a field of dual orthonormal bases on $\mathscr{U}$. Let $\{\mathbf{E}_a\}$, $\{\mathbf{E}^a\}$ be a parallelly propagated dual orthonormal basis on the curve $\lambda(t)$. Let $\tilde{t}$ be a parameter on λ such that in $\mathscr{U}$,

$$\mathrm{d}\tilde{t}/\mathrm{d}t = (\sum_i X^iX^i)^{\frac{1}{2}},$$

where X^i are the components of the tangent vector $\partial/\partial t$ in the basis $\{\mathbf{Y}_i\}$. Then $\tilde{t}$ measures arc-length in the positive definite metric on $\mathscr{U}$ in which the bases $\{\mathbf{Y}_i\}$, $\{\mathbf{Y}^i\}$ are orthonormal.

Since $R_{ab}K^aK^b \neq 0$ at p for any non-spacelike vector K^a, there is a neighbourhood $\mathscr{V} \subset \mathscr{U}$ such that $R_{ab} = CZ_aZ_b + R'_{ab}$, where $C \neq 0$ is a constant, Z_a is a unit timelike vector, and R'_{ab} is such that $CR'_{ab}K^aK^b > 0$ for any non-spacelike vector K^a. Suppose that after some value $\tilde{t}_0$ of $\tilde{t}$ the curve λ intersects $\mathscr{V}$. Since λ has no endpoint and since p is a limit point of λ, the part of λ in $\mathscr{V}$ will have infinite length as measured by $\tilde{t}$. However, the generalized affine parameter is given by

$$\mathrm{d}u/\mathrm{d}\tilde{t} = \{\sum_a (E^a{}_i\tilde{X}^i)^2\}^{\frac{1}{2}},$$

where $\tilde{X}^i$ are the components of the tangent vector $(\partial/\partial\tilde{t})_\lambda$, so $\sum_i \tilde{X}^i\tilde{X}^i = 1$, and $E^a{}_i$ are the components of the basis $\{\mathbf{E}^a\}$ in the basis $\{\mathbf{Y}^i\}$. Since u is finite on the curve, the modulus of the column vector $E^a{}_i\tilde{X}^i$ must go to zero, and so the Lorentz transformation represented by the components $E^a{}_i$ must become unboundedly large. Since $\mathbf{Z}$ is a unit timelike vector, the components of $\mathbf{Z}$ in the basis $\{\mathbf{E}_a\}$ will therefore become unboundedly large and hence some component of the Ricci tensor in the basis $\{\mathbf{E}_a\}$ will become unboundedly large. □

This result shows that an observer whose history was a b-incomplete imprisoned non-spacelike curve in a generic space–time would be torn apart by unboundedly large curvature forces in a finite time. However another observer could travel through the same region without experiencing any such effects. An interesting example in this connection is provided by Taub–NUT space in which the metric has been altered by a conformal factor Ω which differs from one only in a small neighbourhood of a point p on the horizon. This conformal transformation would not alter the causal structure of the space and would not affect

the incompleteness of the closed null curve through the point p. However in general $R_{ab}K^aK^b \neq 0$ where K^a is the tangent vector to the closed null geodesic. After each cycle, $R_{ab}K^aK^b$ increases by a factor a^2 and so there is a p.p. curvature singularity. Yet the metric is perfectly regular on a compact neighbourhood of the horizon and so there is no s.p. curvature singularity associated with the incompleteness.

One would like to rule out this kind of situation in which the incomplete curves are totally imprisoned in a compact region. This kind of behaviour might occur in a countably infinite number of different regions of space–time. Thus one cannot describe it by saying that *all* the incomplete curves are totally imprisoned in one compact set. Instead one wants to describe it by saying that a set of incomplete curves which are compact in some sense are totally imprisoned in a compact region of $\mathscr{M}$. To make this concept precise, we define b-boundedness as follows.

We define the space $B(\mathscr{M})$ to be the set of all pairs (λ, u), where u is a point in the bundle of linear frames $L(\mathscr{M})$ and λ is a C^1 curve in $\mathscr{M}$ which has only one endpoint, which is at $\pi(u)$. Let $\mathscr{U}$ be an open set in $\mathscr{M}$ and $\mathscr{V}$ be an open set in $L(\mathscr{M})$. We define the open set $O(\mathscr{U}, \mathscr{V})$ to be the set of all elements of $B(\mathscr{M})$ such that λ intersects $\mathscr{U}$ and $u \in \mathscr{V}$. The sets of the form $O(\mathscr{U}, \mathscr{V})$ for all $\mathscr{U}$, $\mathscr{V}$ form a sub-basis for the topology of $B(\mathscr{M})$. Recall that the map exp: $T(\mathscr{M}) \to \mathscr{M}$ is defined by taking a vector $\mathbf{X}$ at a point p and proceeding along the geodesic from p in the direction of $\mathbf{X}$ a unit distance as measured in the affine parameter defined by $\mathbf{X}$. Similarly we may define a map Exp: $B(\mathscr{M}) \to \mathscr{M}$ by proceeding from $\pi(u)$ along the curve λ a unit distance as measured in the generalized affine parameter on λ defined by u. The map Exp is continuous and will be defined for all of $B(\mathscr{M})$ if $\mathscr{M}$ is b-complete. We shall say that $(\mathscr{M}, \mathbf{g})$ is *b-bounded* if for every compact set $W \subset B(\mathscr{M})$, Exp (W) has a compact closure in $\mathscr{M}$. Since Exp is continuous, $(\mathscr{M}, \mathbf{g})$ is b-bounded if it is b-complete. However, Taub–NUT space is an example which is b-bounded but not b-complete. We shall show that this can be possible only because Taub–NUT space is completely empty. The presence of any matter on the surface $\mathscr{S}$ in theorem 4 will mean that the space is both b-incomplete and b-unbounded.

Theorem 5

Space–time is not b-bounded if conditions (1)–(3) of theorem 4 hold, and

(4) the energy–momentum tensor is non-zero somewhere on $\mathscr{S}$,

(5) the energy–momentum tensor obeys a slightly stronger form of the dominant energy condition (§4.3): if K^a is a non-spacelike vector, then $T^{ab}K_a$ is zero or non-spacelike and $T_{ab}K^aK^b \geqslant 0$, equality holding only if $T^{ab}K_b = 0$.

Remark. Condition (4) could be replaced by the generic condition (see Theorem 2).

Proof. Consider the covering space $\mathscr{M}_G$ (§6.5) defined as the set of all pairs $(p, i[\lambda])$, where λ is a curve from q to p, $p, q \in M$, and $i[\lambda]$ is the number of times λ cuts $\mathscr{S}$ in the future direction minus the number of times it cuts it in the past direction. For each integer a,

$$\mathscr{S}_a \equiv \{(p, i[\lambda]) : p \in \mathscr{S},\ i[\lambda] = a\}$$

is diffeomorphic to $\mathscr{S}$ and is a partial Cauchy surface in $\mathscr{M}_G$. In general $\mathscr{M}_G$ need not be b-bounded if $\mathscr{M}$ is, but in the situation under consideration we have the following result:

Lemma 8.5.3

Let conditions (1)–(3) hold and let $D^+(\mathscr{S}_0)$ not have compact closure in $\mathscr{M}_G$; then if ψ is the covering projection $\psi : \mathscr{M}_G \to \mathscr{M}$, $\psi(D^+(\mathscr{S}_0))$ will not have compact closure in $\mathscr{M}$.

$\mathscr{M}$ is either diffeomorphic to $\mathscr{M}_G$ or to $\mathscr{M}_a$, the portion of $\mathscr{M}_G$ between $\mathscr{S}_a$ and $\mathscr{S}_{a+1}$ with $\mathscr{S}_a$ and $\mathscr{S}_{a+1}$ identified. If for any $a \geqslant 0$, $\mathscr{M}_a \cap D^+(\mathscr{S}_0)$ does not have compact closure in $\mathscr{M}_G$, then $\psi(D^+(\mathscr{S}_0))$ will not have compact closure in $\mathscr{M}$. If however $\mathscr{M}_a \cap D^+(\mathscr{S}_0)$ had compact closure for all $a \geqslant 0$ it would also have to be non-empty for all $a \geqslant 0$ since $\overline{D}^+(\mathscr{S}_0)$ is non-compact. But for $p \in \mathscr{S}_a$, the proper volume of $I^-(p) \cap \mathscr{M}_{a-1}$ has some lower bound c. Thus for every $a \geqslant 0$ the proper volume of $\mathscr{M}_a \cap D^+(\mathscr{S}_0)$ could not be less than c. But this is impossible since by conditions (1)–(3) and proposition 6.7.1, the proper volume of $D^+(\mathscr{S}_0)$ is less than $3/(-\theta_0) \times$ (area of $\mathscr{S}$), where θ_0 is the negative upper bound of $\chi^a{}_a$ on $\mathscr{S}$. □

Using this result, one can prove:

Lemma 8.5.4

If $D^+(\mathscr{S}_0)$ does not have compact closure, $\mathscr{M}$ is not b-bounded.

Let $\mathscr{W}$ be the subset of $B(\mathscr{M}_G)$ consisting of all pairs (λ, u) where λ is any future-inextendible timelike geodesic curve in $\mathscr{M}_G$ orthogonal to

$\mathscr{S}_0$ with endpoint $r \in \mathscr{S}_0$, and $u \in \pi^{-1}(r)$ is any basis at r, one of whose vectors is tangent to λ and of length $-3/\theta_0$, the remaining vectors being an orthonormal basis in $\mathscr{S}_0$.

Let $\{\mathscr{P}_\alpha\}$ be a collection of open sets which cover $\mathscr{W}$. Each $\mathscr{P}_\alpha$ will be the union of finite intersections of sets of the form $O(\mathscr{U}, \mathscr{V})$. It is sufficient to consider the case when the $\mathscr{P}$ can be represented as

$$\mathscr{P}_\alpha = \bigcap_\beta O(\mathscr{U}_{\alpha\beta}, \mathscr{V}_\alpha),$$

where for each α the $\mathscr{U}_{\alpha\beta}$ are a finite number of open sets in $\mathscr{M}_G$, and $\mathscr{V}_\alpha$ is an open set in $L(\mathscr{M}_G)$. Let $(\mu, v) \in \mathscr{W}$. Then there is some α such that $(\mu, v) \in \mathscr{P}_\alpha$. This means that the geodesic μ intersects the open set $\mathscr{U}_{\alpha\beta}$ for each value of β and that $v \in \mathscr{V}_\alpha$. Since geodesics depend continuously on their initial conditions there will be some neighbourhood $\mathscr{Y}_\alpha$ of $\pi(v)$ such that every future-inextendible geodesic through $\mathscr{Y}_\alpha$ orthogonal to $\mathscr{S}_0$ will intersect $\mathscr{U}_{\alpha\beta}$ for each value of β. Let $\mathscr{V}'_\alpha$ be an open set contained in $\mathscr{V}_\alpha$ such that $\pi(\mathscr{V}'_\alpha) \subset \mathscr{Y}_\alpha$. Then

$$(\mu, v) \in O(\pi(\mathscr{V}'_\alpha), \mathscr{V}'_\alpha)$$

is contained in $\mathscr{P}_\alpha$. Thus the sets $\{O(\pi(\mathscr{V}'_\alpha), \mathscr{V}'_\alpha)\}$ form a refinement of the covering $\mathscr{P}_\alpha$.

Consider the subset $\mathscr{Q}$ of $L(\mathscr{M}_G)$ consisting of all bases over $\mathscr{S}_0$ where one of the basis vectors is orthogonal to $\mathscr{S}_0$ and of length $-3/\theta_0$, and the remaining vectors are an orthonormal basis of $\mathscr{S}_0$. Since $\mathscr{Q}$ is compact, it can be covered by a finite number of the sets $\mathscr{V}'_\alpha$. Thus $\mathscr{W}$ is compact since it can be covered by a finite number of the sets $O(\pi(\mathscr{V}'_\alpha), \mathscr{V}'_\alpha)$.

By proposition 6.7.1 each point of $D^+(\mathscr{S}_0)$ lies within a proper distance $-3/\theta_0$ along the future-directed geodesic orthogonal to $\mathscr{S}_0$. This means that $\mathrm{Exp}\,(\mathscr{W})$ contains $D^+(\mathscr{S}_0)$. Let $\psi_*: B(\mathscr{M}_G) \to B(\mathscr{M})$ be the map which takes $(\lambda, u) \in B(\mathscr{M}_G)$ to $(\psi(\lambda), \psi_* u) \in B(\mathscr{M})$. Then $\psi_* \mathscr{W}$ will be a compact subset of $B(\mathscr{M})$ such that

$$\mathrm{Exp}\,(\psi_* W) \supset \psi(D^+(\mathscr{S}_0)).$$

Thus if $\overline{D^+(\mathscr{S}_0)}$ is not compact, $\overline{\psi}(D^+(\mathscr{S}_0))$ is not compact, so $(\mathscr{M}, \mathbf{g})$ is not b-bounded. □

This shows that it is sufficient to prove $\overline{D^+}(\mathscr{S}_0)$ non-compact. Suppose it were compact. Then $H^+(\mathscr{S}_0)$ would also be compact. We show below that this would imply that the divergence of the null geodesic generators would have to be zero everywhere on $H^+(\mathscr{S}_0)$. This would be impossible if the matter density were non zero somewhere on $H^+(\mathscr{S}_0)$.

Lemma 8.5.5

If $H^+(\mathscr{2})$ is compact for a partial Cauchy surface $\mathscr{2}$, then the null geodesic generating segments of $H^+(\mathscr{2})$ are geodesically complete in the past direction.

From proposition 6.5.2 it follows that the generating segments have no past endpoints. They must therefore form 'almost closed' curves in the compact set $H^+(\mathscr{2})$. If they formed actual closed curves, one could use proposition 6.4.4 to show that if they were incomplete in the past direction, they could be varied towards the past to give closed timelike curves. This however would be impossible since such curves would lie in $D^+(\mathscr{2})$. The proof in the case when the null geodesic generators of $H^+(\mathscr{2})$ are only 'almost closed' is similar though a little more delicate.

Introduce a future-directed timelike unit vector field $\mathbf{V}$ which is geodesic in a neighbourhood $\mathscr{U}$ of $H^+(\mathscr{2})$ with compact closure. Define the positive definite metric $\mathbf{g}'$ as in proposition 6.4.4 by

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

and let t be a parameter which measures proper distance in the metric $\mathbf{g}'$ along a null geodesic generating segment γ of $H^+(\mathscr{2})$, and which is zero at some point $q \in \gamma$. Then $g(\mathbf{V}, \partial/\partial t) = -2^{-\frac{1}{2}}$. As γ has no past endpoint, t will have no lower bound. Let f and h be given by

$$f\frac{\partial}{\partial t} = \frac{\mathrm{D}}{\partial t}\left(\frac{\partial}{\partial t}\right), \quad \frac{\partial}{\partial v} = h\frac{\partial}{\partial t},$$

where v is an affine parameter. Suppose γ were geodesically incomplete in the past, then the affine parameter

$$v = \int_0^t h^{-1}\, \mathrm{d}t'$$

would have a lower bound v_0 as $t \to -\infty$. Now consider a variation α of γ whose variation vector $\partial/\partial u$ is equal to $-x\mathbf{V}$. Then

$$\frac{\partial}{\partial u} g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right)\Big|_{u=0} = 2^{-\frac{1}{2}}\left(\frac{\mathrm{d}x}{\mathrm{d}t} + xh^{-1}\frac{\mathrm{d}h}{\mathrm{d}t}\right). \tag{8.3}$$

Since $h \to \infty$ as $t \to -\infty$, one could find a bounded function $x(t)$ such that (8.3) was negative for all $t \leqslant 0$. However this would not be sufficient to ensure that the variation gave an everywhere timelike curve since it could be that the range of u for which (8.3) remained negative

tended to zero as $t \to -\infty$. To deal with this we shall consider the second derivative under the variation:

$$\frac{\partial^2}{\partial u^2} g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right) = \frac{\partial}{\partial u}\left(g\left(\frac{\partial}{\partial t}, \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial u}\right)\right)$$

$$= g\left(\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial u}, \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial u}\right) + g\left(\frac{\partial}{\partial t}, \frac{\mathrm{D}}{\partial t}\frac{\mathrm{D}}{\partial u}\frac{\partial}{\partial u}\right) + g\left(\frac{\partial}{\partial t}, R\left(\frac{\partial}{\partial u}, \frac{\partial}{\partial t}\right)\frac{\partial}{\partial u}\right).$$

Choosing $\partial x/\partial u$ to be zero and using the fact that $\mathbf{V}$ is a geodesic in a neighbourhood $\mathscr{U}$ of $H^+(\mathscr{Q})$ this reduces to

$$-\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 + x^2\left[g\left(\frac{\mathrm{D}\mathbf{V}}{\partial t}, \frac{\mathrm{D}\mathbf{V}}{\partial t}\right) + g\left(\frac{\partial}{\partial t}, R\left(\mathbf{V}, \frac{\partial}{\partial t}\right)\mathbf{V}\right)\right]$$

for $0 \leqslant u \leqslant \epsilon$. In any basis orthonormal with respect to the metric $\mathbf{g}'$, the components of the Riemann tensor and of the covariant derivative of $\mathbf{V}$ (with respect to $\mathbf{g}$) will be bounded on $\mathscr{U}$. Thus there is some $C > 0$ such that

$$\frac{\partial^2}{\partial u^2} g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right) \leqslant C^2 x^2 g'\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right).$$

Now

$$\frac{\partial}{\partial u}\left(g\left(\mathbf{V}, \frac{\partial}{\partial t}\right)\right) = -\frac{\mathrm{d}x}{\mathrm{d}t},$$

so

$$g\left(\mathbf{V}, \frac{\partial}{\partial t}\right) = -2^{-\frac{1}{2}} - u\frac{\mathrm{d}x}{\mathrm{d}t}.$$

Therefore

$$g'\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right) = g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right) + 1 - (2\sqrt{2})\, u\frac{\mathrm{d}x}{\mathrm{d}t} + 2u^2\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2$$

$$\leqslant g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right) + d$$

for $0 \leqslant u \leqslant \epsilon$, where $d = (2\sqrt{2})\,\epsilon C_1 + 2\epsilon^2 C_1^2 + 1$, and C_1 is an upper bound to $|\mathrm{d}x/\mathrm{d}t|$. Thus we have

$$\frac{\partial^2 y}{\partial u^2} \leqslant C^2 x^2 (y + d)$$

and

$$\left.\frac{\partial y}{\partial u}\right|_{u=0} = 2^{-\frac{1}{2}} h^{-1}\frac{\mathrm{d}}{\mathrm{d}t}(hx), \quad y|_{u=0} = 0,$$

where $y = g(\partial/\partial t, \partial/\partial t)$. Therefore

$$y \leqslant d(\cosh Cxu - 1) + a \sinh Cxu$$

$$\leqslant \sinh Cxu(d \tanh \tfrac{1}{2}Cxu + a),$$

where $a = 2^{-\frac{1}{2}} C^{-1}\,\mathrm{d}(\log hx)/\mathrm{d}t$.

Now take
$$x = h^{-1}\left[-\int_t^0 h^{-1}\,\mathrm{d}t' + K\right]^{-1},$$

where
$$K = 2\int_{-\infty}^0 h^{-1}\,\mathrm{d}t';$$

then $a = -2^{-\frac{1}{2}}C^{-1}hx$. Since $f = -h^{-1}(\mathrm{d}h/\mathrm{d}t)$ is bounded on the compact set $H^+(\mathscr{Q})$ and since
$$\int_t^0 h^{-1}\,\mathrm{d}t' = -v$$

was assumed to converge as $t \to -\infty$, there would be upper bounds for x and $|\mathrm{d}x/\mathrm{d}t|$ and a positive lower bound C_2 for h when $-\infty < t \leqslant 0$. Then for $0 < u < \min(\epsilon, 2C^{-2}d^{-1}C_2)$, y would be negative when $-\infty < t \leqslant 0$.

In other words, the variation α would give a past-inextendible timelike curve which lay in $\operatorname{int} D^+(\mathscr{Q})$ and which was totally imprisoned in the compact set $\overline{\mathscr{U}}$. But this is impossible, since by lemma 6.6.5 the strong causality condition holds on $\operatorname{int} D^+(\mathscr{Q})$. Thus γ must be geodesically complete in the past direction. □

Consider the expansion $\hat{\theta}$ of the tangent vectors $\partial/\partial t$ to the null geodesic generators of $H^+(\mathscr{S}_0)$. Suppose that $\hat{\theta} > 0$ at some point q on a generator γ and let $\mathscr{T}$ be a spacelike two-surface through q in a neighbourhood of q in $H^+(\mathscr{S}_0)$. The generators of $H^+(\mathscr{S}_0)$ will be orthogonal to $\mathscr{T}$ and would be converging into the past. Then by condition (1) and the above lemma there would be a point $r \in \gamma$ conjugate to $\mathscr{T}$ along γ (proposition 4.4.6). Points on γ beyond r could be joined to $\mathscr{T}$ by timelike curves (proposition 4.5.14). But this would be impossible since $H^+(\mathscr{S}_0)$ is an achronal set. Therefore $\hat{\theta} \leqslant 0$ on $H^+(\mathscr{S}_0)$.

Now consider the family of differentiable maps β_z: $H^+(\mathscr{S}_0) \to H^+(\mathscr{S}_0)$ defined by taking a point $q \in H^+(\mathscr{S}_0)$ a distance z (measured in the metric $\mathbf{g}'$) to the past along the null geodesic generator through q. Let $\mathrm{d}A$ be the area measured in the metric $\mathbf{g}'$ of a small element of $H^+(\mathscr{S}_0)$. Under the map β_z,
$$\frac{\mathrm{d}}{\mathrm{d}z}\mathrm{d}A = -\hat{\theta}\,\mathrm{d}A.$$

Thus
$$\frac{\mathrm{d}}{\mathrm{d}z}\int_{\beta_z(H^+(\mathscr{S}_0))} \mathrm{d}A = -\int_{\beta_z(H^+(\mathscr{S}_0))} \hat{\theta}\,\mathrm{d}A. \tag{8.4}$$

But β_z maps $H^+(\mathscr{S}_0)$ into $H^+(\mathscr{S}_0)$ (and onto if the generating segments have no future endpoints). Thus (8.4) must be less than or equal to

zero. Together with the previous result this would imply $\hat{\theta} = 0$ on $H^+(\mathscr{S}_0)$. By the propagation equation (4.35) this is possible only if $R_{ab}K^aK^b = 0$ everywhere on $H^+(\mathscr{S}_0)$, where $\mathbf{K}$ is the tangent vector to the null geodesic generator. However by the conservation theorem of §4.3 condition (5) implies that $T_{ab}K^aK^b$ is non-zero somewhere on $H^+(\mathscr{S})$ and by the Einstein equations (with or without Λ), $T_{ab}K^aK^b$ equals $R_{ab}K^aK^b$. (Strictly, the form of the conservation theorem required is slightly different from that in §4.3. Since there are no suitable spacelike surfaces which intersect $H^+(\mathscr{S}_0)$, one uses instead a family of surfaces one of which is $H^+(\mathscr{S}_0)$, the others being spacelike. These surfaces can be defined by taking the value of the function t at the point $p \in \overline{D^+}(\mathscr{S}_0)$ to be minus the proper volume of $J^+(p) \cap D^+(\mathscr{S}_0)$. Since $t_{;a}$ becomes null on $H^+(\mathscr{S}_0)$, it is no longer necessarily true that there is a constant $C > 0$ such that on $\overline{D^+}(\mathscr{S}_0)$,

$$T^{ab}t_{;ab} \leqslant CT^{ab}t_{;a}t_{;b}.$$

However if V^a is a timelike vector field on $\overline{D^+}(\mathscr{S}_0)$, there is a constant C such that

$$T^{ab}t_{;ab} \leqslant CT^{ab}(t_{;a}t_{;b} + t_{;a}V_b)$$

and

$$T^{ab}V_{a;b} \leqslant CT^{ab}(t_{;a}t_{;b} + t_{;a}V_b).$$

One can then proceed as in §4.3 using $T^{ab}(t_{;ab} + V_{a;b})$ in place of $T^{ab}t_{;ab}$, and proving that $T^{ab}(t_{;a}t_{;b} + t_{;a}V_b)$ cannot be zero on $H^+(\mathscr{S}_0)$ if it is non-zero on $\mathscr{S}_0$. The result then follows from (5).) □