Lectures on Lorentzian causality

ESI-EMS-IAMP Summer School on Mathematical Relativity

Gregory J. Galloway
Department of Mathematics
University of Miami

August 4, 2014

Contents

Resources for Causal Theory

References

[1] G. Galloway, *Notes on Lorentzian causality*, ESI-EMS-IAMP Summer School on Mathematical Relativity.

[2] B. O'Neill, *Semi-Riemannian geometry*, Pure and Applied Mathematics, vol. 103, Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1983.

****

[3] J. K. Beem, P. E. Ehrlich, and K. L. Easley, *Global Lorentzian geometry*, second ed., Monographs and Textbooks in Pure and Applied Mathematics, vol. 202, Marcel Dekker Inc., New York, 1996.

[4] S. W. Hawking and G. F. R. Ellis, *The large scale structure of space-time*, Cambridge University Press, London, 1973, Cambridge Monographs on Mathematical Physics, No. 1.

[5] R. Penrose, *Techniques of differential topology in relativity*, Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1972, Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 7.

[6] Robert M. Wald, *General relativity*, University of Chicago Press, Chicago, IL, 1984.

1 Lorentzian manifolds

In General Relativity, the space of events is represented by a *Lorentzian manifold*, which is a smooth manifold M^{n+1} equipped with a metric g of Lorentzian signature. Thus, at each $p \in M$,

$$g : T_pM \times T_pM \to \mathbb{R} \tag{1.1}$$

is a scalar product of signature $(-, +, ..., +)$. With respect to an orthonormal basis $\{e_0, e_1, ..., e_n\}$, as a matrix,

$$[g(e_i, e_j)] = \mathrm{diag}(-1, +1, ..., +1)\,. \tag{1.2}$$

Example: Minkowski space, the spacetime of Special Relativity. Minkowski space is $\mathbb{R}^{n+1}$, equipped with the Minkowski metric η: For vectors $X = X^i \frac{\partial}{\partial x^i}$, $Y = Y^i \frac{\partial}{\partial x^i}$ at p, (where x^i are standard Cartesian coordinates on $\mathbb{R}^{n+1}$),

$$\eta(X, Y) = -X^0Y^0 + \sum_{i=1}^{n} X^iY^i\,. \tag{1.3}$$

Similarly, for the Lorentzian metric g, we have for vectors $X = X^ie_i$, $Y = Y^je_j$ at p,

$$g(X, Y) = X^iY^jg(e_i, e_j) = -X^0Y^0 + \sum_{i=1}^{n} X^iY^i\,. \tag{1.4}$$

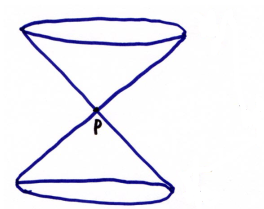
Thus, each tangent space of a Lorentzian manifold is isometric to Minkowski space. Hence, one may say that Lorentzian manifolds are locally modeled on Minkowski space, just as Riemannian manifolds are locally modeled on Euclidean space.

1.1 Causal character of vectors.

At each point, vectors fall into three classes, as follows:

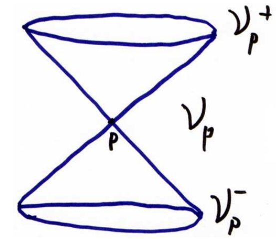
$$X \text{ is } \begin{cases} \text{timelike} & \text{if } g(X,X) < 0 \\ \text{null} & \text{if } g(X,X) = 0 \\ \text{spacelike} & \text{if } g(X,X) > 0\ . \end{cases}$$

We see that the set of null vectors $X \in T_pM$ forms a double cone $\mathcal{V}_p$ in the tangent space T_pM:



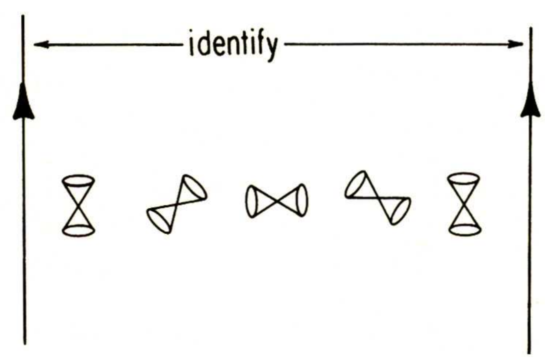
called the null cone (or light cone) at p. Timelike vectors point inside the null cone and spacelike vectors point outside.

Time orientability. Consider at each point of p in a Lorentzian manifold M the null cone $\mathcal{V}_p \subset T_pM$. $\mathcal{V}_p$ is a double cone consisting of two cones, $\mathcal{V}_p^+$ and $\mathcal{V}_p^-$:



We may designate one of the cones, $\mathcal{V}_p^+$, say, as the *future* null cone at p, and the other half cone, $\mathcal{V}_p^-$, as the *past* null cone at p. If this assignment can made in a continuous manner over all of M (this can always be done locally) then we say that M is *time-orientable.*

The following figure illustrates a Lorentzian manifold that is *not* time-orientable (even though the underlying manifold is orientable).



There are various ways to make the phrase "continuous assignment" precise (see e.g., [10, p. 145]), but they all result in the following, which we adopt as the definition of of time-orientability.

Definition 1.1. *A Lorentzian manifold M^{n+1} is time-orientable iff it admits a smooth timelike vector field T.*

If M is time-orientable, the choice of a smooth time-like vector field T fixes a time orientation on M:

For any $p \in M$, a (nonzero) causal (timelike or null) vector $X \in T_pM$ is

(1) future directed provided $g(X,T) < 0$, and

(2) past directed provided $g(X,T) > 0$.

Thus X is future directed if it points into the same half cone at p as T. (We remark that if M is not time-orientable, it admits a double cover that is.)

By a **spacetime** we mean a connected time-oriented Lorentzian manifold (M^{n+1}, g). We will usually restrict attention to spacetimes.

Lorentzian inequalities. $X \in T_pM$ is *causal* if it is timelike or null, $g(X,X) \le 0$. If X is causal, define its length as

$$|X| = \sqrt{-g(X,X)}\,.$$

Proposition 1.1. *The following basic inequalities hold.*

(1) *(Reverse Schwarz inequality) For all causal vectors $X, Y \in T_pM$,*

$$|g(X,Y)| \ge |X||Y| \tag{1.5}$$

(2) *(Reverse triangle inequality) For all causal vectors X, Y that point into the same half cone of the null cone at p,*

$$|X+Y| \ge |X| + |Y|\,. \tag{1.6}$$

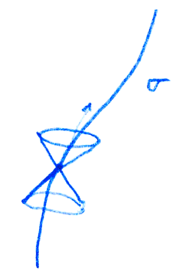
Proof hints: Note (1.5) trivially holds if X is null. For X timelike, decompose Y as $Y = \lambda X + Y^\perp$, where $Y^\perp$ (the component of Y perpendicular to X) is necesarily spacelike. Inequality (1.6) follows easily from (1.5).

The reverse triangle inequality is the source of the twin paradox.

1.2 Causal character of curves:

Let $\gamma : I \to M$, $t \to \gamma(t)$ be a smooth curve in M.

γ is said to be *timelike* provided $\gamma'(t)$ is timelike for all $t \in I$.



In GR, a timelike curve corresponds to the history (or *worldline*) of an observer.

Null curves and *spacelike curves* are defined analogously.

A *causal curve* is a curve whose tangent is either timelike or null at each point. Heuristically, in accordance with relativity, information flows along causal curves, and so such curves are the focus of our attention.

The notion of a causal curve extends in a natural way to piecewise smooth curves; require when two segments join, the end point tangent vectors must point into the same half cone of the null cone $\mathcal{V}_p$ at p.

The length of a causal curve $\gamma : [a, b] \to M$, is defined by

$$L(\gamma) = \text{Length of } \gamma = \int_a^b |\gamma'(t)| dt = \int_a^b \sqrt{-\langle \gamma'(t), \gamma'(t) \rangle}\, dt\,.$$

If γ is timelike one can introduce arc length parameter along γ. In general relativity, the arc length parameter along a timelike curve is called proper time, and corresponds to time kept by the observer.

1.3 The Levi-Civita connection and geodesics.

Recall that a Lorentzian manifold M (like any pseudo-Riemannian manifold) admits a unique covariant derivative operator ∇ called the *Levi-Civita connection*. Thus for smooth vector fields X, Y on M, $\nabla_X Y$ is a vector field on M (the directional derivative of Y in the direction X) satisfying:

(1) $\nabla_X Y$ is linear in Y over the reals.

(2) $\nabla_X Y$ is linear in Y over the space of smooth functions. (In particular, $\nabla_{fX} Y = f\nabla_X Y$).

(3) (Product rule) $\nabla_X fY = X(f)Y + f\nabla_X Y$.

(4) (Symmetric) $[X, Y] = \nabla_X Y - \nabla_Y X$.

(5) (Metric product rule) $X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle$.

∇ is uniquely determined by these properties. With respect to a coordinate neighborhood (U, x^i), one has,

$$\nabla_X Y = (X(Y^k) + \Gamma^k_{ij} X^i Y^j)\partial_k \,, \tag{1.7}$$

where $\partial_k = \frac{\partial}{\partial x^k}$, $X = X^i \partial_i$, $Y = Y^j \partial_j$, and the Γ^k_{ij}'s are the Christoffel symbols,

$$\Gamma^k_{ij} = \frac{1}{2} g^{km} \left(g_{jm,i} + g_{im,j} - g_{ij,m} \right) \,,$$

where $g_{ij} = g(\partial_i, \partial_j)$, etc.

We see from the coordinate expression in (1.7) that $\nabla_X Y$ depends only on the value of X at a point and only on the values of Y along a curve, defined in neighborhood of the point, having X as a tangent vector.

Thus the Levi-Civita connection enables one to compute the covariant derivative of a vector field $t \xrightarrow{Y} Y(t) \in T_{\gamma(t)}M$ defined along a curve $\gamma : I \to M$, $t \to \gamma(t)$. In local coordinates $\gamma(t) = (x^1(t), ..., x^n(t))$, and from (1.7) we have

$$\nabla_{\gamma'} Y = \left(\frac{dY^k}{dt} + \Gamma^k_{ij} \frac{dx^i}{dt} Y^j \right) \partial_k \,. \tag{1.8}$$

where $\gamma' = \frac{dx^i}{dt} \partial_i|_\gamma$ is the tangent (or velocity) vector field along γ and $Y(t) = Y^i(t)\partial_i|_{\gamma(t)}$.

Geodesics. Given a curve $t \to \gamma(t)$ in M, $\nabla_{\gamma'}\gamma'$ is called the *covariant acceleration* of γ. In local coordinates,

$$\nabla_{\gamma'}\gamma' = \left(\frac{d^2x^k}{dt^2} + \Gamma^k_{ij}\frac{dx^i}{dt}\frac{dx^j}{dt}\right)\partial_k\,, \tag{1.9}$$

as follows by setting $Y^k = \frac{dx^k}{dt}$ in Equation (1.8).

By definition, a *geodesic* is a curve of zero covariant acceleration,

$$\nabla_{\gamma'}\gamma' = 0 \qquad \text{(Geodesic equation)} \tag{1.10}$$

In local coordinates the geodesic equation becomes a system of $n+1$ second order ODE's in the coordinate functions $x^i = x^i(t)$,

$$\frac{d^2x^k}{dt^2} + \Gamma^k_{ij}\frac{dx^i}{dt}\frac{dx^j}{dt} = 0\,, \qquad k = 0, ..., n\,. \tag{1.11}$$

The basic existence and uniqueness result for systems of ODE's guarantees the following.

Proposition 1.2. *Given $p \in M$ and $v \in T_pM$, there exists an interval I about $t = 0$ and a unique geodesic $\sigma : I \to M$, $t \to \sigma(t)$, satisfying,*

$$\sigma(0) = p\,, \qquad \frac{d\sigma}{dt}(0) = v\,.$$

Convex neighborhoods: By a more refined analysis it can be shown that each $p \in M$ is contained in a *(geodesically) convex* neighborhood U, which has the property that any two points in U can be joined by a unique geodesic contained in U. In fact U can be chosen so as to be a normal neighborhood of each of its points; cf. [10], p. 129. (Recall, a normal neighborhood of $p \in M$ is the diffeomorphic image under the exponential map of a star-shaped domain about $\mathbf{0} \in T_pM$.)

Finally, note if γ is a geodesic then by the metric product rule

$$\gamma'(g(\gamma', \gamma')) = 2g(\nabla_{\gamma'}\gamma', \gamma') = 0\,,$$

and hence geodesics are always constant speed curves. In GR timelike geodesics correspond to *freely falling* observers and null geodesics correspond to the paths of photons.

1.4 Local Lorentz geometry.

In Minkowski space the geodesics are straight lines (the Christoffel symbols vanish in Cartesian coordinates). Moreover the following holds:

(1) If there is a timelike curve γ from p to q then $\overline{pq}$ is timelike.

(2) $L(\overline{pq}) \geq L(\gamma)$, for all causal curves γ from p to q.

Although it can be very different in the large, *locally* the geometry and causality of a Lorentzian manifold is similar to Minkowski space.

Let U be a convex neighborhood in a Lorentzian manifold. Hence for each pair of points $p, q \in U$ there exists a unique geodesic segment from p to q in U, which we denote by $\overline{pq}$.

Proposition 1.3 ([10], p. 146). *Let U be a convex neighborhood in a Lorentzian manifold M^{n+1}.*

(1) If there is a timelike (resp., causal) curve in U from p to q then $\overline{pq}$ is timelike (causal).

(2) If $\overline{pq}$ is timelike then $L(\overline{pq}) \geq L(\gamma)$ for all causal curves γ in U from p to q. Moreover, the inequality is strict unless, when suitable parametrized, $\gamma = \overline{pq}$.

Thus, within a convex neighborhood U, timelike geodesics are *maximal*, i.e., are causal curves of greatest length.) Moreover, within U null geodesics are *achronal*, i.e., no two points can be joined by a timlike curve. Both of these features can fail in the large.

1.5 Curvature and the Einstein equations

The Riemann curvature tensor of (M, g) is defined in terms of second covariant derivatives anti-symmetrized: For vector fields $X, Y, Z \in \mathfrak{X}(M)$,

$$R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z\,. \tag{1.12}$$

The components ${R^l}_{kij}$ of the Riemann curvature tensor in a coordinate chart (U, x^i) are defined by the following equation,

$$R(\partial_i, \partial_j)\partial_k = {R^l}_{kij}\partial_l$$

The Ricci tensor is obtained by contraction,

$$R_{ij} = {R^l}_{ilj}$$

Symmetries of the Riemann curvature tensor imply that the Ricci tensor is symmetric, $R_{ij} = R_{ji}$. By tracing the Ricci tensor, we obtain the scalar curvature,

$$R = g^{ij}R_{ij}\ .$$

The Einstein equations, the field equations of GR, are given by:

$$R_{ij} - \frac{1}{2}Rg_{ij} = 8\pi T_{ij}\ ,$$

where T_{ij} is the energy-momentum tensor.

The Einstein equations describe how spacetime curves in the presence of matter, and it is this curvature that is responsible for the effects of gravity. The left hand side is a purely geometric tensor, the Einstein tensor. On the right hand side is the is the energy momentum tensor T, which describes the energy-momentum content of matter and all other nongravitational fields.

- From the PDE point of view, the Einstein equations form a system of second order quasi-linear equations for the g_{ij}'s. This system may be viewed as a (highly complicated!) generalization of Poisson's equation in Newtonian gravity.

- The *vacuum* Einstein equations are obtained by setting $T_{ij} = 0$. It is easily seen that this equivalent to setting $R_{ij} = 0$.

We will sometimes require that a spacetime satisfying the Einstein equations, obeys an *energy condition*.

- The *null energy condition* (NEC) is the requirement that

$$T(X,X) = \sum_{i,j} T_{ij} X^i X^j \geq 0 \quad \text{for all null vectors } X\,. \tag{1.13}$$

- The stronger *dominant energy condtion* (DEC) is the requirement,

$$T(X,Y) = \sum_{i,j} T_{ij} X^i Y^j \geq 0 \quad \text{for all future directed causal vectors } X, Y\,. \tag{1.14}$$

  The DEC is satisfied by most classical fields. Physically, the DEC requires that the speed of energy flow is less than the speed of light.

2 Futures and pasts

We begin the study of causal theory in earnest. Causal theory is the study of the causal relations '$\ll$' and '$<$'.

Let (M, g) be a spacetime. A timelike (resp. causal) curve $\gamma : I \to M$ is said to be *future directed* provided each tangent vector $\gamma'(t)$, $t \in I$, is future directed. (*Past-directed* timelike and causal curves are defined in a time-dual manner.)

Definition 2.1. *For $p, q \in M$,*

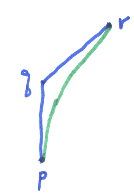
(1) $p \ll q$ means there exists a future directed timelike curve in M from p to q (we say that q is in the timelike future of p),

(2) $p < q$ means there exists a future directed (nontrivial) causal curve in M from p to q (we say that q is in the causal future of p),

We shall use the notation $p \leq q$ to mean $p = q$ or $p < q$.

The causal relations $\ll$ and $<$ are clearly transitive. Also, from variational considerations, it is heuristically clear that the following holds,

$$\text{if } p \ll q \text{ and } q < r \text{ or if } p < q \text{ and } q \ll r \text{ then } p \ll r\,.$$



The above statement is a consequence of the following fundamental causality result; see [10, p. 294] for a careful proof.

Proposition 2.1. *In a spacetime M, if q is in the causal future of p ($p < q$) but is not in the timelike future of p ($p \not\ll q$) then any future directed causal curve γ from p to q must be a null geodesic (when suitably parameterized).*

Definition 2.2. *Given any point p in a spacetime M, the* timelike future *and* causal future *of p, denoted $I^+(p)$ and $J^+(p)$, respectively, are defined as,*

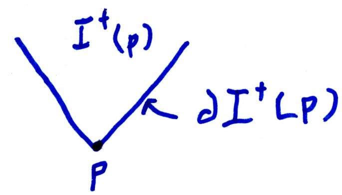
$$I^+(p) = \{q \in M : p \ll q\} \quad \textit{and} \quad J^+(p) = \{q \in M : p \leq q\}\,.$$

Hence, $I^+(p)$ consists of all points in M that can be reached from p by a future directed timelike curve, and $J^+(p)$ consists of the point p and all points in M that can be reached from p by a future directed causal curve.

The timelike and causal *pasts* of p, $I^-(p)$ and $J^-(p)$, respectively, are defined in a time-dual manner in terms of past directed timelike and causal curves.

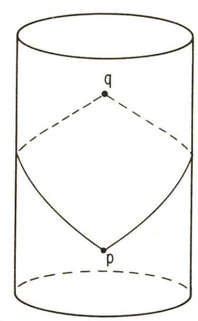
Note by Proposition 2.1, if $q \in J^+(p) \setminus I^+(p)$ ($q \neq p$) then there exists a future directed null geodesic from p to q.

Ex. Minkowski space. For p any point in Minkowski space, $I^+(p)$ is open, $J^+(p)$ is closed and $\partial I^+(p) = J^+(p) \setminus I^+(p)$ is just the future null cone at p. $I^+(p)$ consists of all points inside the future null cone, and $J^+(p)$ consists of all points on and inside the future null cone.



We note, however, that curvature and topology can drastically change the structure of 'null cones' in spacetime.

Ex. Consider the following example of a flat spacetime cylinder, closed in space.



For any point p, $\partial I^+(p)$, is compact and consists of the two future directed null geodesic segments emanating from p that meet to the future at a point q.

In some situations it is convenient to restrict the causal relations $\ll$ and $<$ to open subsets U of a spacetime M. For example,

$$I^+(p,U) = \{q \in U : \exists \text{ future directed timelike curve } \textit{within}\, U \text{ from } p \text{ to } q\}\,.$$

Note that, in general $I^+(p,U) \neq I^+(p) \cap U$.

In general the sets $I^+(p)$ in a spacetime M are open. This is heuristically rather clear: A sufficiently small smooth perturbation of a timelike curve is still timelike. A rigorous proof is based on the causality of convex neighborhoods.

Proposition 2.2. *Let U be a convex neighborhood in a spacetime M. Then, for each $p \in U$,*

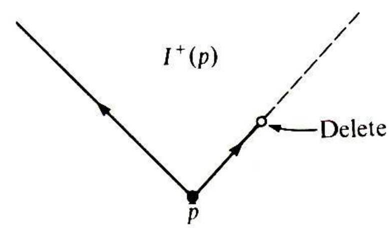
(1) $I^+(p,U)$ is open in U (and hence M),

(2) $J^+(p,U)$ is the closure in U of $I^+(p,U)$.

This proposition follows essentially from part (1) of Proposition 1.3.

Exercise: Prove that for each p in a spacetime M, $I^+(p)$ is open.

In general, sets of the form $J^+(p)$ need not be closed: This can be seen by removing a point from Minkowski space, as illustrated in the figure below.



For any subset $S \subset M$, we define the timelike and causal future of S, $I^+(S)$ and $J^+(S)$, respectively by

$$I^+(S) = \bigcup_{p \in S} I^+(p) \quad \text{and} \quad J^+(S) = \bigcup_{p \in S} J^+(p)\,.$$

Note:

(1) $S \subset J^+(S)$.

(2) $I^+(S)$ is open (union of open sets).

$I^-(S)$ and $J^-(S)$ are defined in a time-dual manner.

Although in general $J^+(S) \neq \overline{I^+(S)}$, the following relationships always hold between $I^+(S)$ and $J^+(S)$.

Proposition 2.3. *For all subsets $S \subset M$,*

(1) $\operatorname{int} J^+(S) = I^+(S)$*,*

(2) $J^+(S) \subset \overline{I^+(S)}$*.*

Proof. Exercise.

3 Achronal boundaries

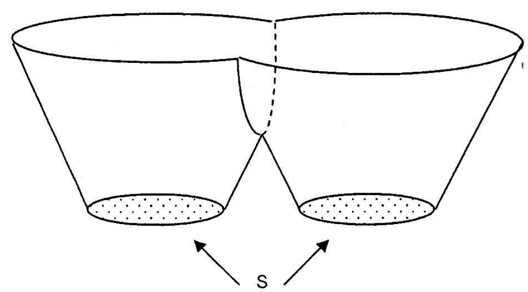
Achronal sets play an important role in causal theory.

Definition 3.1. *A subset $S \subset M$ is* **achronal** *provided no two of its points can be joined by a timelike curve.*

Of particular importance are *achronal boundaries.*

Definition 3.2. *An* **achronal boundary** *is a set of the form $\partial I^+(S)$ (or $\partial I^-(S)$), for some $S \subset M$.*

The following figure illustrates some of the important structural properties of achronal boundaries. S is the disjoint union of two spacelike disks; the achronal boundary $\partial I^+(S)$ consists of S and the merging of two future light cones.



Proposition 3.1. *An achronal boundary $\partial I^+(S)$, if nonempty, is a closed achronal C^0 hypersurface in M.*

We discuss the proof of this proposition, beginning with the following basic lemma.

Lemma 3.2. *If $p \in \partial I^+(S)$ then $I^+(p) \subset I^+(S)$, and $I^-(p) \subset M \setminus \overline{I^+(S)}$.*

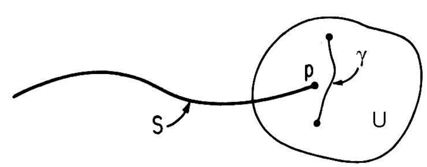
Proof. To prove the first part of the lemma, note that if $q \in I^+(p)$ then $p \in I^-(q)$, and hence $I^-(q)$ is a neighborhood of p. Since p is on the boundary of $I^+(S)$, it follows that $I^-(q) \cap I^+(S) \neq \emptyset$, and hence $q \in I^+(S)$. The second part of the lemma, which can be proved similarly, is left as an exercise. $\square$

Claim 1: An achronal boundary $\partial I^+(S)$ is achronal.

Proof of the claim: Suppose there exist $p, q \in \partial I^+(S)$, with $q \in I^+(p)$. By Lemma 3.2, $q \in I^+(S)$. But since $I^+(S)$ is open, $I^+(S) \cap \partial I^+(S) = \emptyset$. Thus, $\partial I^+(S)$ is achronal. $\square$

Lemma 3.2 also implies that achronal boundaries are *edgeless*. We need to introduce the edge concept.

Definition 3.3. *Let $S \subset M$ be achronal. Then $p \in \overline{S}$ is an edge point of S provided every neighborhood U of p contains a timelike curve γ from $I^-(p, U)$ to $I^+(p, U)$ that does* not *meet S (see the figure).*



We denote by edge S the set of edge points of S. We note that

$$\overline{S} \setminus S \subset \operatorname{edge} S \subset \overline{S}$$

(exercise). If $\operatorname{edge} S = \emptyset$ we say that S is *edgeless*.

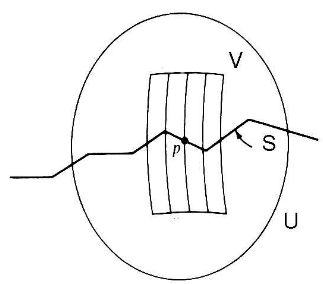
Claim 2: An achronal boundary $\partial I^+(S)$ is edgeless.

Proof of the claim: Lemma 3.2 implies that for any $p \in \partial I^+(S)$, any timelike curve from $I^-(p)$ to $I^+(p)$ must meet $\partial I^+(S)$. It follows that $\partial I^+(S)$ is edgless. $\square$

Proposition 3.1 now follows from the following basic result.

Proposition 3.3. *Let S be achronal. Then $S \setminus \operatorname{edge} S$, if nonempty, is a C^0 hypersurface in M. In particular, an edgeless achronal set is a C^0 hypersurface in M.*

Proof. We sketch the proof; for details, see [10, p. 413]. It suffices to show that in a neighborhood of each $p \in S \setminus \operatorname{edge} S$, $S \setminus \operatorname{edge} S$ can be expressed as a C^0 graph over a smooth hypersurface.

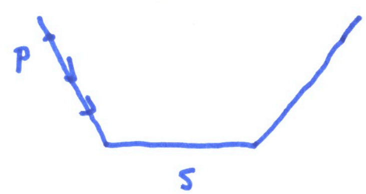


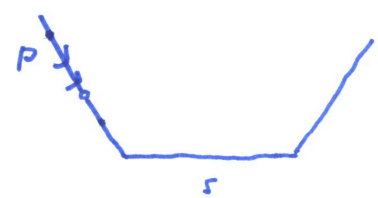
Fix $p \in S \setminus \operatorname{edge} S$. Since p is not an edge point there exists a neighborhood U of p such that every timelike curve from $I^-(p, U)$ to $I^+(p, U)$ meets S.

Let X be a future directed timelike vector field on M, and let $\mathcal{N}$ be a smooth hypersurface in U transverse to X near p. Then, by choosing $\mathcal{N}$ small enough, each integral curve of X through $\mathcal{N}$ will meet S, and meet it exactly once since S is achronal. Using the flow generated by X, it follows that there is a neighborhood $V \approx (t_1, t_2) \times \mathcal{N}$ of p such that $S \cap V$ is given as the graph of a function $t = h(x)$, $x \in \mathcal{N}$ One can now show that a discontinuity of h at some point $x_0 \in \mathcal{N}$ leads to an achronality violation of S. Hence h must be continuous. $\square$

The next result shows that, in general, large portions of achronal boundaries are ruled by null geodesics. A future (resp., past) directed causal curve $\gamma : (a, b) \to M$ is said to be **future (resp., past) inextendible** in M if $\lim_{t \to b^-} \gamma(t)$ does not exist.

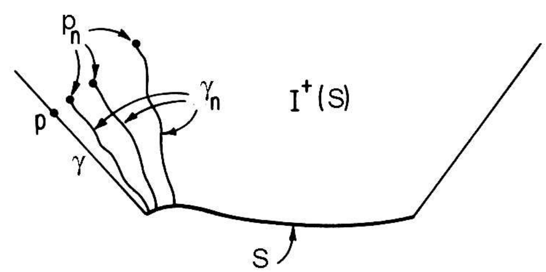
Proposition 3.4. *Let $S \subset M$ be closed. Then each $p \in \partial I^+(S) \setminus S$ lies on a null geodesic contained in $\partial I^+(S)$, which either has a past end point on S, or else is past inextendible in M.*





Proof. The proof uses an important tool in causal theory, namely that of taking a limit of causal curves. This is treated carefully in the notes. Here we just want to get the main idea across. For presentation purposes we are changing the proof a little.

- Fix $p \in \partial I^+(S) \setminus S$. Since $p \in \partial I^+(S)$, there exists a sequence of points $p_n \in I^+(S)$, such that $p_n \to p$. For each n, let γ_n be a past directed timelike curve from p_n to $q_n \in S$



- Let $\hat{M} = M \setminus S$. Let $\hat{\gamma}_n = \gamma_n \setminus \{q_n\}$; $\hat{\gamma}_n$ is past inextendible in $\hat{M}$. Let h be a complete Riemannian metric on $\hat{M}$, and parametrize $\hat{\gamma}_n$ with respect to h-arc length; then have $\hat{\gamma}_n : [0, \infty) \to \hat{M}$. By the Limit Curve Lemma (Lemma 3.5, p. 13 of the notes), there exists a subsequence $\{\hat{\gamma}_m\}$ that converges uniformly on compact subsets to a past inextendible causal curve $\gamma : [0, \infty) \to \hat{M}$.

NB: γ may not be a smooth causal curve - in general it will be C^0 causal curve (see the notes); C^0 causal curves can be approximated by piecewise smooth causal curves to arbitrary precision. WLOG, can assume γ is piecewise smooth causal curve.

- $\gamma \subset \partial I^+(S)$: Clearly $\gamma \subset \overline{I^+(S)}$. Since $p \in \partial I^+(S)$, $\gamma \subset \partial I^+(S)$. (Otherwise if γ entered $I^+(S)$ then p would be in $I^+(S)$, contradicting p being on the boundary.)
- γ is a null geodesic: Consider $q \in \gamma$; $q \in J^-(p) \setminus I^-(p) \implies$ segment of γ from p to q is a null geodesic.
- γ past inextendible in $\hat{M} \implies \gamma$ is either past inextendible in M or has past end point on S (this is treated carefully in the notes).

Achronal boundaries will play an important role in out treatment of the Penrose singularity theorem.

(Achronal boundaries have been recently employed in a fundamental way to study Lorentzian splitting problems, cf. [5].)

4 Causality conditions

A number of results in Lorentzian geometry and general relativity require some sort of causality condition. It is perhaps natural on physical grounds to rule out the occurrence of closed timelike curves. Physically, the existence of such a curve signifies the existence of an observer who is able to travel into his/her own past, which leads to variety of paradoxical situations.

Chronology condition: A spacetime M satisfies the *chronology condition* provided there are no closed timelike curves in M.

Compact spacetimes have limited interest in general relativity since they all violate the chronology condition.

Proposition 4.1. *Every compact spacetime contains a closed timelike curve.*

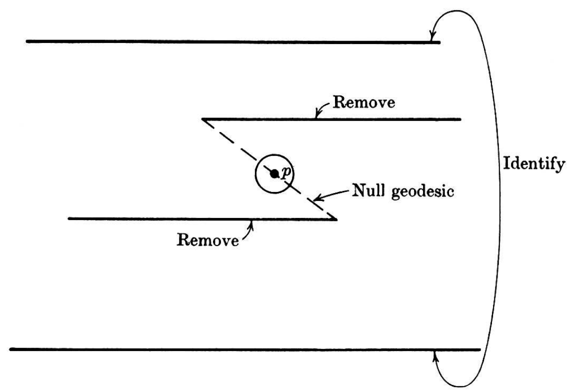
Proof. The sets $\{I^+(p); p \in M\}$ form an open cover of M from which we can abstract a finite subcover: $I^+(p_1)$, $I^+(p_2)$, ..., $I^+(p_k)$. We may assume that this is the minimal number of such sets covering M. Since these sets cover M, $p_1 \in I^+(p_i)$ for some i. It follows that $I^+(p_1) \subset I^+(p_i)$. Hence, if $i \neq 1$, we could reduce the number of sets in the cover. Thus, $p_1 \in I^+(p_1)$ which implies that there is a closed timelike curve through p_1. □

A somewhat stronger condition than the chronology condition is the

Causality condition: A spacetime M satisfies the causality condition provided there are no closed (nontrivial) causal curves in M.

Exercise: Construct a spacetime that satisfies the chronology condition but not the causality condition.

A spacetime that satisfies the causality condition can nontheless be on the verge of failing it, in the sense that there exist causal curves that are "almost closed", as illustrated by the following figure.



Strong causality is a condition that rules out almost closed causal curves.

Definition 4.1. *An open set U in spacetime M is said to be causally convex provided no causal curve in M meets U in a disconnected set.*

Note the neighborhood of p above is *not* causally convex.

Definition 4.2. *Given $p \in M$, strong causality is said to hold at p provided p has arbitrarily small causally convex neighborhoods, i.e., for each neighborhood V of p there exists a causally convex neighborhood U of p such that $U \subset V$.*

Note that strong causality fails at the point p in the figure above. In fact strong causality fails at all points along the dashed null geodesic. It can be shown that the set of points at which strong causality holds is open.

Strong causality condition: A spacetime M is said to be strongly causal if strong causality holds at all of its points.

This is the "standard" causality condition in spacetime geometry, and, although there are even stronger causality conditions, it is sufficient for most applications. (There is an interesting connection between strong causality and the so-called *Alexandrov topology*; see the notes.)

The following lemma is often useful.

Lemma 4.2. *Suppose strong causality holds at each point of a compact set K in a spacetime M. If $\gamma : [0, b) \to M$ is a future inextendible causal curve that starts in K then eventually it leaves K and does not return, i.e., there exists $t_0 \in [0, b)$ such that $\gamma(t) \notin K$ for all $t \in [t_0, b)$.*

Proof. Exercise.

In referring to the property described by this lemma, we say that a future inextendible causal curve cannot be "imprisoned" or "partially imprisoned" in a compact set on which strong causality holds.

4.1 Global hyperbolicity

We now come to a fundamental condition in spacetime geometry, that of global hyperbolicity. Mathematically, global hyperbolicity is a basic 'niceness' condition that often plays a role analogous to geodesic completeness in Riemannian geometry. Physically, global hyperbolicity is closely connected to the issue of classical determinism and the strong cosmic censorship conjecture (that, generically, spacetime solutions to the Einstein equations do not admit *naked singularities*.

Definition 4.3. *A spacetime M is said to be globally hyperbolic provided*

(1) M is strongly causal.

(2) (Internal Compactness) The sets $J^+(p) \cap J^-(q)$ are compact for all $p, q \in M$.

Condition (2) says roughly that M has no holes or gaps.

Ex. Minkowski space is globally hyperbolic but the spacetime obtained by removing one point from it is not.

We consider a few basic consequences of global hyperbolicity.

Proposition 4.3. *Let M be a globally hyperbolic spacetime. Then,*

(1) The sets $J^{\pm}(A)$ are closed, for all compact $A \subset M$.

(2) The sets $J^{+}(A) \cap J^{-}(B)$ are compact, for all compact $A, B \subset M$.

Proof. The following argument shows that $J^{\pm}(p)$ are *closed* for all $p \in$M. The rest we leave as an exercise.

Suppose $q \in \overline{J^{+}(p)} \setminus J^{+}(p)$ for some $p \in M$. Choose $r \in I^{+}(q)$, and $\{q_n\} \subset J^{+}(p)$, with $q_n \to q$. Since $I^{-}(r)$ is an open neighborhood of q, $\{q_n\} \subset J^{-}(r)$ for n large. It follows that $q \in \overline{J^{+}(p) \cap J^{-}(r)} = J^{+}(p) \cap J^{-}(r)$, since $J^{+}(p) \cap J^{-}(r)$ is compact and hence closed. But this contradicts $q \notin J^{+}(p)$. Thus, $J^{+}(p)$ is closed, and similarly so is $J^{-}(p)$. □

Analogously to the case of Riemannian geometry, one can learn much about the global structure of spacetime by studying its causal geodesics.

Basic question: Given $q \in I^{+}(p)$ under what circumstances does there exist a maximal future directed timelike geodesic γ from p to q? *Maximal means:* $L(\gamma) \geq L(\sigma)$ for all future directed causal curves σ from p to q.

Maximality can be conveniently expressed in terms of the Lorentzian *distance function*, $d : M \times M \to [0, \infty]$. For $p < q$, let $\Omega_{p,q}$ denote the collection of future directed causal curves from p to q. Then, for any $p, q \in M$, define

$$d(p,q) = \begin{cases} \sup\{L(\sigma) : \sigma \in \Omega_{p,q}\}, & \text{if } p < q \\ 0, & \text{if } p \not< q \end{cases}$$

While the Lorentzian distance function is not a distance function in the usual sense of metric spaces, and may not even be finite valued, it does have a few nice properties. For one, it obeys a *reverse triangle inequality*,

$$\text{if } p < r < q \quad \text{then} \quad d(p,q) \geq d(p,r) + d(r,q)\,.$$

<u>Exercise:</u> Prove this.

We have the following basic fact.

Proposition 4.4. *The Lorentzian distance function is lower semi-continuous.*

Proof. See the notes.

Though the Lorentzian distance function is not continuous in general, it is continuous (and finite valued) for globally hyperbolic spacetimes; cf., [10, p. 412].

Note that a timelike geodesic from p to $q \in I^+(p)$ is maximal provided

$$L(\gamma) = d(p,q)\,.$$

Global hyperbolicity is the standard condition to ensure the existence of maximal causal geodesic segments.

Proposition 4.5. *Let M be a globally hyperbolic spacetime. Given $p, q \in M$, with $q \in I^+(p)$, there exists a maximal future directed timelike geodesic γ from p to q ($L(\gamma) = d(p,q)$).*

Proof. We sketch the idea. The proof involves a standard limit curve argument, together with the fact that the Lorentzian arc length functional is upper semi-continuous; see [12, p. 54].

- For each n, let γ_n be a future directed timelike curve from p to q, parameterized with respect to h-arc length, such that $L(\gamma_n) \to d(p,q)$.

- Each γ_n is contained in the *compact set* $J^+(p) \cap J^-(q)$. Together with strong causality, the limit curve lemma can be used to show there exists a subsequence γ_m that converges to a C^0 causal curve γ from p to q.

- By usc of L,
$$L(\gamma) \geq \limsup_{m \to \infty} L(\gamma_m) = d(p,q)\,,$$
and so $L(\gamma) = d(p,q)$.

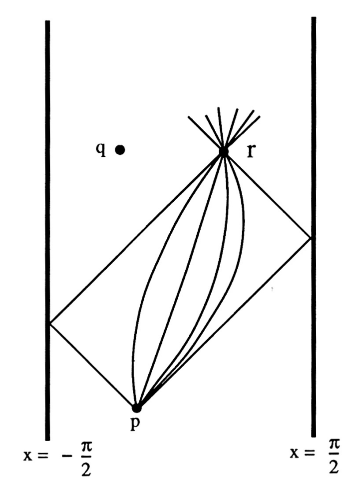
- Each sub-segment of γ must have maximal length. Using the maximality of timelike geodesics in convex neighborhoods (Proposition 1.3), one can show that γ is a timelike geodesic.

Remarks:

(1) There are simple examples showing that if either of the conditions (1) or (2) fail to hold in the definition of global hyperbolicity then maximal segments may fail to exist.

(2) Contrary to the situation in Riemannian geometry, geodesic completeness does not guarantee the existence of maximal segments. as is well illustrated by anti-de Sitter space which is geodesically complete.

Two-dimensional anti-de Sitter space:
$M = \{(t,x) : -\pi/2 < x < \pi/2\}$, $ds^2 = \sec^2 x(-dt^2 + dx^2)$
It can be shown that all future directed timelike geodesics emanating from p refocus at r. The points p and q are timelike related, but there is no timelike geodesic segment from p to q.



4.2 Cauchy hypersurfaces.

Global hyperbolicity is closely related to the existence of certain ‘ideal initial value hypersurfaces’, called Cauchy surfaces. There are slight variations in the literature in the definition of a Cauchy surface. Here we adopt the following definition.

Definition 4.4. *A Cauchy surface for a spacetime M is an achronal subset S of M which is met by every inextendible causal curve in M.*

Observations:

(1) If S is a Cauchy surface for M then $\partial I^+(S) = S$. (Exercise! Also $\partial I^-(S) = S$.) It follows from Proposition 3.1 that a Cauchy surface S is a closed achronal C^0 hypersurface in M.

(2) If S is Cauchy then every inextendible timelike curve meets S exactly once.

Theorem 4.6 (Geroch, [7]). *If a spacetime M is globally hyperbolic then it has a Cauchy surface S, and conversely.*

Comments on the proof: The converse will be discussed in the next section.

- Introduces a measure μ on M such that $\mu(M) = 1$, and consider the function $f : M \to \mathbb{R}$ defined by

$$f(p) = \frac{\mu(J^-(p))}{\mu(J^+(p))}\,.$$

- Internal compactness is used to show that f is continuous, and strong causality is used to show that f is strictly increasing along future directed causal curves.
- Moreover, if $\gamma : (a, b) \to M$ is a future directed inextendible causal curve in M, one shows $f(\gamma(t)) \to 0$ as $t \to a^+$, and $f(\gamma(t)) \to \infty$ as $t \to b^-$.
- It follows that 'slices' of f, e.g., $S = \{p \in M : f(p) = 1\}$, are Cauchy surfaces for M. $\square$

<u>Remark:</u> The function f constructed in the proof is what is referred to as a *time function*, namely, a continuous function that is strictly increasing along future directed causal curves. See e.g, [1, 3] for recent developments concerning the construction of *smooth time functions*, i.e., smooth functions with (past directed) timelike gradient (which hence are necessarily time functions) and smooth spacelike Cauchy surfaces.

Proposition 4.7. *Let M be gobally hyperbolic.*

(1) If S is a Cauchy surface for M then M is homeomorphic to $\mathbb{R} \times S$.

(2) Any two Cauchy surfaces in M are homeomorphic.

Proof. To prove (1), one introduces a future directed timelike vector field X on M. X can be scaled so that the time parameter t of each integral curve of X extends from $-\infty$ to ∞, with $t = 0$ at points of S. Each $p \in M$ is on an integral curve of X that meets S in exactly one point q. This sets up a correspondence $p \leftrightarrow (t, q)$, which gives the desired homeomorphism. A similar technique may be used to prove (2) □

In view of Proposition 4.7, any nontrivial topology in a globally hyperbolic spacetime must reside in its Cauchy surfaces.

The following fact is often useful.

Proposition 4.8. *If S is a compact achronal C^0 hypersurface in a globally hyperbolic spacetime M then S must be a Cauchy surface for M.*

The proof will be discussed in the next section.

5 Domains of dependence

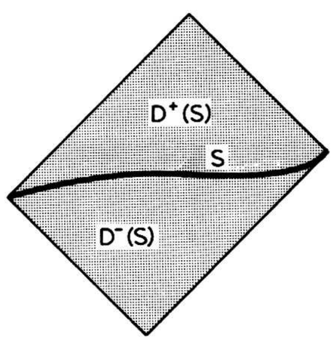
Definition 5.1. *Let S be an achronal set in a spacetime M. The future domain of dependence of $D^+(S)$ of S is defined as follows,*

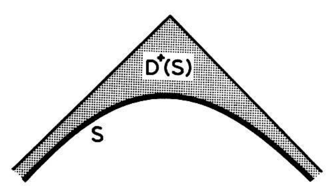
$$D^+(S) = \{p \in M : \text{ every past inextendible causal curve from } p \text{ meets } S\}$$

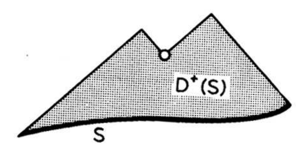
The past domain of dependence of $D^-(S)$ is defined in a time-dual manner. The (total) domain of dependence of S is the union, $D(S) = D^+(S) \cup D^-(S)$.

In physical terms, since information travels along causal curves, a point in $D^+(S)$ only receives information from S. Thus if physical laws are suitably causal, initial data on S should determine the physics on $D^+(S)$.

Below we show a few examples of future and past domains of dependence.







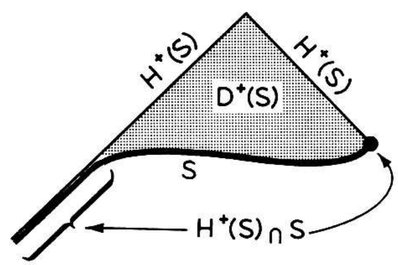
Note: $D^+(S) \supset S$.

If S is achronal, the *future Cauchy horizon* $H^+(S)$ of S is the future boundary of $D^+(S)$. This is made precise in the following definition.

Definition 5.2. *Let $S \subset M$ be achronal. The future Cauchy horizon $H^+(S)$ of S is defined as follows*

$$\begin{aligned} H^+(S) &= \{p \in \overline{D^+(S)} : I^+(p) \cap D^+(S) = \emptyset\} \\ &= \overline{D^+(S)} \setminus I^-(D^+(S)) \, . \end{aligned}$$

The past Cauchy horizon $H^-(S)$ is defined time-dually. The (total) Cauchy horizon of S is defined as the union, $H(S) = H^+(S) \cup H^-(S)$. (See the figure below.)



We record some basic facts about domains of dependence and Cauchy horizons.

Proposition 5.1. *Let S be an achronal subset of M. Then the following hold.*

(1) If $p \in D^+(S)$ then $I^-(p) \cap I^+(S) \subset D^+(S)$.

(2) $\partial D^+(S) = H^+(S) \cup S$.

(3) $H^+(S)$ is achronal.

(4) $\operatorname{edge} H^+(S) \subset \operatorname{edge} \mathrm{S}$, with equality holding if S is closed.

Comment on the proof: The achronality of $H^+(S)$ follows almost immediately from the definition: Suppose $p, q \in H^+(S)$ with $q \in I^+(p)$. Since $q \in \overline{D^+(S)}$, and $I^+(p)$ is a neighborhood of q, $I^+(p)$ meets $D^+(S)$, contradicting the definition of $H^+(S)$. We leave the proofs of the other parts as an exercise.

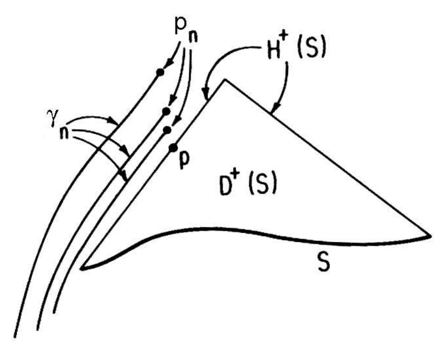
Cauchy horizons have structural properties similar to achronal boundaries, as indicated in the next two results. From Proposition 3.3, we obtain the following.

Proposition 5.2. *Let $S \subset M$ be achronal. Then $H^+(S) \setminus \operatorname{edge} H^+(S)$, if nonempty, is an achronal C^0 hypersurface in M.*

In a similar vein to Proposition 3.4, we have the following.

Proposition 5.3. *Let S be an achronal subset of M. Then $H^+(S)$ is ruled by null geodesics, i.e., every point of $H^+(S) \setminus \operatorname{edge} S$ is the future endpoint of a null geodesic in $H^+(S)$ which is either past inextendible in M or else has a past end point on* $\operatorname{edge} S$.

Comments on the proof. The proof uses a limit curve argument. Consider the case $p \in H^+(S) \setminus S$. Since $I^+(p) \cap D^+(S) = \emptyset$, we can find a sequence of points $p_n \notin D^+(S)$, such that $p_n \to p$. For each n, there exists a past inextendible causal curve γ_n that does not meet S. By the limit curve lemma there exists a subsequence γ_m that converges to a past inextendible C^0 causal curve γ starting at p. Near p this defines the desired null geodesic (see the figure below).



The case $p \in S \setminus \operatorname{edge} S$ is handled somewhat differently; for details see [13, p. 203]. □

Proposition 5.4. *Let S be an achronal subset of a spacetime M. Then, S is a Cauchy surface for M if and only if $D(S) = M$ if and only if $H(S) = \emptyset$.*

Proof. Follows straight-forwardly from definitions, together with the fact that, for S achronal, $\partial D(S) = H(S)$ (exercise).

The following basic result ties domains of dependence to global hyperbolicity.

Proposition 5.5. *Let $S \subset M$ be achronal.*

(1) Strong causality holds on $\operatorname{int} D(S)$.

(2) Internal compactness holds on $\operatorname{int} D(S)$, *i.e., for all $p, q \in \operatorname{int} D(S)$, $J^+(p) \cap J^-(q)$ is compact.*

Proof. See the notes.

We can now address the converse part of Theorem 4.6.

Corollary 5.6. *If S is a Cauchy surface for M then M is globally hyperbolic.*

Proof. This follows immediately from Propositions 5.4 and 5.5: S Cauchy $\implies D(S) = M \implies \operatorname{int} D(S) = M \implies M$ is globally hyperbolic. □

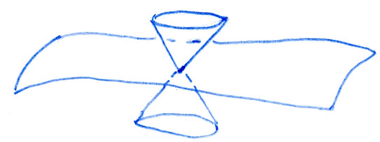
Re: Proposition 4.8.: *If S is a compact achronal C^0 hypersurface in a globally hyperbolic spacetime M then S must be a Cauchy surface for M.*

Proof of Proposition 4.8. It suffices to show that $H(S) = H^+(S) \cup H^-(S) = \emptyset$. Suppose there exists $p \in H^+(S)$. Since S is edgeless, it follows from Proposition 5.3 that p is the future endpoint of a past inextendible null geodesic $\gamma \subset H^+(S)$. Then since $\gamma \subset D^+(S) \cap J^-(p)$ (exercise: show this), we have that γ is contained in the set $J^+(S) \cap J^-(p)$, which is compact by Proposition 4.3. By Lemma 4.2 strong causality must be violated at some point of $J^+(S) \cap J^-(p)$. Thus $H^+(S) = \emptyset$, and time-dually, $H^-(S) = \emptyset$.

The notes conclude this section with the statement of several lemmas that are useful in proving some of the results described here, as well as other results concerning domains of dependence.

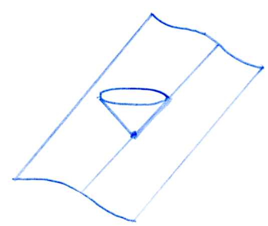
6 The geometry of null hypersurfaces

In addition to curves, one can discuss the causality of certain higher dimensional submanifolds. For example, a *spacelike* hypersurface is a hypersurface all of whose tangent vectors are spacelike, or, equivalently, whose normal vectors are timelike:



More precisely, a hypersurface is spacelike if the induced metric is positive definite (i.e. Riemannian). In GR, a spacelike hypersurface represents space at a given instant of time.

A null hypersurface is a hypersurface such that the null cone is tangent to at each of its points:



Null hypersurfaces play an important role in GR as they represent horizons of various sorts. For example the event horizons in the Schwarzschild and Kerr spacetimes are null hypersurfaces. Null hypersurfaces have an interesting geometry which we would like to discuss in this section.

Definition 6.1. *A null hypersurface in a spacetime* (M, g) *is a smooth co-dimension one submanifold* S *of* M*, such that at each* $p \in S$*,* $g : T_pS \times T_pS \to \mathbb{R}$ *is degenerate.*

This means that there exists a nonzero vector $K_p \in T_pS$ (the direction of degeneracy) such that

$$\langle K_p, X\rangle = 0 \quad \text{for all} \quad X \in T_pS$$

where we have introduced the shorthand metric notation: $\langle U, V\rangle = g(U, V)$. In particular,

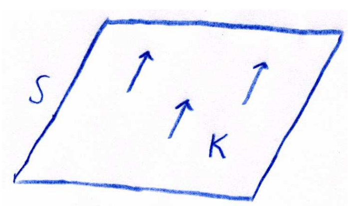
(1) K_p is a null vector, $\langle K_p, K_p\rangle = 0$, which we can choose to be future pointing, and

(2) $[K_p]^\perp = T_pS$.

(3) Moreover, every (nonzero) vector $X \in T_pS$ that is not a multiple of K_p is spacelike.

Thus, every null hypersurface S gives rise to a smooth future directed null vector field K on S, unique up to a positive pointwise scale factor.

$$p \in S \xrightarrow{K} K_p \in T_pS,$$



Ex. $\mathbb{M}^{n+1}$ = Minkowski space.

(1) Null hyperplanes in $\mathbb{M}^{n+1}$: Each nonzero null vector $X \in T_p\mathbb{M}^{n+1}$ determines a null hyperplane $\Pi = \{q \in \mathbb{M}^{n+1} : \langle \overline{pq}, X\rangle = 0\}$.

(2) Null cones in $\mathbb{M}^{n+1}$: The past and future cones, $\partial I^-(p)$ and $\partial I^+(p)$, respectively, are smooth null hypersurfaces away from the vertex p.

The following fact is fundamental.

Proposition 6.1. *Let S be a smooth null hypersurface and let K be a smooth future directed null vector field on S. Then the integral curves of K are null geodesics (when suitably parameterized),*

Remark: The integral curves of K are called the null generators of S. Apart from parameterizations, the null generators are intrinsic to the null hypersurface.

Proof. Suffices to show:

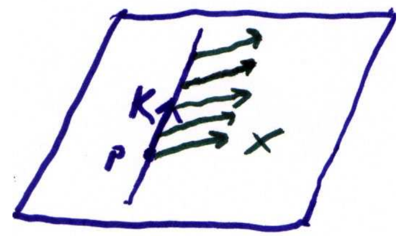
$$\nabla_K K = \lambda K$$

This follows by showing at each $p \in S$,

$$\nabla_K K \perp T_pS\,, \quad \text{i.e.,} \quad \langle \nabla_K K, X\rangle = 0 \quad \forall X \in T_pS$$

Extend $X \in T_pS$ by making it invariant under the flow generated by K,

$$[K, X] = \nabla_K X - \nabla_X K = 0$$



X remains tangent to S, so along the flow line through p,

$$\langle K, X \rangle = 0 \tag{6.15}$$

Differentiating,

$$K\langle K, X \rangle = \langle \nabla_K K, X \rangle + \langle K, \nabla_K X \rangle = 0$$

$$\langle \nabla_K K, X \rangle = -\langle K, \nabla_X K \rangle = -\frac{1}{2} X \langle K, K \rangle = 0.$$

□

To study the 'shape' of the null hypersurface S we study how the null vector field K varies along S. Since K is actually orthogonal to S, this is somewhat analogous to how we study the shape of a hypersurface in a Riemannian manifold, or spacelike hypersurface in a Lorentzian manifold, by introducing the shape operator (or Weingarten map) and associated second fundamental form. We proceed to introduce null analogues of these objects. For technical reasons one works "mod K", as described below.

Null Weingarten Map/Null 2nd Fundamental Form.

- We introduce the following equivalence relation on tangent vectors: For $X, Y \in T_pS$,

$$X = Y \text{ mod } K \iff X - Y = \lambda K \tag{6.16}$$

  Let $\overline{X}$ denote the equivalence class of $X \in T_pS$ and let,

$$T_pS/K = \{\overline{X} : X \in T_pS\} \tag{6.17}$$

  Then,

$$TS/K = \cup_{p\in S} T_pS/K \tag{6.18}$$

  is a rank $n-1$ vector bundle over S ($n = \dim S$). This vector bundle does not depend on the particular choice of null vector field K.

- There is a natural positive definite metric h on TS/K induced from $\langle\,,\rangle$: For each $p \in S$, define $h : T_pS/K \times T_pS/K \to \mathbb{R}$ by

$$h(\overline{X}, \overline{Y}) = \langle X, Y\rangle.$$

Well-defined: If $X' = X \mod K$, $Y' = Y \mod K$ then

$$\begin{aligned}\langle X', Y' \rangle &= \langle X + \alpha K, Y + \beta K \rangle \\ &= \langle X, Y \rangle + \beta \langle X, K \rangle + \alpha \langle K, Y \rangle + \alpha\beta \langle K, K \rangle \\ &= \langle X, Y \rangle \,.\end{aligned}$$

- The *null Weingarten map* $b = b_K$ of S with respect to K is, for each point $p \in S$, a linear map $b : T_pS/K \to T_pS/K$ defined by

$$b(\overline{X}) = \overline{\nabla_X K} \,.$$

b is well-defined: $X' = X \mod K \Rightarrow$

$$\begin{aligned}\nabla_{X'} K &= \nabla_{X+\alpha K} K \\ &= \nabla_X K + \alpha \nabla_K K = \nabla_X K + \alpha\lambda K \\ &= \nabla_X K \quad \mod K\end{aligned}$$

- b is self adjoint with respect to h, i.e., $h(b(\overline{X}), \overline{Y}) = h(\overline{X}, b(\overline{Y}))$, for all $\overline{X}, \overline{Y} \in T_pS/K$.

Proof. Extend $X, Y \in T_pS$ to vector fields tangent to S near p. Using $X\langle K, Y\rangle = 0$ and $Y\langle K, X\rangle = 0$, we obtain,

$$\begin{aligned} h(b(\overline{X}), \overline{Y}) &= \langle \nabla_X K, Y\rangle = -\langle K, \nabla_X Y\rangle = -\langle K, \nabla_Y X\rangle + \langle K, [X, Y]\rangle \\ &= \langle \nabla_Y K, X\rangle = h(\overline{X}, b(\overline{Y}))\,. \end{aligned}$$

- The *null second fundamental form* $B = B_K$ of S with respect to K is the bilinear form associated to b via h: For each $p \in S$, $B : T_pS/K \times T_pS/K \to \mathbb{R}$ is defined by,

$$B(\overline{X}, \overline{Y}) = h(b(\overline{X}), \overline{Y}) = h(\overline{\nabla_X K}, \overline{Y}) = \langle \nabla_X K, Y\rangle\,.$$

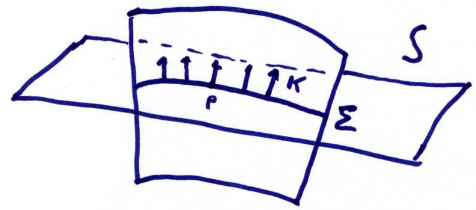
  Since b is self-adjoint, B is symmetric.

- The *null mean curvature* (or *null expansion scalar*) of S with respect to K is the smooth scalar field θ on S defined by,

$$\theta = \operatorname{tr} b$$

  .

  θ has a natural geometric interpretation. Let Σ be the intersection of S with a hypersurface in M which is transverse to K near $p \in S$; Σ will be a co-dimension two spacelike submanifold of M, along which K is orthogonal.

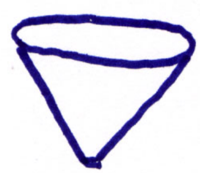


Let $\{e_1, e_2, \cdots, e_{n-1}\}$ be an orthonormal basis for $T_p\Sigma$ in the induced metric. Then $\{\bar{e}_1, \bar{e}_2, \cdots, \bar{e}_{n-1}\}$ is an orthonormal basis for T_pS/K. Hence at p,

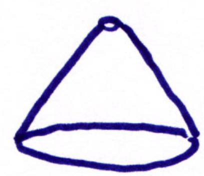
$$\begin{aligned}\theta = \operatorname{tr} b = \sum_{i=1}^{n-1} h(b(\bar{e}_i), \bar{e}_i) = \sum_{i=1}^{n-1} \langle \nabla_{e_i} K, e_i \rangle. \\ = \operatorname{div}_\Sigma K\,. \end{aligned} \tag{6.19}$$

where $\operatorname{div}_\Sigma K$ is the divergence of K along Σ.

Thus, θ measures the overall expansion of the null generators of S towards the future.

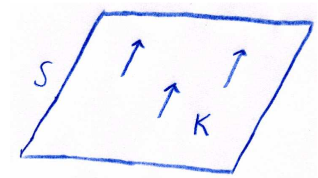


$\theta > 0$



$\theta < 0$

Re: The Geometry of Null Hypersurfaces



- S = null hypersurface in spacetime M
  K = associated future directed null vector field
- Equivalence relation:

$$X = Y \text{ mod } K \iff X - Y = \lambda K$$

- mod K tangent space: $T_pS/K = \{\overline{X} : X \in T_pS\}$
  mod K tangent bundle: $TS/K = \cup_{p\in S} T_pS/K$
- Metric h on TS/K: $h(\overline{X}, \overline{Y}) = \langle X, Y\rangle$ (positive definite).
- Null Weingarten map $b : T_pS/K \to T_pS/K$, defined by

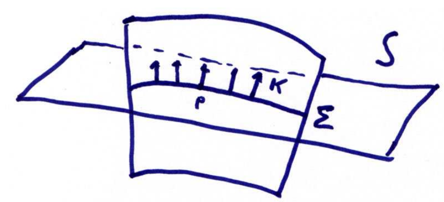
$$b(\overline{X}) = \overline{\nabla_X K}$$

- Associated null second fundamental form $B : T_pS/K \times T_pS/K \to \mathbb{R}$,

$$B(\overline{X}, \overline{Y}) = h(b(\overline{X}), \overline{Y}) = h(\overline{\nabla_X K}, \overline{Y}) = \langle \nabla_X K, Y\rangle$$

- Null expansion scalar θ,

$$\theta := \operatorname{tr} b = \operatorname{tr} B = \operatorname{div}_\Sigma K$$



Exercise: Effect of scaling. Show that if $\widetilde{K} = fK$, $f \in C^\infty(S)$, is any other future directed null vector field on S, then $b_{\tilde{K}} = f b_K$, and hence, $\widetilde{\theta} = f\theta$. It follows that the Weingarten map $b = b_K$ at a point p is uniquely determined by the value of K at p.

Comparison Theory

We now study how the null Weingarten map propagates along the null geodesic generators of S.

Let $\eta : I \to M$, $s \to \eta(s)$, be a future directed affinely parameterized null geodesic generator of S. For each $s \in I$, consider the Weingarten map $b = b(s)$ based at $\eta(s)$ with respect to the null vector field $K = \eta'(s)$ at $\eta(s)$,

$$b(s) = b_{\eta'(s)} : T_{\eta(s)}S/\eta'(s) \to T_{\eta(s)}S/\eta'(s)$$

Proposition 6.2. *The one parameter family of Weingarten maps $s \to b(s)$, obeys the following Riccati equation,*

$$b' + b^2 + R = 0\,, \qquad {}' = \nabla_{\eta'} \tag{6.20}$$

where $R : T_{\eta(s)}S/\eta'(s) \to T_{\eta(s)}S/\eta'(s)$ is given by $R(\overline{X}) = \overline{R(X,\eta'(s))\eta'(s)}$.

Remark on notation: In general, if $Y = Y(s)$ is a vector field along η tangent to S, we define, $(\overline{Y})' = \overline{Y'}$. Then, if $X = X(s)$ is a vector field along η tangent to S, b' is defined by,

$$b'(\overline{X}) := b(\overline{X})' - b(\overline{X'})\,. \tag{6.21}$$

Proof. Fix a point $p = \eta(s_0)$, $s_0 \in (a, b)$, on η. On a neighborhood U of p in S we can scale the null vector field K so that K is a geodesic vector field, $\nabla_K K = 0$, and so that K, restricted to η, is the velocity vector field to η, i.e., for each s near s_0, $K_{\eta(s)} = \eta'(s)$. Let $X \in T_pM$. Shrinking U if necessary, we can extend X to a smooth vector field on U so that $[X, K] = \nabla_X K - \nabla_K X = 0$. Then,

$$R(X, K)K = \nabla_X \nabla_K K - \nabla_K \nabla_X K - \nabla_{[X,K]} K = -\nabla_K \nabla_K X$$

Hence along η we have, $X'' = -R(X, \eta')\eta'$ (which implies that X, restricted to η, is a Jacobi field along η).

Thus, from Equation (6.21), at the point p we have,

$$\begin{aligned}
b'(\overline{X}) &= \overline{\nabla_X K}\,' - b(\overline{\nabla_K X}) = \overline{\nabla_K X}\,' - b(\overline{\nabla_X K}) \\
&= \overline{X''} - b(b(\overline{X})) = -\overline{R(X, \eta')\eta'} - b^2(\overline{X}) \\
&= -R(\overline{X}) - b^2(\overline{X}),
\end{aligned}$$

which establishes Equation 6.20.

By taking the trace of (6.20) we obtain the following formula for the derivative of the null mean curvature $\theta = \theta(s)$ along η,

$$\theta' = -\mathrm{Ric}(\eta', \eta') - \sigma^2 - \frac{1}{n-1}\theta^2, \tag{6.22}$$

where $\sigma := (\mathrm{tr}\, \hat{b}^2)^{1/2}$ is the *shear scalar*, $\hat{b} := b - \frac{1}{n-1}\theta \cdot \mathrm{id}$ is the trace free part of the Weingarten map, and $\mathrm{Ric}(\eta', \eta')$ is the spacetime Ricci tensor evaluated on the tangent vector η'.

Equation 6.22 is known in relativity as the **Raychaudhuri equation** (for an irrotational null geodesic congruence). This equation shows how the Ricci curvature of spacetime influences the null mean curvature of a null hypersurface.

We consider a basic application of the Raychaudhuri equation.

Proposition 6.3. *Let M be a spacetime which obeys the null enery condition (NEC), $\mathrm{Ric}\,(X, X) \geq 0$ for all null vectors X, and let S be a smooth null hypersurface in M. If the null generators of S are future geodesically complete then S has nonnegative null expansion, $\theta \geq 0$.*

Proof. Suppose $\theta < 0$ at $p \in S$. Let $s \to \eta(s)$ be the null generator of S passing through $p = \eta(0)$, affinely parametrized. Let $b(s) = b_{\eta'(s)}$, and take $\theta = \mathrm{tr}\, b$. By the invariance of sign under scaling, one has $\theta(0) < 0$.

Raychaudhuri's equation and the NEC imply that $\theta = \theta(s)$ obeys the inequality,

$$\frac{d\theta}{ds} \leq -\frac{1}{n-1}\theta^2\,, \tag{6.23}$$

and hence $\theta < 0$ for all $s > 0$. Dividing through by θ^2 then gives,

$$\frac{d}{ds}\left(\frac{1}{\theta}\right) \geq \frac{1}{n-1}\,, \tag{6.24}$$

which implies $1/\theta \to 0$, i.e., $\theta \to -\infty$ in finite affine parameter time, contradicting the smoothness of θ. $\square$

<u>Exercise.</u> Let Σ be a local cross section of the null hypersurface S, (as in hte figure on p. 53), with volume form ω. If Σ is moved under flow generated by K, show that $L_K\omega = \theta\,\omega$, where L = Lie derivative.

Thus, Proposition 6.3 implies, under the given assumptions, that cross sections of S are nondecreasing in area as one moves towards the future. Proposition 6.3 is the simplest form of Hawking's black hole area theorem [8]. For a study of the area theorem, with a focus on issues of regularity, see [2].

7 Trapped surfaces and the Penrose Singularity Theorem

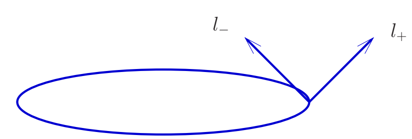
In this section we introduce the important notion of a trapped surface and present the classical Penrose singularity theorem.

- Let (M^{n+1}, g) be an $(n+1)$-dimensional spacetime, $n \geq 3$.

  Let Σ^{n-1} be a closed (i.e., compact without boundary) co-dimension two spacelike submanifold of M.

- Each normal space of Σ, $[T_p\Sigma]^\perp$, $p \in \Sigma$, is timelike and 2-dimensional, and hence admits two future directed null directions orthogonal to Σ.

  Thus, under suitable orientation assumptions, Σ admits two smooth nonvanishing future directed null normal vector fields l_+ and l_- (unique up to positive rescaling).

  By convention, we refer to l_+ as outward pointing and l_- as inward pointing.

- Associated to l_+ and l_-, are the two *null second fundamental forms*, χ_+ and χ_-, respectively, defined as

$$\chi_\pm : T_p\Sigma \times T_p\Sigma \to \mathbb{R}, \qquad \chi_\pm(X,Y) = g(\nabla_X l_\pm, Y)\,. \tag{7.25}$$

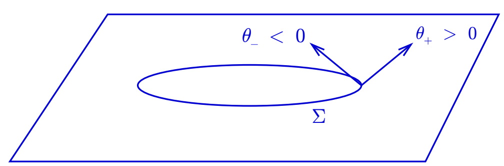
- *The null expansion scalars (or null mean curvatures)* $\theta_\pm$ of Σ are obtained by tracing $\chi_\pm$ with respect to the induced metric γ on Σ,

$$\theta_\pm = \mathrm{tr}_\gamma \chi_\pm = \gamma^{AB}\chi_{\pm AB} = \mathrm{div}\,_\Sigma l_\pm\,. \tag{7.26}$$

  The sign of $\theta_\pm$ does not depend on the scaling of $l_\pm$. Physically, θ_+ (resp., θ_-) measures the divergence of the outgoing (resp., ingoing) light rays emanating orthogonally from Σ.

  (There is a natural connection between these null expansion scalars and the null expansion of null hypersurfaces; see the notes.)

- For round spheres in Euclidean slices in Minkowski space (and, more generally, large "radial" spheres in AF spacelike hypersurfaces),



- However, in regions of spacetime where the gravitational field is strong, one can have both

$$\theta_- < 0 \quad \text{and} \quad \theta_+ < 0\,,$$

  in which case Σ is called a **trapped surface**.

  As we now discuss, under appropriate energy and causality conditions, the occurrence of a trapped surface signals the onset of gravitational collapse

7.1 The Penrose Singularity Theorem.

The Penrose singularity theorem [11] is the first of the famous singularity theorems of general relativity. The singularity theorems establish, under generic circumstances, the existence in spacetime of incomplete timelike or null geodesics.

Such incompleteness indicates that spacetime comes to an end either in the past or future. Future incompleteness is often associated with gravitational collapse and the formation of a black hole (such as in Schwarzschild and Kerr).

All the classical singularity theorems require energy conditions. The Penrose singularity theorem assumes the NEC, $\mathrm{Ric}(X, X) \geq 0$ for all null vectors X, holds.

It also assumes the existence of a *noncompact* Cauchy surface. This is consistent with modeling an isolated gravitating system, where one takes spacetime (or space) to be asymptotically flat.

What Penrose proves is that if the gravitational field becomes sufficiently strong that trapped surfaces appear (as they do in the Schwarzschild solution) then the development of singularities is inevitable.

Theorem 7.1 (Penrose singularity theorem)**.** *Let M be a globally hyperbolic spacetime which satisfies the null energy condition and which has noncompact Cauchy surfaces. If M contains a trapped surface Σ then M is future null geodesically incomplete.*

Proof. The proof is by contradiction. Assume M is future null geodesically complete. Then what we actually show is the following.

Claim: If M is future null geodesically complete then the $\partial I^+(\Sigma)$ is compact.

But $\partial I^+(\Sigma)$ is an achronal C^0 hypersurface (Proposition 3.1), and hence, by Proposition 4.8, is a compact Cauchy surface. This contradicts the assumption that M admits a noncompact Cauchy surface.

Proof of Claim: Suppose to the contrary that $\partial I^+(\Sigma)$ is noncompact.

- By Proposition 3.4 (or Proposition 2.1) each $q \in \partial I^+(\Sigma)$ lies on a null geodesic in $\partial I^+(\Sigma)$ with past end point on Σ. Moreover this null geodesic meets Σ orthogonally (due to achronality, cf. [10, Lemma 50, p. 298])).

- Since $\partial I^+(\Sigma)$ is closed and noncompact, there exists a sequence of points $\{q_n\} \subset \partial I^+(\Sigma)$ that diverges to infinity. For each n, there is a null geodesic η_n from Σ to q_n, which is contained in $\partial I^+(\Sigma)$ and meets Σ orthogonally.

- By compactness of Σ, some subsequence converges to a future inextendible, and hence future complete, null geodesic $\eta : [0, \infty) \to M$ contained in $\partial I^+(\Sigma)$, and meeting Σ orthogonally (at p, say).

- By achronality of $\partial I^+(\Sigma)$,

  - No other null geodesic in $\partial I^+(\Sigma)$ starting on Σ can meet η.

  - There can be no *null focal point* to Σ along η (cf. [10, Prop. 48, p. 296]).

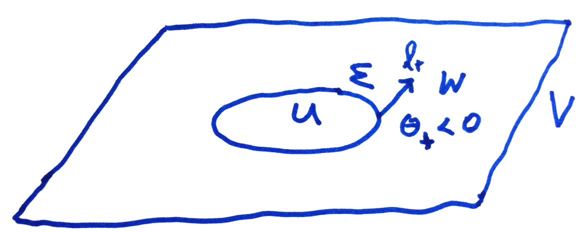
- It follows that η is contained in a smooth (perhaps very thin) null hypersurface $H \subset \partial I^+(\Sigma)$.

- Let θ be the null expansion of H along η. Since Σ is a trapped surface $\theta(p) < 0$. Arguing just as in the "area theorem" (Proposition 6.3), using Raychaudhuri + NEC, θ must go to $-\infty$ in finite affine parameter time $\rightarrow\leftarrow$ $\square$

7.2 A variant of the Penrose singularity theorem.

For certain applications, the following variant of the Penrose singularity theorem is useful.

Theorem 7.2. *Let M be a globally hyperbolic spacetime satisfying the null energy condition, with smooth spacelike Cauchy surface V. Let Σ be a smooth closed (compact without boundary) hypersurface in V which separates V into an "inside" U and an "outside" W, i.e., $V \setminus \Sigma = U \cup W$ where $U, W \subset V$ are connected disjoint sets. Suppose, further, that $\overline{W}$ is non-compact. If Σ is* **outer-trapped** *($\theta_+ < 0$) then M is future null geodesically incomplete.*



Proof. Exercise. Hint: Consider the achronal boundary $\partial I^+(\overline{U})$ and argue similarly to the proof of the Penrose singularity theorem.

This version of the Penrose singularity theorem may be used to prove the following beautiful result of Gannon [6] and Lee [9].

Theorem 7.3. *Let M be a globally hyperbolic spacetime which satisfies the null energy condition and which contains a smooth asymptotically flat spacelike Cauchy surface V. If V is not simply connected ($\pi_1(V) \neq 0$) then M is future null geodesically incomplete.*

Thus, as suggested by this theorem, nontrivial topology tends to induce gravitational collapse. In the standard collapse scenario (based on the *weak cosmic censorship conjecture*) the process of gravitational collapse leads to the formation of an event horizon which shields the singularities from view. According to the *principle of topological censorship* the nontrivial topology that induced collapse should end up behind hidden the event horizon, and the region outside the black hole should have simple topology. There are a number of results supporting this view. See [4] for further discussion, relevant references and related results.

<u>Exercise:</u> Let M be a globally hyperbolic spacetime which satisfies the null energy condition and which contains a smooth asymptotically flat spacelike Cauchy surface V. Use Theorem 7.2 to show that if V has more than one asymptotically flat end then M is future null geodesically incomplete. Thus, in Theorem 7.3 one might as well assume that V has only one asymptotically flat end.

References

[1] A. N. Bernal and M. Sánchez, *Smoothness of time functions and the metric splitting of globally hyperbolic spacetimes*, Comm. Math. Phys. **257** (2005), no. 1, 43–50.

[2] P. T. Chruściel, E. Delay, G. J. Galloway, and R. Howard, *Regularity of horizons and the area theorem*, Ann. Henri Poincaré **2** (2001), no. 1, 109–178.

[3] P. T. Chruściel, D. E. Grant, and E. Minguzzi, *On differentiability of volume time functions*, 2013, arXiv:1301.2909v1.

[4] M. Eichmair, G. J. Galloway, and D. Pollack, *Topological censorship from the initial data point of view*, J. Differential Geom. **95** (2013), no. 3, 389–405.

[5] G. J. Galloway and C. Vega, *Achronal limits, lorentzian spheres, and splitting*, 2012, arXiv:1211.2460v2, to appear in Annales Henri, Poincaré.

[6] D. Gannon, *Singularities in nonsimply connected space-times*, J. Mathematical Phys. **16** (1975), no. 12, 2364–2367.

[7] R. Geroch, *Domain of dependence*, J. Mathematical Phys. **11** (1970), 437–449.

[8] S. W. Hawking and G. F. R. Ellis, *The large scale structure of space-time*, Cambridge University Press, London, 1973, Cambridge Monographs on Mathematical Physics, No. 1.

[9] C. W. Lee, *A restriction on the topology of Cauchy surfaces in general relativity*, Comm. Math. Phys. **51** (1976), no. 2, 157–162.

[10] Barrett O'Neill, *Semi-Riemannian geometry*, Pure and Applied Mathematics, vol. 103, Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1983, With applications to relativity.

[11] R. Penrose, *Gravitational collapse and space-time singularities*, Phys. Rev. Lett. **14** (1965), 57–59. MR MR0172678 (30 #2897)

[12] ———, *Techniques of differential topology in relativity*, Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1972, Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 7. MR MR0469146 (57 #8942)

[13] Robert M. Wald, *General relativity*, University of Chicago Press, Chicago, IL, 1984.