Causal Structure & Singularities

Advanced Topics of Gravity - Spring 2021

Wei Song

Prepared by: Kevin Loo

Contents

This is the lecture note from Professor Wei Song's Advanced Topics of Gravity course given in Spring 2021, at Tsinghua University. This lecture note is based on Professor Wei Song's handwritten notes, Wald [1], and Witten [2].

Part I

Causal Structure

1 Lightcones and Causal Diamonds

In General Relativity, one usually considers a Riemannian geometry with Lorentzian signature. This geometry is somewhat different from our everyday life Euclidean geometry. However, to an excellent approximation, we may think of a two-dimensional curved surface to make sense of the higher dimensional Riemannian geometry. An important consequence of Lorentzian signature geometry is the constraint of causality: one can travel along a worldline only inside the lightcone.

Let (M, g_{ab}) be a spacetime. At each point (event) $q \in M$, its tangent space V_q is isomorphic to Minkowski spacetime. The lightcone passing through the origin of V_q, as a subset of V_q, will be referred to as the **lightcone of** q. As in Special Relativity, at each $q \in M$, we designate half of the lightcone as "future" and the other half as as "past". A simply connected manifold which can have a continuous designation of "future" and "past" as q varies over M, is said to be a **time orientable** spacetime. An important property satisfied by every time orientable spacetime is that there exists a (nonunique) smooth nonvanishing timelike vector field t^a on M (Wald, Lemma 8.1.1).

To understand the causal structure of spacetime, we will have to study causal paths. A path is usually expressed in parametric form as $x^\mu(s)$, where x^μ are local coordinates in spacetime and s is a parameter. We consider two paths to be equivalent if they differ only by a reparametrisation $s \to \tilde{s}(s)$. We require that the tangent vector $\frac{dx^\mu}{ds}$ is nonzero for all s. A path $x^\mu(s)$ is **causal** if its tangent vector $\frac{dx^\mu}{ds}$ is everywhere timelike or null.

In an arbitrary spacetime (M, g_{ab}) we define the **chronological future** of $q \in M$, denoted $I^+(q)$, by

$$I^+(q) = \{p \in M \mid \text{There exists a future directed timelike curve } \lambda(t) \text{ with } \lambda(0) = q \text{ and } \lambda(1) = p\}\,. \quad (1.1)$$

Of course, this has to be inside the future lightcone. Furthermore, if S is any subset that contains the point q, then the chronological future of the subset is $I^+(S) = \bigcup_{q \in S} I^+(q)$. We also define the **causal future** of $q \in M$, denoted $J^+(q)$, by

$$J^+(q) = \{p \in M \mid \text{There exists a future directed causal path } \lambda(t) \text{ with } \lambda(0) = q \text{ and } \lambda(1) = p\}\,. \quad (1.2)$$

Here the points are inside or on the future lightcone of q. If $p \in J^+(q) - I^+(q)$, then any causal path connecting q to p must be a **null geodesic** (Wald, Cor. of Th. 8.1.2). The subset of M generated by null geodesics from q is the boundary of $J^+(q)$, denoted $\partial J^+(q)$ (or $\dot{J}^+(q)$); it is also known as the lightcone in some literature. Likewise, one can define the **chronological past** $I^-(q)$ and **causal past** $J^-(q)$ of $q \in M$.

All causal paths lie within a subset of M called the **causal diamond** $D_q^p = J^+(q) \cap J^-(p)$, as depicted in Fig. 1. When D_q^p is compact, the space of all causal paths from q to p is compact, so that there exists a geodesic that *maximises* the proper time from q to p.

1. The first essential point is that the space of causal paths from q to p is *compact*. Causality plays an important role here. Without it, a sequence of paths, even if confined to a compact region of

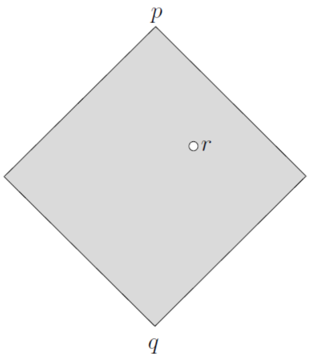

Figure 1: For two spacetime points q, p, with p to the future of q, the causal diamond D_q^p consists of all points that are in the causal future of q and the causal past of p. If a point r is omitted from the causal diamond D_q^p, then D_q^p is not compact.

spacetime (like D_q^p), could oscillate more and more wildly, with no convergent subsequence. This can be understood more concretely using an example of a two-dimensional Minkowski space M with metric $ds^2 = -dt^2 + dx^2$, as argued in WITTEN [2] pp. 6-7.

2. The second essential point is that the proper time is maximised by a unique causal path called the **geodesic**, following the compactness of the space of causal paths. The proper time elapsed along a causal path is

$$\tau = \int ds \sqrt{\left(\frac{dt}{ds}\right)^2 - \left(\frac{d\vec{x}}{ds}\right)^2}. \tag{1.3}$$

To be more precise, there must be an upper bound on τ among all causal paths from q to p, because a sequence of causal paths $\mathbf{x}_n(s) = (t_n(s), \vec{x}_n(s))$ whose elapsed proper time τ_n grows without limit for $n \to \infty$ could not have a convergent subsequence. This can be understood better in the *twin paradox*: a twin who travels from q to p along a non-geodesic path, accelerating at some point along the way, comes back younger than a twin whose trajectory from q to p is a geodesic. In a suitable Lorentz frame, the twin whose path is a geodesic was always at rest, and thus the proper time along his worldline equals the coordinate time difference ($\tau = \Delta t$). However, there is no such frame for the twin whose trajectory involved acceleration.

There are a few situations where D_q^p is *not compact*. Some examples are shown as follows.

Example. *Consider a Minkowski spacetime M' with a point r being removed from the interior of D_q^p. Accordingly the causal diamond D_q^p is not compact and the space of causal paths from q to p is not compact. In the usual Minkowski spacetime M, a sequence of causal paths passing through r has a limit. However, in this new spacetime M' the limit does not exist and there is no geodesic from q to p, since the point r is missing.*

Example. *Consider a Weyl-transformed metric. In particular, we replace the usual Minkowski space metric by $e^{\phi}(-dt^2 + d\vec{x}^2)$. If the function ϕ is chosen to blow up at the point r, producing a singularity, this provides a rationale for omitting that point from the spacetime. Again, this is a spacetime in which causal diamonds D_q^p are not compact.*

2 Causality Conditions

While all spacetimes in General Relativity locally have the same qualitative causal structure as in Special Relativity, globally very significant differences can occur. We would like to have a "causally well behaved" spacetime. This notion can be formulated in terms of the following causality conditions. The upshot here is that we will want a *globally hyperbolic spacetime*, which is the strongest causality condition in wide use.

1. The most obvious causality condition is the **absence of closed causal curves** from a point q to itself, since these curves are unphysical.

2. A stronger condition is that a spacetime (M, g_{ab}) has to be **strongly causal**, i.e. for all points $p \in M$ and every neighbourhood O of p, there exists a neighbourhood $V \subset O$ of p such that no causal curve intersects V more than once. Strong causality roughly says that there are no causal curves that are arbitrarily close to being closed.

   A violation of strong causality can be seen in a spacetime that has no closed timelike curves but is not strongly causal. For example, consider the two-dimensional spacetime M with metric $ds^2 = -dvdu + v^2du^2$, where v is real-valued, but u is an angular variable, $u \cong u + 2\pi$. A short calculation shows that the only closed causal curve in M is the curve $v = 0$. If we remove a point p on that curve, to make a new spacetime M' (Fig. 2). Then strong causality is violated at any point $q \in M'$ that has $v = 0$. There is no closed causal curve from q to itself (since the point p was removed), but there are closed causal curves from q that come arbitrarily close to returning to q; these are curves that remain everywhere at very small v. It is reasonable to consider such behaviour to be unphysical.

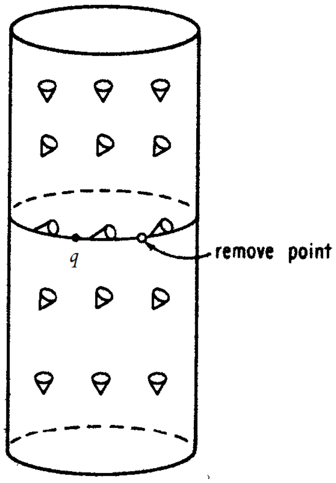

Figure 2: A spacetime which violates strong causality. The light cones "tip over" sufficiently that the curve drawn through q is a null geodesic. Since a point is removed from this manifold, there are no closed causal curves in this spacetime. There are, however, causal curves through p which come arbitrarily close to intersecting themselves.

3. An even stronger condition is known as **stable causality**. A spacetime (M, g_{ab}) is stably causal if there exists a continuous , nonvanishing timelike vector field t^a, such that the spacetime $(M, \tilde{g}_{ab}), \tilde{g}_{ab} = g_{ab} - t_a t_b$ possesses no closed timelike curves (CTCs). This is equivalent to saying that a spacetime (M, g_{ab}) is stably causal if and only if there exists a differentiable function f on M such that $\nabla^a f$ is a past directed timelike vector field (Wald, Th. 8.2.2). This means that stable causality is equivalent to the existence of a "global time function" on the spacetime. In fact, *stable causality implies strong causality* (Wald, Cor. of Th. 8.2.2).

The strongest causality condition is, as we have mentioned, the global hyperbolicity of spacetime. To define it, we would the notion of achronal. A subset $S \subset M$ is **achronal** if there is no timelike path in M connecting distinct points $q, p \in S$. This idea applies to a spatial hypersurface S (a submanifold of codimension 1 in M).

In general, a spatial hypersurface is not necessarily achronal. Consider a two-dimensional cylindrical ($\mathbb{R} \times S^1$) spacetime with flat metric $ds^2 = -dt^2 + d\phi^2$, where t is a real variable, but ϕ is an angular variable, $\phi \cong \phi + 2\pi$. The hypersurface S_ε defined by $t = \varepsilon\phi$, for nonzero ε, wraps infinitely many times around the cylinder (Fig. 3). S_ε is spacelike if ε is small, but it is not achronal. For instance, the points $(t, \phi) = (0, 0)$ and $(t, \phi) = (2\pi\varepsilon, 0)$ in S_ε can be connected by an obvious timelike geodesic. Thus a spacelike hypersurface is not necessarily achronal.

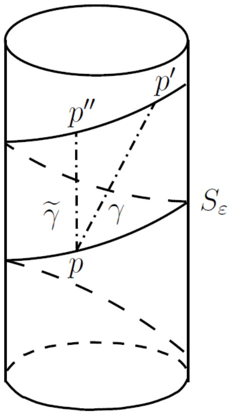

Figure 3: If S_ε is a spacelike hypersurface and γ is a null curve connecting points $p, p' \in S_\varepsilon$, then by moving p' along S_ε "towards" γ, one gets a point $p'' \in S_\varepsilon$ such that there is a strictly timelike curve $\tilde{\gamma}$ from p to p''.

If the subset S is an achronal spacelike hypersurface, then there actually is no causal path from p to p'. In other words, γ does not exist even if it is allowed to be null rather than timelike. This can be seen by considering $p, p' \in S$ connected by a causal path γ that is null (in whole or in part). After displacing p or p' along S in the appropriate direction, γ can be deformed to a timelike path, which does not exist for the achronal S.

Now, let S be a closed, achronal set in M, we can define the **future domain of dependence**, denoted $D^+(S)$, as

$$D^+(S) = \{p \in M \mid \text{Every past inextendible causal curve through } p \text{ intersects } S\}\,. \tag{2.1}$$

Note that we always have $S \subset D^+(S) \subset J^+(S)$ and, since S is achronal, we also have $D^+(S) \cap I^-(S) = \emptyset$. $D^+(S)$ is of interest because if nothing can travel faster than light, then any signal sent to $p \in D^+(S)$ must have "registered" on S. Thus, if we are given appropriate information about *initial conditions* on S, we should be able to predict what happens at $p \in D^+(S)$. Likewise, we can also define the **past domain of dependence** $D^-(S)$ by interchanging "future" and "past". The **(full) domain of dependence** of S, denoted $D(S)$, is defined as simply, $D(S) = D^+(S) \cup D^-(S)$. Therefore, $D(S)$ represents the complete set of events for which all conditions should be determined by a knowledge of conditions on S.

A closed achronal set Σ for which $D(\Sigma) = M$ is called a **Cauchy surface** (or initial value surface). Technically, a closed achronal set Σ is a Cauchy surface if p is a point in M, not in Σ, then every inextendible causal path γ through p intersects Σ. Every Cauchy surface is an embedded C^0 submanifold of M, hence justifying the term "surface". Intuitively, we may think of Σ as representing an "instant of time" throughout the universe, since Σ is achronal.

A spacetime (M, g_{ab}) which possesses a Cauchy surface Σ is said to be **globally hyperbolic**. The intuitive idea is that in a globally hyperbolic spacetime, one can predict the future from the past. Some examples of globally hyperbolic spacetimes are Minkowski and anti-de Sitter (AdS) spacetimes. Some examples of non-globally hyperbolic spacetimes are the globally hyperbolic spacetimes with a point removed, and Reissner-Nordström or Kerr solutions which extend beyond their horizons.

3 Cauchy Horizons

As defined previously, the domain of dependence of S, $D(S)$, is the largest region in M in which the physics can be predicted from a knowledge of initial conditions on S. In fact, $D(S)$ can be regarded as a spacetime in its own right; it is a manifold since it is open in M.

The closure of the future domain of dependence of S, $\overline{D^+(S)}$ is characterised by a property: $p \in \overline{D^+(S)}$ if and only if every past inextendible timelike curve from p intersects S (Wald, Prop. 8.3.2). Some properties of the interiors of $D^+(S)$ and $D(S)$ are

$$\begin{aligned} \operatorname{int}\left[D^+(S)\right] &= I^-\left[D^+(S)\right] \cap I^+(S), \\ \operatorname{int}\left[D(S)\right] &= I^-\left[D^+(S)\right] \cap I^+\left[D^-(S)\right] \end{aligned} \tag{3.1}$$

(Wald, Lemma 8.3.3). It follows immediately from the definition of a Cauchy surface that if Σ is a Cauchy surface, then every inextendible causal curve meets Σ. In fact, we have a stronger result: for a Cauchy surface Σ and an inextendible causal curve λ, λ intersects Σ, $I^+(\Sigma)$, and $I^-(\Sigma)$ (Wald, Prop. 8.3.4).

Let S be a closed achronal set, we define the **future Cauchy horizon** of S, denoted $H^+(S)$, by

$$H^+(S) = \overline{D^+(S)} - I^-\left[D^+(S)\right]. \tag{3.2}$$

The **(full) Cauchy horizon** of a closed achronal set S is defined by

$$H(S) = H^+(S) \cup H^-(S), \tag{3.3}$$

where $H^-(S)$ is the **past Cauchy horizon** of S. It can be shown that the Cauchy horizon marks the boundary of the domain of dependence of S, denoted $\partial D(S)$ or $\dot{D}(S)$ (Wald, Prop. 8.3.6).

There are some important results that we shall mention.

1. (Wald, Th. 8.3.5) If every point $p \in H^+(S)$ lies on a null geodesic λ contained entirely within $H^+(S)$, then the null geodesic is past inextendible or has a past endpoint on the edge of S.
2. (Wald, Th. 8.3.10) For a globally hyperbolic spacetime (M, g_{ab}) and two distinct points $p, q \in M$, the causal diamond $D^p_q = J^+(q) \cap J^-(p)$ is compact.
3. (Wald, Prop. 8.3.13) If Σ and Σ' are both Cauchy surfaces for the globally hyperbolic spacetime (M, g_{ab}), then Σ and Σ' are homeomorphic, i.e. have the same topology.
4. (Wald, Th. 8.3.14) Let (M, g_{ab}) be a globally hyperbolic spacetime. Then
   (a) (M, g_{ab}) is stably causal.
   (b) A global time function, f, can be chosen such that each surface of constant f is a Cauchy surface.
   (c) M can be foliated by Cauchy surfaces.

(d) M has the topology of $\mathbb{R} \times \Sigma$, where Σ denotes any Cauchy surface.

In fact, a globally hyperbolic spacetime (M, g_{ab}) implies that (M, g_{ab}) is strongly causal (Wald, Lemma 8.3.8). We have the following comparison:

$$\text{global hyperbolicity} \geq \text{stable causality} \geq \text{strong causality} \geq \text{no closed causal curves}, \tag{3.4}$$

where $\geq$ denotes that the causality condition on the left is stronger than that on the right.

Furthermore, a spacetime (M, g_{ab}) is **strongly asymptotically predictable** if in the unphysical spacetime $(\tilde{M}, \tilde{g}_{ab})$, there is an open region $\tilde{V} \subset \tilde{M}$ with $\overline{M \cap J^-(\mathscr{I}^+)} \subset \tilde{V}$, such that $(\tilde{V}, \tilde{g}_{ab})$ is globally hyperbolic. A strongly asymptotically predictable spacetime is said to contain a black hole if M is not contained in the causal past of the future null infinity, $J^-(\mathscr{I}^+)$.

For an asymptotically flat spacetime (Minkowski spacetime), the black hole region in M is the region $B = M \backslash J^-(\mathscr{I}^+)$ that is not visible to an outside observer (Fig. 4). The boundary of B in M, $H = \partial B = \partial J^-(\mathscr{I}^+) \cap M$, is called the **event horizon**. For asymptotically AdS spacetime (having a timelike asymptotic boundary), the black hole region in M is the region $B = M \backslash J^-(\mathcal{I})$, and the black hole horizon H is again the boundary of B in M, defined by $H = \partial B = \partial J^-(\mathcal{I}) \cap M$, where $\mathcal{I}$ is the worldline of timelike observer at infinity.

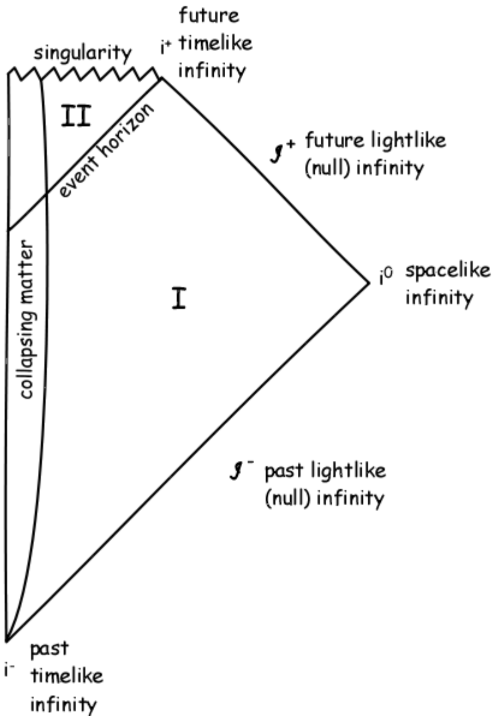

Figure 4: Penrose diagram for the formation of a black hole in 4-dimensional Minkowski spacetime. Figure is adapted from *A Better Picture of Black Holes*.

Part II

Singularities

As discussed previously, in a globally hyperbolic spacetime, the causal diamond D_q^p is compact and the timelike geodesic from q to p maximises the proper time. Why should we care about this causality condition? The answer can be found by investigating the singularities in General Relativity.

4 Geodesics and Focal Points

In Riemannian geometry, is a geodesic the shortest distance between two points? For a sufficiently short geodesic, the answer is always "yes". However, in general if one follows a geodesic for too far, it is no longer length minimising since it has passed a *focal point*.

Example (Two-sphere with its round metric)**.** *A geodesic between two points q and p that goes less than half way around the sphere is the unique shortest path between those two points (Fig. 5(a)). But any geodesic that leaves q and goes more than half way around the sphere is no longer length minimising. In Fig. 5(b) for the case that q is the north pole N. The geodesics that emanate from N initially separate, but after going half way around the sphere, they reconverge at the south pole S. The point of reconvergence is called a* ***focal point*** *or a* ***conjugate point****. A geodesic that is continued past a focal point and thus has gone more than half way around the sphere is no longer length minimising. By "slipping it around the sphere," one can replace it with a shorter geodesic between the same two points that goes around the same great circle on the sphere in the opposite direction.*

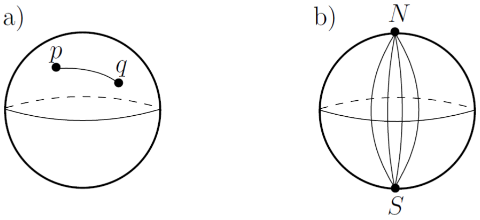

Figure 5: (a) A geodesic between two points p and q in a two-sphere that goes less than half way around the sphere minimises the length between those two points. (b) Any geodesic through the north pole N reaches the south pole S after going half-way around the sphere; the point S is called a focal point for these geodesics. The geodesics from N to S are lines of constant longitude, as drawn. When continued more than half-way around the sphere, these geodesics are no longer length minimising.

In fact, the phenomenon of a geodesic being not length minimising does not depend on any details of the sphere.

Example (Riemannian manifold)**.** *Consider a geodesic qp that originates at a point q in some Riemannian manifold M (Fig. 6). If the geodesic segment qq' can be deformed – at least to first order – to a nearby geodesic connecting the same two points, then q' is called a* ***focal point*** *for geodesics that emanate from q. This displaced geodesic automatically has the same length as the first one since geodesics are stationary points of the length function. Then the displaced path $qq'p$ has a "kink" and its length can be reduced by rounding out the kink. Quantitatively, if the displaced path $qq'p$ "bends" by a small angle at the point q, then there exists a shorter path rs by the triangle inequality in Euclidean space (Fig. 7). So the original geodesic qp was not length minimising.*

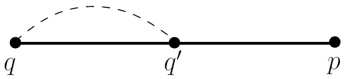

Figure 6: The qq' part of the geodesic qp can be deformed slightly to another nearby geodesic that also connects the two points q and q'.

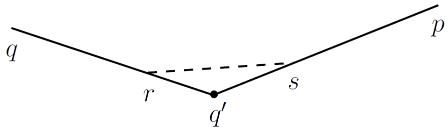

Figure 7: By bending the displaced path $qq'p$ by a small angle at point q, we emphasise that this path is made of two geodesic segments qq' and $q'p$ that meet at a "kink." A small neighborhood of q', containing the nearby points r and s, can be approximated by a portion of flat Euclidean space. The triangle inequality of Euclidean space says that the portion $rq's$ of $qq'p$ can be shortened by replacing it by a straight line rs, shown as a dashed line. So rounding out the kink of $qq'p$ reduces its length.

Example (Lorentzian geometry)**.** *A spacelike geodesic in a spacetime of Lorentz signature is never a minimum or a maximum of the length function, since oscillations in spatial directions tend to increase the length and oscillations in the time direction tend to reduce it. Two points at spacelike separation can be separated by an everywhere spacelike path that is arbitrarily short or arbitrarily long.*

For timelike geodesics, we should discuss the elapsed proper time of a geodesic (not the length) and spatial fluctuations tend to reduce it. A sufficiently short segment of any timelike geodesic maximises the elapsed proper time. But if we continue a timelike geodesic past a focal point, it no longer maximises the proper time.

To see this, consider a future-going timelike geodesic that originates at a point q in spacetime, as in the Euclidean signature case. Rounding out the kink will increase the proper time, which is basically the "twin paradox" of Special Relativity (Fig. 8), in the same sense that the analogous statement in Euclidean signature is the triangle inequality of Euclidean geometry (Fig. 7).

Example (Anti de Sitter (AdS) spacetime)**.** *In the AdS spacetime, future-going timelike geodesics from q meet at the focal point q' (Fig. 9). These geodesics fail to be proper time maximising when continued past q'. For example, the timelike geodesic qw shown in Fig. 9 is not proper time maximising. Indeed, there is no upper bound on the proper time of a causal path from q to w, since a timelike path from q that travels very close to the edge of the figure, lingers there for a while, and then goes on to w can have an arbitrarily large elapsed proper time.*

In summary, a geodesic that emanates from q is no longer length minimising in Euclidean signature (proper time maximising in Lorentzian signature) once it is continued past its first focal point. The absence of a focal point is, however, only a *necessary* condition for a geodesic to be length minimising (proper time maximising), **not** a *sufficient* one. For example, on a torus with a flat metric, geodesics have no focal points no matter how far they are extended. On the other hand, any two points on the torus can be connected by infinitely many different (and homotopically inequivalent) geodesics. Most of those geodesics are not length minimising.

Remark. *A useful statement in Hawking's singularity theorem (we will discuss later) is as follows. Let (M, g_{ab}) be a globally hyperbolic spacetime, and let S be a Cauchy surface. Then every point $p \in M$ is connected to S by a causal path of maximal proper time. Such a path is a timelike geodesic orthogonal to S without focal points.*

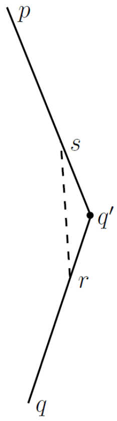

Figure 8: The displaced path $qq'p$ is drawn in a way that emphasises that the segments qq' and $q'p$ are timelike geodesics (Time runs vertically here). A small neighborhood of q' containing the points r and s can be approximated by a portion of Minkowski space. The "twin paradox" of Special Relativity says that the proper time elapsed along the portion $rq's$ of $qq'p$ is less than the proper time elapsed along the geodesic rs, which is shown as a dashed line. (In other words, the twin who takes a trip on the worldline $rq's$ comes back younger than the twin who stays home on the worldline rs.) Thus, the proper time of $qq'p$ can be increased by rounding out the kink, replacing $rq's$ with rs.

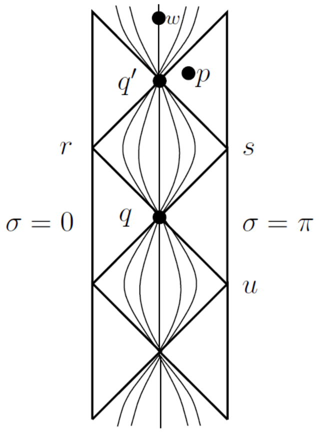

Figure 9: The Penrose diagram of AdS$_2$ spacetime. The causal structure is that of the strip $0 < \sigma < \pi$ in Minkowski spacetime with metric $ds^2 = -dt^2 + d\sigma^2$. Causal curves make an angle no greater than $\pi/4$ from the vertical. A causal curve from q can travel to the right edge of the figure, linger for a while very near the boundary, and then proceed on to p. A causal curve of this kind can have an arbitrarily large elapsed proper time. By contrast, if p is replaced by a point p' inside the quadrilateral $qsq'r$, then a causal curve from q to p' cannot reach the boundary and has a maximum possible elapsed proper time. The curved lines in the figure are the timelike geodesics through the point q; every point p' inside the quadrilateral $qsq'r$ is on such a geodesic (reflecting compactness of $D_q^{p'}$ for such p'). These timelike geodesics all focus at the point q', as shown. Their continuation to the future of q' or the past of q is indicated. No geodesic from q reaches p. There is a timelike geodesic from q to w, but it does not maximise the proper time.

5 Energy Conditions

In General Relativity, we only want to consider "physical matter" which means the stress tensor $T^{\alpha\beta}$ should satisfy some conditions. We define the stress tensor in $(d+1)$-dimensional spacetime as $T^{\alpha\beta} = \rho e_0^\alpha e_0^\beta + \sum_{i=1}^d p_i e_i^\alpha e_i^\beta$.

Name	**Statement**	**Condition**
Weak	$T_{\alpha\beta}V^\alpha V^\beta \geq 0, \forall$ future-directed timelike vector V^α	$\rho \geq 0,\ \rho + p_i \geq 0$
Null	$T_{\alpha\beta}k^\alpha k^\beta \geq 0, \forall$ future-directed lightlike vector k^α	$\rho + p_i \geq 0$
Strong	$\left(T_{\alpha\beta} - \frac{1}{2}Tg_{\alpha\beta}\right)V^\alpha V^\beta \geq 0, \forall$ future-directed timelike vector V^α	$\rho + \Sigma_i p_i \geq 0,\ \rho + p_i \geq 0$
Dominant	$-T_{\alpha\beta}V^\beta$ future-directed, non-spacelike ($T_{\alpha\beta}V^\alpha V^\beta \geq 0$), $\forall$ future-directed timelike vector V^α	$\rho \geq 0,\ \rho \geq \lvert p_i \rvert$

Weak energy condition (WEC): Energy density $T_{\alpha\beta}V^\alpha V^\beta$ of any matter distribution, as measured by any observer in spacetime whose 4-velocity is V^α, must be nonnegative, i.e. $T_{\alpha\beta}V^\alpha V^\beta \geq 0$, for any normalised, future-directed, timelike vector V^α. We can write, in orthogonal basis,

$$V^\alpha = \gamma\left(e_0^\alpha + a e_1^\alpha + b e_2^\alpha + c e_3^\alpha\right), \tag{5.1}$$

where $\gamma = \left(1 - a^2 - b^2 - c^2\right)^{-1/2}$, $a^2 + b^2 + c^2 < 1$ for arbitrary constants a, b, and c. Using the 4-dimensional stress tensor

$$T^{\alpha\beta} = \rho e_0^\alpha e_0^\beta + \sum_{i=1}^3 p_i e_i^\alpha e_i^\beta, \tag{5.2}$$

the WEC gives

$$\rho + a^2 p_1 + b^2 p_2 + c^2 p_3 \geq 0, \quad \forall a, b, c, \tag{5.3}$$

which will be satisfied if and only if

$$\rho \geq 0, \quad \rho + p_i \geq 0. \tag{5.4}$$

Null energy condition (NEC): Energy density $T_{\alpha\beta}k^\alpha k^\beta$ of any matter distribution, as measured by any future-directed null vector k^α, is nonnegative, i.e. $T_{\alpha\beta}k^\alpha k^\beta \geq 0$. Let

$$k^\alpha = e_0^\alpha + a' e_1^\alpha + b' e_2^\alpha + c' e_3^\alpha, \tag{5.5}$$

where $a'^2 + b'^2 + c'^2 = 1$ for arbitrary constants a', b', and c'. Then the NEC gives

$$\rho + a'^2 p_1 + b'^2 p_2 + c'^2 p_3 \geq 0, \quad \forall a', b', c', \tag{5.6}$$

which will be satisfied if and only if

$$\rho + p_i \geq 0 \quad (i = 1, 2, 3). \tag{5.7}$$

So, WEC implies NEC.

Strong energy condition (SEC): Define $T = g^{\alpha\beta}T_{\alpha\beta}$. Then for all normalised, future-directed, timelike vector V^α,

$$\left(T_{\alpha\beta} - \frac{1}{2}Tg_{\alpha\beta}\right)V^\alpha V^\beta \geq 0, \tag{5.8}$$

which is equivalent to

$$T_{\alpha\beta}V^{\alpha}V^{\beta} \geq -\frac{1}{2}T. \tag{5.9}$$

This ensures that the left-hand side of the Einstein equation $T_{\alpha\beta} - \frac{1}{2}g_{\alpha\beta}T = \frac{R_{\alpha\beta}}{8\pi}$ does not become negative, which means that $R_{\alpha\beta}V^{\alpha}V^{\beta} \geq 0$ or "gravity is attractive." Expressing in orthogonal basis, we have

$$\gamma^2\left(\rho + a^2p_1 + b^2p_2 + c^2p_3\right) \geq \frac{1}{2}\left(\rho - p_1 - p_2 - p_3\right). \tag{5.10}$$

For $a = b = c = 0, \gamma = 1$, we find $\rho + p_1 + p_2 + p_3 \geq 0$. For $b = c = 0, \gamma^2 = 1/(1 - a^2)$, we have $\rho + p_1 + p_2 + p_3 \geq a^2\left(p_2 + p_3 - \rho - p_1\right), \; \forall a^2 < 1$, which implies that $\rho + p_1 \geq 0$. So,

$$\rho + \Sigma_{i=1}^{3}p_i \geq 0, \quad \rho + p_i \geq 0 \quad (i = 1, 2, 3). \tag{5.11}$$

Note that SEC does **not** imply WEC, but SEC implies NEC.

Dominant energy condition (DEC): Matter should flow along timelike or null vector field. $T^{\alpha}{}_{\beta}V^{\beta}$ is a future-directed, timelike or null, vector field for an arbitrary future-directed timelike vector field V^{α}. This means

$$\rho^2 - a^2p_1^2 - b^2p_2^2 - c^2p_3^2 \geq 0 \tag{5.12}$$

which can only be satisfied if and only if

$$\rho \geq 0, \quad \rho \geq |p_i| \quad (i = 1, 2, 3). \tag{5.13}$$

We have DEC $\Rightarrow$ WEC $\Rightarrow$ NEC, SEC $\Rightarrow$ NEC. But all these conditions can be violated!

There exists another version of energy condition on $T^{\alpha\beta}$.
Average null energy condition (ANEC): $\int_{\gamma} T_{\alpha\beta}k^{\alpha}k^{\beta}d\lambda \geq 0$ for a null geodesic γ. There are lots of gravitational theorems that can be proved with ANEC. It is known to hold in flat spacetime, and non-interacting scalar and electromagnetic fields in arbitrary states. The proof of ANEC is a work in progress. A further discussion on ANEC can be found in Section 13.

6 Raychaudhuri Equation

Congruence of timelike geodesics

Let M be a manifold and let $O \subset M$ be an open set. A **congruence** in O is a family of curves such that through each point $p \in O$ there passes one and only one curve in the family. Thus, the tangents to a congruence yield a vector field in O. Conversely, every (smooth) continuous vector field generates a congruence of (smooth) curves. Note that the curves do not intersect (Fig. 10).

The timelike congruence is parameterised by the proper time τ with the vector field u^{α} of tangents satisfying the conditions:

$$u^{\alpha}u_{\alpha} = -1, \quad u^{\alpha}{}_{;\beta}u^{\beta} = 0, \quad u^{\alpha}\eta_{\alpha} = 0 \tag{6.1}$$

for an arbitrary (orthogonal) deviation vector η^{α}. The geodesic deviation equation is given by

$$\frac{D^2\eta^{\alpha}}{d\tau^2} = -R^{\alpha}{}_{\beta\rho\sigma}u^{\beta}\eta^{\rho}u^{\sigma}, \tag{6.2}$$

where D is the covariant derivative operator.

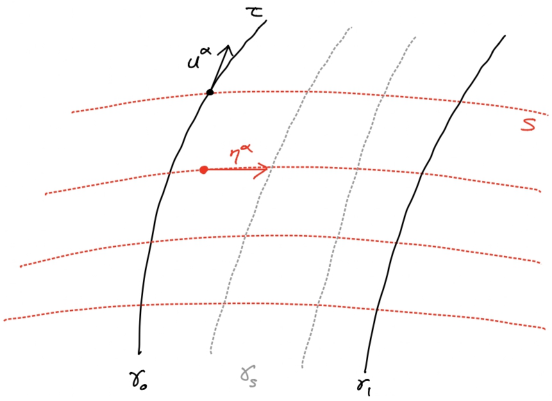

Figure 10: Congruence of timelike geodesics.

Define a tensor field,

$$B_{\alpha\beta} = \nabla_\alpha u_\beta, \tag{6.3}$$

which is spatial:

$$B_{\alpha\beta}u^\alpha = u^\alpha \nabla_\alpha u^\beta = 0, \quad B_{\alpha\beta}u^\beta = \nabla_\alpha(u^2) = 0. \tag{6.4}$$

The *physical interpretation* of $B_{\alpha\beta}$ may be seen from the following considerations. Consider a smooth one-parameter subfamily $\gamma_S(\tau)$ of geodesics in the congruence. We have the deviation vector η^α from γ_0 for this subfamily, representing an infinitesimal spatial displacement from γ_0 to a nearby geodesic in the subfamily. The Lie derivatives of the vector fields are $\mathcal{L}_u \eta^\alpha = \mathcal{L}_\eta u^\alpha = 0$, which imply that $\eta^\alpha{}_{;\beta}u^\beta = u^\alpha{}_{;\beta}\eta^\beta = B^\alpha{}_\beta \eta^\beta$, where $B^\alpha{}_\beta$ measures the failure of η^α to be parallelly transported along the geodesics.

Define a spatial metric on the surface S perpendicular to u^α ($h_{\alpha\beta}u^\beta = 0$) by

$$h_{ab} \coloneqq g_{ab} + u_a u_b. \tag{6.5}$$

Then we can decompose the tensor field B_{ab} in d spatial dimensions as

$$B_{ab} = \frac{\theta}{d} h_{ab} + \sigma_{ab} + \omega_{ab}, \tag{6.6}$$

where $\theta = B^{ab}h_{ab}$ is the **expansion**, $\sigma_{ab} = B_{(ab)} - \frac{1}{d}\theta h_{ab}$ is the **shear**, and $\omega_{ab} = B_{[ab]}$ is the **twist or rotation** of the congruence. If the expansion is positive, $\theta > 0$, the congruence will be diverging, while if it is negative, $\theta < 0$, the congruence will be converging. By virtue of (6.4) and **Frobenius theorem**, the congruence is (locally) hypersurface orthogonal if and only if $u_{[\alpha;\beta}u_{\gamma]} = 0$. This is equivalent to saying that $\omega_{\alpha\beta} = u_{[\alpha;\beta]} = 0$ for timelike geodesics.

Raychaudhuri equation

Now we would like to derive the equations for the rate of change of the expansion, shear, and twist of timelike geodesics, as one moves along the curves in the congruence. We have

$$\begin{aligned} B_{\mu\nu;\alpha}u^\alpha &= \left(u_{\mu;\nu}\right)_{;\alpha} u^\alpha = \left(u_{\mu;\alpha}\right)_{;\nu} u^\alpha + R_{\mu\alpha\nu\beta}u^\alpha u^\beta \\ &= \left(u_{\mu;\alpha}u^\alpha\right)_{;\nu} - u_{\mu;\alpha}u^\alpha{}_{;\nu} + R_{\mu\alpha\nu\beta}u^\alpha u^\beta \\ &= -B_{\mu\alpha}B^\alpha{}_\nu - R_{\mu\alpha\nu\beta}u^\alpha u^\beta. \end{aligned} \tag{6.7}$$

Taking the trace, we obtain

$$\frac{d\theta}{d\tau} = -B^{\alpha\beta}B_{\beta\alpha} - R_{\alpha\beta}u^{\alpha}u^{\beta}, \tag{6.8}$$

which can be written explicitly as

$$\frac{d\theta}{d\tau} = -\frac{1}{d}\theta^2 - \sigma^{\alpha\beta}\sigma_{\alpha\beta} + \omega^{\alpha\beta}\omega_{\alpha\beta} - R_{\alpha\beta}u^{\alpha}u^{\beta}. \tag{6.9}$$

This is known as **Raychaudhuri equation** and is the key equation used in the proof of the (Hawking's) singularity theorem.

(Classical) focussing theorem

Under the following three assumptions:

1. in a congruence of timelike *hypersurface orthogonal* geodesics, $\omega_{ab} = 0$,
2. SEC ($T_{ab}u^a u^b \geq 0$) and Einstein equation imply that $-R_{ab}u^a u^b \leq 0$, which means physically that gravity is attractive and geodesics get focussed as a result of this attraction,
3. $-\sigma^{ab}\sigma_{ab}$ is manifestly nonpositive,

we find that (6.9) reduces to

$$\frac{d\theta}{d\tau} \leq -\frac{1}{d}\theta^2 \;\;\Rightarrow\;\; \theta^{-1}(\tau) \geq \theta_0^{-1} + \frac{\tau}{d}, \quad \theta_0 = \theta(0). \tag{6.10}$$

If $\theta_0 < 0$, then $\theta(\tau) \to -\infty$ within a proper time $\tau \leq d/\left|\theta_0\right|$. Here the congruence develops a **caustic** (focal point/conjugate point), which is a singularity of the congruence.

Example ((Flat) FLRW cosmological model)**.** *The metric is*

$$ds^2 = -dt^2 + a^2(t)d\vec{x}^2. \tag{6.11}$$

The Hubble parameter $H = \frac{\dot{a}}{a}$ is observed to be positive now. Define $V = a^d$, then the expansion is $\theta = \frac{\dot{V}}{V} = d \cdot H > 0$. By Raychaudhuri equation, $\theta \to -\infty$ in the past at finite proper time — there was Big Bang! An open question is, what if our universe is not perfectly homogeneous and isotropic? Will there still be a big bang singularity?

Example (Initial value surface)**.** *Consider a spacetime M with an initial value surface S with local coordinates $\vec{x} = (x^1, \cdots, x^d)$ (Fig. 11). By looking at timelike geodesics orthogonal to S, we can construct a coordinate system in a neighborhood of S. If a point q is on a timelike geodesic that meets S orthogonally at $\vec{x}$, and the proper time from S to q (measured along the geodesic) is t, then we assign to q the coordinates $(t, \vec{x})$ if q is to the future of S, or $(-t, \vec{x})$ if it is to the past.*

The line element of M is

$$ds^2 = -dt^2 + h_{ij}(t, \vec{x})dx^i dx^j. \tag{6.12}$$

Since $h_{ij}(t, \vec{x})$ measures the distance between nearby orthogonal geodesics, a necessary and sufficient criterion for a focal point is $\det h_{ij}(t, \vec{x}) = 0$. Raychaudhuri equation gives a useful criterion for predicting that $\det h_{ij}$ will go to 0 within a known time. Note that in general, this will represent only a breakdown of the coordinate system, not a true spacetime singularity.

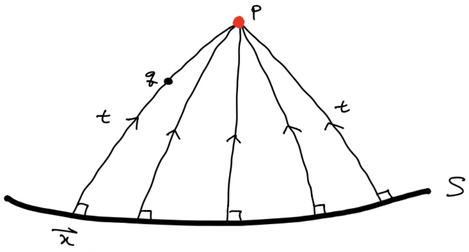

Figure 11: If orthogonal geodesics from a spacelike hypersurface S are focussed at point p to the future of S, then the coordinate system based on the orthogonal geodesics will break down at p.

Raychaudhuri's equation is just the Einstein equation

$$R_{tt} = 8\pi G_N \left(T_{tt} - \frac{1}{d-1} g_{tt} T^{\alpha}{}_{\alpha} \right) \tag{6.13}$$

in the coordinate system defined by the orthogonal geodesics. A straightforward computation in the metric (6.12) shows that

$$\begin{aligned} R_{tt} &= -\partial_t \Gamma^i{}_{ti} - \Gamma^i{}_{tj}\Gamma^j{}_{ti} = -\frac{1}{2}\partial_t \left(h^{ik}\partial_t h_{ik} \right) - \frac{1}{4}\left(h^{ik}\partial_t h_{kj} \right) \left(h^{jm}\partial_t h_{mi} \right) \\ &= -\frac{1}{2}\partial_t \operatorname{Tr}\left(h^{-1}\dot{h} \right) - \frac{1}{4}\operatorname{Tr}\left(h^{-1}\dot{h} \right)^2, \end{aligned} \tag{6.14}$$

where the dot represents a derivative with respect to t.

It is convenient to define

$$V = \sqrt{\det h}, \tag{6.15}$$

which measures the volume occupied by a little bundle of geodesics. The quantity

$$\theta = \frac{\dot{V}}{V} = \frac{1}{2}\operatorname{Tr}\left(h^{-1}\dot{h} \right) \tag{6.16}$$

is the expansion. The traceless part of $h^{-1}\dot{h}$, defined by

$$\sigma^i{}_j = \frac{1}{2}\left(h^{ik}\dot{h}_{kj} - \frac{1}{d}\delta^i{}_j \operatorname{Tr} h^{-1}\dot{h} \right), \tag{6.17}$$

is the shear, where the factor of $1/2$ is conventional. So

$$R_{tt} = -\partial_t \left(\frac{\dot{V}}{V} \right) - \frac{1}{d}\left(\frac{\dot{V}}{V} \right)^2 - \operatorname{Tr}\sigma^2 = -\dot{\theta} - \frac{\theta^2}{d} - \operatorname{Tr}\sigma^2. \tag{6.18}$$

Recall if $\theta < 0$ at some point, then

$$\begin{aligned} \frac{d\theta}{dt} \le -\frac{1}{d}\theta^2 \;\; &\Rightarrow \;\; \left(\frac{\dot{V}}{V} \right)^{-1} = \theta^{-1} \ge \theta_0^{-1} + \frac{t}{d} \\ &\Rightarrow \;\; \log\left(\frac{V(t)}{V(0)} \right) \le d \log \frac{\theta_0^{-1} + t/d}{\theta_0^{-1}}, \end{aligned} \tag{6.19}$$

showing that $\log V(t) \to -\infty$ and thus $V(t) \to 0$ at a time no later than $t = -d/\theta_0$.

In many situations the vanishing of V predicted by the Raychaudhuri equation represents only a focal point, a breakdown of the coordinate system, and **not** a spacetime singularity. To predict a spacetime singularity requires a more precise argument, with some input beyond Raychaudhuri's equation.

7 Hawking's Big Bang Singularity Theorem

In proving a singularity theorem, Hawking made the following assumptions:

1. the universe is globally hyperbolic with Cauchy surface S,
2. strong energy condition (SEC),
3. the local Hubble parameter of our universe is everywhere positive (local Hubble parameter, for time flowing to the future, is $H = \frac{\dot{V}}{d \cdot V} = \frac{\theta}{d} \geq h_{\text{min}}$; for time flowing to the past, $\frac{\dot{V}}{d \cdot V} \leq -h_{\text{min}}$).

Note that Raychaudhuri equation and SEC imply that $V \to 0$ in the finite past. Hawking's proof of the existence of Big Bang singularity consists of comparing two statements.

1. Since the universe is globally hyperbolic, every point p is connected to S by a causal path of maximal proper time. Such a path is a timelike geodesic without focal points that is orthogonal to S.
2. But the assumption that the initial value of $\frac{\dot{V}}{V}$ on the surface S is everywhere $\leq -d \cdot h_{\text{min}}$ implies that any past-going timelike geodesic orthogonal to S develops a focal point within a proper time at most $1/h_{\text{min}}$.

Combining the two statements, we see that there is no point in spacetime that is to the past of S by a proper time more than $1/h_{\text{min}}$, along any causal path. Thus (given Hawking's assumptions) the minimum value of the local Hubble parameter, h_{min}, gives an upper bound on how long anything in the universe could have existed in the past. This is **Hawking's theorem** about the Big Bang singularity.

An alternative statement of Hawking's theorem is that no timelike geodesic γ from S can be continued into the past for a proper time greater than $1/h_{\text{min}}$. Otherwise, there would be a point $p \in \gamma$ that is to the past of S by a proper time measured along γ that is greater than $1/h_{\text{min}}$, contradicting what was just proved.

Although Hawking's theorem is generally regarded as a statement about the Big Bang singularity, singularities are not directly involved in the statement or proof of the theorem. In fact, to the present day, one has only a limited understanding of the implications of Einstein's equations concerning singularities. In the classic singularity theorems, going back to Penrose, only the smooth part of spacetime is studied, or to put it differently, "spacetime" is taken to be, by definition, a manifold with a smooth metric of Lorentz signature. Then "singularity theorems" are really statements about *geodesic incompleteness* of spacetime.

8 Promptness and Null Geodesics

Having discussed timelike geodesics, let us now have an analogous study of the null geodesics. These are needed for applications such as Penrose's singularity theorem and an understanding of black holes.

Causal paths and in particular null geodesics will be assumed to be future-going unless otherwise specified. Of course, similar statements apply to past-going causal paths, with the roles of the future and the past exchanged.

Promptness

Any null geodesic has zero elapsed proper time. Nevertheless, there is a good notion that has properties somewhat similar to "maximal elapsed proper time" for timelike geodesics with no focal points. A causal path from q to p is said to be **prompt** if no causal path from q to p arrives sooner. To be precise, the path γ from q to p is prompt if there is no causal path γ' from q to r, a point to the past of p (Fig. 12). In other words, for a causal path to be prompt, it has to be an achronal null geodesic with no focal points.

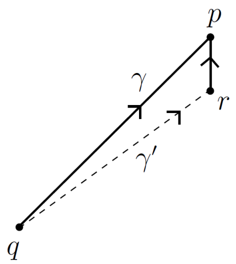

Figure 12: Here time runs vertically, so r is to the past of p. A causal path γ from q to p is "prompt" if there does not exist any causal path γ' from q to a point such as r that is to the past of p.

Properties:

1. A prompt causal path from q to p can only exist if $p \in \partial J^+(q) = J^+(q) - I^+(q)$.
2. A prompt causal path has to be a null geodesic.
3. With SEC, short enough null geodesics are prompt.
4. A prompt segment of a causal path can only be the initial segment.
5. A prompt causal path must contain no focal point.
6. A null geodesic is prompt if and only if it is achronal.

So far we have discussed prompt causal paths between two points. More generally, if $W \subset M$ is any set in spacetime M, we say that a future-going causal path from W to $p \in M$ is **prompt** if it arrives as soon as possible, in the sense that no causal path from W arrives to the past of p.

Properties:

1. A prompt causal path ℓ has to be an achronal null geodesic.
2. If W is a spacelike submanifold of spacetime, then ℓ has to be orthogonal to W at the intersection point (Fig. 13). The orthogonality is only possible if W has real codimension at least 2.
3. ℓ must have no focal point.

Boundary of the future

$J^+(W)$ is closed if W is compact and M is globally hyperbolic. For any W, $\partial J^+(W)$ is always achronal (Fig. 14). If $p \in \partial J^+(W)$, then p can be reached from W by a future-going prompt causal path and more specifically, a null geodesic without focal points.

If $W \subset M$ is compact, $\partial J^+(W)$ is always a codimension 1 submanifold in M, though generally not smooth (Fig. 15). Moreover, if M is globally hyperbolic, then $\partial J^+(W)$ is closed in M, so it is actually a closed submanifold of M. The topology of $\partial J^+(W)$ is equivalent topologically to the initial value surface $t = 0$. In particular, $\partial J^+(W)$ is never compact and Minkowski space has no compact achronal hypersurface.

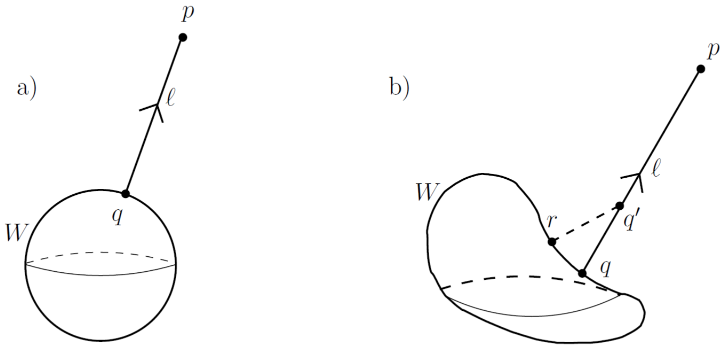

Figure 13: (a) ℓ is a null geodesic from a spacelike submanifold W to a point $p \in M$. For ℓ to be prompt, it must be orthogonal to W at the point q at which it departs W. Otherwise, by moving q slightly "towards" p, one can replace ℓ with another causal curve to p that gets a "head start" and arrives in the past of p. (b) Even if orthogonal to W, ℓ is nonetheless not prompt if a segment qq' of ℓ can be displaced to a nearby null geodesic, meeting W orthogonally at some other point r. This displacement gives a causal path from W to p that is not a null geodesic, so it can be modified to a causal path that is more prompt. In this situation, we say that the point q' is a focal point of the orthogonal null geodesics from W. This situation does not arise in Minkowski space if W is a round two-sphere supported at time $t = 0$ and the geodesic ℓ is "outgoing," but it can arise for a more general choice of W, as shown, or for a round sphere embedded in a more general spacetime, or for "incoming" geodesics from a round sphere in Minkowski space.

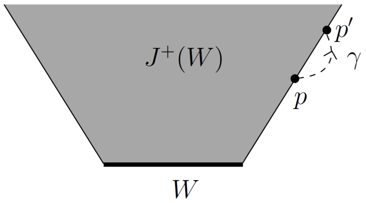

Figure 14: For any subset W of spacetime, $\partial J^+(W)$ is always achronal. For if there is a future-going timelike path γ from p to p' with $p, p' \in \partial J^+(W)$, then a neighborhood of p' can be reached from p, and therefore from W, by a timelike path. So p' is in the interior of $J^+(W)$, not in its boundary $\partial J^+(W)$. Thus in the diagram, the curve γ cannot be everywhere timelike. If γ is a causal curve from p to p', it must be a null geodesic entirely contained in $\partial J^+(W)$.

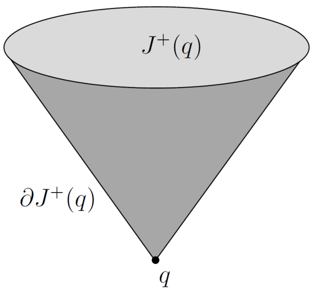

Figure 15: For a point q in Minkowski space, $\partial J^+(q)$ consists of the points on the future light cone, including the vertex of the cone. It is a closed submanifold of Minkowski space but (because of the conical singularity at q) it is not smoothly embedded in Minkowski space. Here we consider q itself to be contained in the causal future $J^+(q)$ and in its boundary $\partial J^+(q)$.

Congruence of null geodesics

Just as in the case of timelike geodesics, we define a null vector field k^α which parameterises the null congruence and have an orthogonal deviation vector field η^α. These vector fields obey the similar conditions:

$$k^\alpha k_\alpha = 0, \quad k^\alpha{}_{;\beta} k^\beta = 0, \quad k^\alpha{}_{;\beta}\eta^\beta = 0, \quad k^\alpha \eta_\alpha = 0. \tag{8.1}$$

Define the transverse metric by

$$h_{\alpha\beta} \coloneqq g_{\alpha\beta} + k_\alpha N_\beta + k_\beta N_\alpha \tag{8.2}$$

with N^α the auxiliary null vector field which is not uniquely defined. It satisfies the conditions:

$$k_\alpha N^\alpha = -1, \quad N_\alpha N^\alpha = 0. \tag{8.3}$$

Then we see that $h_{\alpha\beta}$ is spatial, where

$$h_{\alpha\beta}k^\alpha = h_{\alpha\beta}N^\alpha = 0, \quad h^\alpha{}_\alpha = g^{\alpha\beta}h_{\alpha\beta} = d-1, \quad h^\alpha{}_\mu h^\mu{}_\alpha = h^\alpha{}_\beta. \tag{8.4}$$

The deviation vector is no longer parallelly transported:

$$\eta^\alpha{}_{;\beta}k^\beta = B^\alpha{}_\beta \eta^\beta, \tag{8.5}$$

where $B_{\alpha\beta} = k_{\alpha;\beta}$. It is transverse to k^α ($B_{\alpha\beta}k^\beta = B_{\beta\alpha}k^\alpha = 0$), but not to N^α. Define a tensor field $\tilde{B}_{\alpha\beta} = h^\mu{}_\alpha h^\nu{}_\beta B_{\mu\nu} = h^\mu{}_\alpha h^\nu{}_\beta k_{\mu;\nu}$, which is purely spatial and can be decomposed as

$$\tilde{B}_{\alpha\beta} = \frac{1}{2}\theta h_{\alpha\beta} + \sigma_{\alpha\beta} + \omega_{\alpha\beta}. \tag{8.6}$$

Here $\theta = g^{\alpha\beta}\tilde{B}_{\alpha\beta} = g^{\alpha\beta}B_{\alpha\beta} = k^\alpha{}_{;\alpha}$ is the expansion, that is independent of the choice of N. σ_{ab} is the shear, and ω_{ab} is the twist or rotation of the null congruence. We also define the tensor field $k_{\alpha\beta} = \nabla_\beta \xi_\alpha = \frac{1}{2}\mathcal{L}_\xi h_{\alpha\beta} = \frac{1}{2}\mathcal{L}_\xi g_{ab}$.

Example. *Consider a lightcone from point q (Fig. 16). The metric is $ds^2 = -dt^2 + dr^2 + r^2 d\Omega^2$. We have the Killing vector $\hat{k} = \partial_t - \partial_r$ and the expansion is*

$$\theta = k^\alpha{}_{;\alpha} = \frac{2}{r} = \frac{1}{4\pi r^2}\frac{d(4\pi r^2)}{d\lambda} = \frac{1}{A}\frac{dA}{d\lambda}. \tag{8.7}$$

The coordinates t and r are defined as $t = \lambda, r = \lambda$, where λ is the affine parameter.

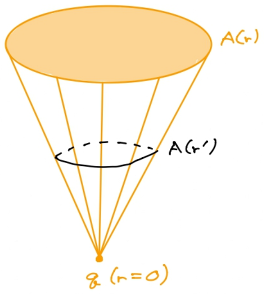

Figure 16: The expansion is the rate of change of the area of the surface swept out by different r's in the lightcone.

Example. *Let W be a codimension two spacelike submanifold of a spacetime M. If M is Minkowski space, then W might be a sphere, embedded in Minkowski space with line element $ds^2 = -dt^2 + d\vec{x}^2$ in the standard fashion, as the submanifold $\vec{x}^2 = R^2, t = 0$. Emanating from W are two families of future-going null geodesics orthogonal to W (Fig. 17), which can be naturally called as "outgoing" and "incoming." In what follows, we focus on the outgoing family.*

The outgoing orthogonal null geodesics emanating from W sweep out a codimension 1 submanifold, Y. We may pick any set of local coordinates x^A, $A = 1, \cdots, d-1$ on W. We have also an affine parameter of a null geodesic, denoted λ, as the additional coordinate (this is similar to the proper time measured along a timelike geodesic, but there is no notion of proper time for a null geodesic). The geodesic equation reads

$$\frac{D^2 x^\mu}{d\lambda^2} = 0 \tag{8.8}$$

and is invariant under affine transformations

$$\lambda \to a\lambda + b, \quad a, b \in \mathbb{R}. \tag{8.9}$$

On each of the null geodesics ℓ that make up Y, we pick an affine outgoing coordinate u that increases towards the future and vanishes at the point $\ell \cap W$. This uniquely determines u up to the possibility of multiplying it by a positive constant. On a neighbourhood of Y, x^A and the affine ingoing coordinate v are constant along the outgoing null geodesics. The geodesic equation is then satisfied with $u = \lambda$ with v and x^A constant.

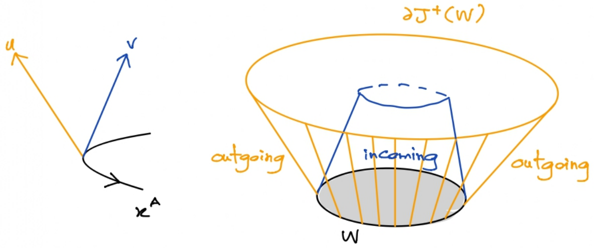

Figure 17: If W is a spacelike submanifold of spacetime of codimension 2, then two families of future-going null geodesics ("outgoing" and "incoming") emanate from W. The two families sweep out two different null hypersurfaces.

Note that, in general, the expansion is coordinate-dependent, i.e.

$$\theta(\vec{x}) = \frac{1}{\delta A(\vec{x})} \frac{d\delta A(\vec{x})}{d\lambda}. \tag{8.10}$$

Null Raychaudhuri equation

In a similar fashion as for the timelike geodesics, the Raychaudhuri equation for null geodesics is

$$\frac{d\theta}{d\lambda} = -\frac{1}{2}\theta^2 - \sigma^{\alpha\beta}\sigma_{\alpha\beta} + \omega^{\alpha\beta}\omega_{\alpha\beta} - R_{\alpha\beta}k^\alpha k^\beta. \tag{8.11}$$

Here the congruence of null geodesics is hypersurface orthogonal if and only if $\omega_{\alpha\beta} = 0$. The NEC ($T_{\alpha\beta}k^\alpha k^\beta \geq 0$) with Einstein equation imply that $R_{ab}k^a k^b \geq 0$. Then the focussing theorem states that

$$\frac{d\theta}{d\lambda} \leq -\frac{\theta^2}{d-1} \leq 0. \tag{8.12}$$

If $\theta = \theta_0 < 0$ at some point, then

$$\theta = \frac{\dot{A}}{A} \leq \left(\frac{1}{\theta_0} + \frac{\lambda}{d-1} \right)^{-1}. \tag{8.13}$$

Trapped surface

Penrose defined a **trapped surface** to be a codimension two, achronal, spacelike submanifold W such that the expansion of each family of orthogonal future-going null geodesics is everywhere negative.

Example. *Consider a spherically symmetric surface behind the horizon of a spherically symmetric black hole. Such a black hole can be conveniently represented by a two-dimensional picture, known as a* ***Penrose diagram****, in which only the radial and time directions are shown (Fig. 19). A point p in the diagram represents a two-sphere W. The future-directed outgoing and incoming radial geodesics that leave p represent future-directed causal paths, so r decreases along each of them. Therefore both null expansions of W are negative; W is a trapped surface.*

Example. *Consider the Vaidya solution:*

$$ds^2 = -\left(1 - \frac{2m(v)}{r}\right) dv^2 + 2dvdr + r^2 d\Omega_2^2. \tag{8.14}$$

The "inward" Killing vector is $k = -\partial_r$ and the expansion is

$$\theta_{in} = \frac{k^\mu \partial_\mu A}{A} \sim -\frac{2}{r} < 0. \tag{8.15}$$

The "outward" Killing vector is $l = \gamma \left[\partial_v + \left(1 - \frac{2m(v)}{r}\right) \partial_r \right]$, where the additional factor γ ensures that the geodesic equation is satisfied. The expansion is then given by

$$\theta_{out} = \frac{l^\mu \partial_\mu A}{A} = \frac{2}{r}\left(1 - \frac{2m(v)}{r}\right)\gamma. \tag{8.16}$$

Here the mass parameter is defined by

$$m(v) = \begin{cases} 0, & v \leq 0 \quad \textit{Minkowski region}, \\ \mu v, & 0 \leq v \leq M/\mu \quad \textit{Vaidya region}, \\ M, & v \geq M/\mu \quad \textit{Schwarzschild region}. \end{cases} \tag{8.17}$$

Trapped surfaces exist if $r < m(v)$. For $\theta \leq 0$, we have the marginally trapped surface. The Penrose diagram for Vaidya spacetime is depicted in Fig. 18.

Solving the geodesic equation, we get $\gamma \propto 1/\left(1 - \frac{2m(v)}{r}\right)$. For future directed null geodesics, $k \cdot l = -\gamma$ implies that $\gamma > 0$ and thus $\gamma = 1/\left|1 - \frac{2m(v)}{r}\right|$. Hence the expansion for future directed outgoing null geodesics is

$$\theta_{out} = \frac{2}{r} \textit{sign} \left(1 - \frac{2m(v)}{r}\right). \tag{8.18}$$

9 Penrose's Theorem

Historically, Penrose's theorem was the first modern singularity theorem; Hawking's singularity theorem came later.

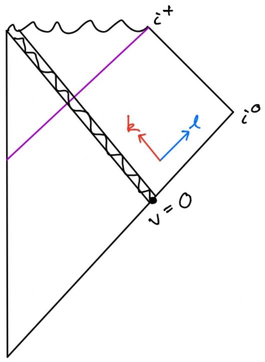

Figure 18: Penrose diagram for Vaidya spacetime.

Penrose's goal was to prove that the formation of a singularity is a generic phenomenon in gravitational collapse, as depicted in Fig. 19. If spherical symmetry is assumed, one can solve Einstein's equations for a collapsing star rather explicitly and show that formation of a singularity is unavoidable. But what happens if the geometry is not quite spherically symmetric? Does infalling matter still collapse to a singularity, or does it "miss" and, perhaps, re-emerge in an explosion? Penrose wanted a robust criterion for formation of a singularity that would not depend on precise spherical symmetry. This goal was the motivation for the concept of a trapped surface.

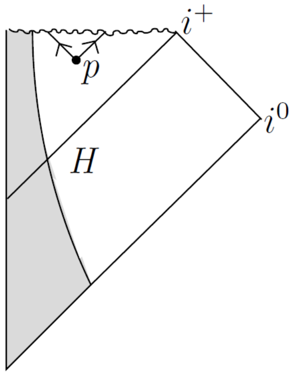

Figure 19: This is a Penrose diagram describing spherically symmetric collapse to a black hole. The area shaded gray represents matter and radiation that is collapsing to form a black hole; the exterior, unshaded region represents part of the Schwarzschild solution. A point in the diagram represents a two-sphere (or in $d+1$ dimensions, a $(d-1)$-sphere) whose radius r is a function of the two coordinates in the diagram. Behind the horizon, r decreases along every future-going causal curve. For example, the point labelled p is behind the horizon and represents a sphere W of area $4\pi r^2$. The future-going orthogonal null geodesics from p are represented in the diagram by ingoing and outgoing null geodesics (the rays emerging from p at $\pi/4$ angles to the vertical). These are future-going causal curves beyond the horizon, so r decreases along each of them and hence the area of W also decreases. Thus, W is a trapped surface. The points in the diagram labeled i^0 and i^+ represent spatial infinity and the future infinity of an outside observer, respectively.

Like the other results that came later, Penrose's singularity theorem is proved without directly studying singularities. It is really a statement about *completeness*. Penrose proved that a spacetime M satisfying certain conditions (assumptions) is null geodesically incomplete, meaning that at least one null geodesic in M cannot be continued in the future to an arbitrarily large value of its affine parameter.

The assumptions made by Penrose in his proof are as follows.

1. The spacetime (M, g) is globally hyperbolic.
2. Noncompact Cauchy surface S (applies to asymptotically flat spacetime).
3. NEC ($T_{\alpha\beta}k^\alpha k^\beta \geq 0$).
4. Classical Einstein equation.
5. The spacetime (M, g) contains a compact trapped surface W (which is stable against perturbations).

Under these assumptions, **Penrose's theorem** states that the spacetime (M, g) is not fully null geodesic complete. The theorem applies to any spacetime that is sufficiently close to Schwarzschild, that is, sufficiently close to spherically symmetric collapse to a black hole in an asymptotically flat spacetime.

The sketch of proof of Penrose's theorem is as follows.

- Suppose on the contrary that every one of the future-going null geodesics ℓ orthogonal to W can be extended to a value of its affine parameter greater than $-(d-1)/\theta$. This means, according to Raychaudhuri equation, that every such ℓ can be extended beyond its first focal point.
- If all null geodesics ℓ from the compact trapped surface W can be extended, it has to go beyond the focal point. The initial segment of ℓ is prompt and is a subset of $\partial J^+(W)$. $\ell \cap \partial J^+(W)$ is closed and compact. (If ℓ cannot be continued until reaching a focal point — for instance because it ends at a singularity or leaves M — then $\ell \cap \partial J^+(W)$ can be noncompact.)
- Every point $p \in \partial J^+(W)$ can be determined by a triple consisting of
  - the choice of intersection point q,
  - the choice of whether ℓ is incoming or outgoing at q, and
  - the affine parameter λ measured along ℓ.

  Since W itself is compact and λ measured along each ℓ ranges over a compact interval, this implies that $\partial J^+(W)$ is compact. (For non-trapped surface, ℓ can be noncompact.)
- On the other hand, $\partial J^+(W)$ is achronal codimension 1 submanifold. In a globally hyperbolic spacetime M with a noncompact Cauchy surface S, there is no compact achronal submanifold of codimension 1.
- Putting this together, our hypothesis was wrong and at least one future-going null geodesic ℓ that is orthogonal to W cannot be extended within M beyond an affine distance $-(d-1)/\theta$ to the future of W.

But what is the fate of incomplete null geodesics? There are two possibilities.

1. The geodesics might terminate on a singularity, as in the case of a Schwarzschild black hole.
2. The geodesics simply leave the globally hyperbolic spacetime M without reaching any singularity. This phenomenon is a *failure of predictivity*: what happens to these geodesics after they leave M cannot be predicted based on initial data on the Cauchy surface S of M.

Part III

Black Holes and Wormholes

10 Cosmic Censorship

What Penrose's theorem actually says about the region inside a black hole is quite limited. With one more assumption, the ideas on which this theorem is based lead rather naturally to an understanding of important properties of black holes. The assumption is the **cosmic censorship hypothesis**. Note that this hypothesis is still an open question in classical General Relativity.

To get a good theory of black holes, one should ensure that something "worse" than collapse to a black hole does not occur. Roughly speaking, "something worse" might be the formation of what Penrose called a **naked singularity**. This is a singularity not enclosed by a horizon, and visible to an outside observer. Formation of a naked singularity might bring the predictive power of classical General Relativity to an end.

Penrose introduced the simplest form of the cosmic censorship hypothesis, which is known as **"weak" cosmic censorship** (it was elaborated later). The hypothesis says that in a globally hyperbolic, asymptotically flat spacetime the evolution seen by an outside observer is supposed to be predictable based on the classical Einstein equations. Any singularity is hidden by a horizon, and does not affect the outside evolution. By now reasonable evidence for cosmic censorship, at least for $D = 4$, has come from the fact that simulations of black hole collisions have not generated naked singularities. This result has also been supported by the LIGO/VIRGO observations of colliding black holes, providing the information that black holes merge to bigger black holes, rather than the formation of naked singularities.

11 Black Hole Region

If cosmic censorship is assumed, one can make a nice theory of black holes in a globally hyperbolic and asymptotically flat spacetime M.

The black hole region in M is the region B that is not visible to an outside observer. To be more exact, the **black hole region** B is the complement of $J^-(\mathcal{I})$ in M: $B = M \backslash J^-(\mathcal{I})$, where $\mathcal{I}$ is the worldline of a timelike observer who remains at rest at a great distance, in the asymptotically flat region observing whatever happens. ($\mathcal{I}$ is actually the future null infinity $\mathscr{I}^+$ in the Penrose diagram of M.) $J^-(\mathcal{I})$ denotes the causal past of the observer, i.e. the set of points from which the observer can receive a signal. $J^-(\mathcal{I})$ is always open and thus B is closed. The **black hole horizon** H is defined to be the boundary of B: $H = \partial B$.

Let us prove that these definitions are sensible by showing that the existence of a black hole region is a generic property of gravitational collapse. We will show that any compact (marginally) trapped surface W is in the black hole region B. In other words, we will show that a signal from a compact (marginally) trapped surface cannot reach the outside observer.

Assume that the last statement is not true, i.e. a causal signal from the compact (marginally) trapped surface W can reach the worldline $\mathcal{I}$ of the distant observer. Then there is a point $p \in \mathcal{I}$ being the earliest point to receive signals from W. The causal path γ from W to p would be prompt, so it would be a future-going null geodesic without focal points, orthogonal to W, and can go very far (Fig. 20). With NEC, since W is a compact (marginally) trapped surface, there is a focal point on γ within a known, bounded affine distance from W. But $\mathcal{I}$, the worldline of the outside observer, can be arbitrarily far away.

This is a contradiction and hence there can be no causal signal from W to $\mathcal{I}$.

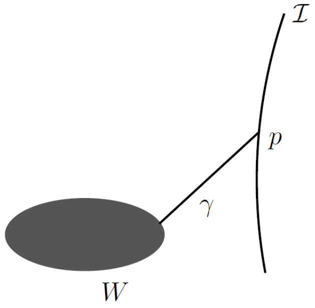

Figure 20: If an observer with worldline $\mathcal{I}$ can receive a signal from a compact set W in spacetime, then the earliest possible such signal arrives on a prompt null geodesic γ that connects W to a point $p \in \mathcal{I}$. If W is a trapped region and $\mathcal{I}$ is sufficiently far away, this is impossible.

If a set W is contained in the black hole region B, then its future $J^+(W)$ is also in B. For if r is in $J^+(W)$ and an event at r can be seen by the distant observer, then that observer can also receive a signal from W (Fig. 21). Now let W be the (marginally) trapped surface or the boundary of a codimension 1 spacelike submanifold Z. A prompt causal path from Z to $\mathcal{I}$ would actually be an orthogonal null geodesic from W to $\mathcal{I}$. But we already know that no such geodesic exists. So Z is in the black hole region.

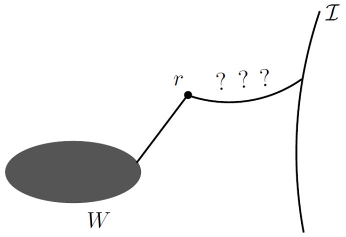

Figure 21: An observer who can receive a signal from a point r in the causal future of a set W can also receive a signal from W. So if W is contained in the black hole set and r is in its causal future, the curve labeled ? ? ? connecting r to the worldline $\mathcal{I}$ of an observer at infinity must not be future-going causal; r must be contained in the black hole set.

There might be several black holes in spacetime, so on a given Cauchy surface S, the black hole region might have several disconnected components (Fig. 22(a)). Black holes can merge, but a black hole cannot split. Let $Z \subset B$ be a connected component of the black hole region on a given Cauchy surface S, then the future of Z intersects any Cauchy surface S' to the future of S in a connected set. To prove this, suppose that the black hole region intersects S' in disconnected components Z'_i, $i = 1, \cdots, r$ (Fig. 22(b)). Consider future-going causal paths from each point in Z to S'. If there is a causal path from Z to a given component Z'_i of the black hole region, then by maximising the elapsed proper time of such a path, we learn that there is a causal geodesic (null or timelike) from Z to Z'_i. But, the space of future-going causal geodesics starting at Z is connected, and cannot be continuously divided into two or more disjoint subsets that would intersect S' in different components of the black hole region. So all causal geodesics from Z arrive at S' in the same component Z'_i of the black hole region. Therefore, there is only one component Z'_i of the black hole region on S' is in the future of Z.

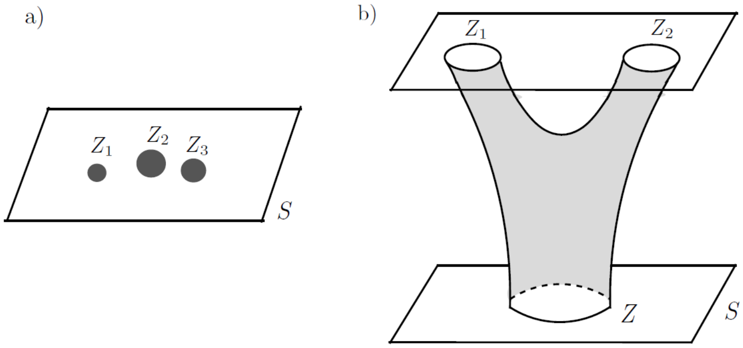

Figure 22: (a) In a spacetime that contains multiple black holes, the black hole region on a given Cauchy surface S might have several connected components Z_i. (b) An impossible situation, in which a black hole splits into two, or more precisely in which a component Z of the black hole region on a Cauchy surface S evolves to two disconnected components Z_1 and Z_2 on a future Cauchy surface S'.

12 The Horizon and Its Generators

Let Σ be a Cauchy surface, and let $C \subset \Sigma \cap M$ be a closed subset such that $\theta \leq 0$ for the outgoing family of null geodesics orthogonal to its surface $S = \partial C$. Such a surface S is called an **outer marginally trapped surface**, and C is called a **trapped region**.

We refer to the region $B \cap \Sigma$ as the **total black hole region at time** Σ ("time" here means the choice of Cauchy surface Σ). Each connected component of $B \cap \Sigma$ will be called a **black hole at time** Σ. The trapped region C has the property that $C \subset B \cap \Sigma$.

We define the **total trapped region**, $\mathcal{T}$, of a Cauchy surface, Σ, to be the closure of the union of all trapped regions, C, on Σ. We call the boundary, $\mathcal{A} = \partial\mathcal{T}$, of $\mathcal{T}$ the **apparent horizon** on Σ. It is a codimension 2 submanifold. The **trapped horizon** is the union of apparent horizons on all Σ's. It is a codimension 1 submanifold (sometimes also referred to as *apparent horizon*).

In a globally hyperbolic spacetime with NEC satisfied, we have $\mathcal{T} \subset B \cap \Sigma$, so the apparent horizon $\mathcal{A}$ always lies inside of or coincides with the true **event horizon**, $H \cap \Sigma$, on Σ. The apparent horizon, $\mathcal{A}$, satisfies the property: if the total trapped region $\mathcal{T}$ on Σ is a manifold with boundary, then $\mathcal{A}$ is an *outer marginally trapped surface* with *vanishing expansion*, $\theta = 0$ (Fig. 23).

Example. *Recall the example of a Vaidya spacetime. From Fig. 24, we see that the event horizon forms "before" the formation of a black hole. After the black hole is formed, the event horizon on Σ_2 is the apparent horizon on Σ_2.*

Horizon generator

Let q be a point on the black hole horizon H, and let $\mathcal{I}$ be the timelike worldline of an observer who is stationary at a great distance (or at infinity). The point $q \in H$ is the limit of a sequence of points $q_1, q_2, q_3, \cdots$ that are outside of the black hole region B. Each of the q_i is connected to the worldline $\mathcal{I}$ by a future-going prompt null geodesic ℓ_i. In a globally hyperbolic spacetime, as $q_i \to q$, the ℓ_i converge to a future-going null geodesic ℓ from q. But since q is contained in the black hole region ($q \in H = \partial B$), it is not true that ℓ connects q to a point in $\mathcal{I}$. Rather, what happens is that as $q_i \to q$, the geodesic ℓ_i arrives at $\mathcal{I}$ later and later; the proper time at which the distant observer can see the point q_i diverges (goes to ∞ along $\mathcal{I}$). The limiting geodesic ℓ does not reach $\mathcal{I}$ at all (Fig. 25).

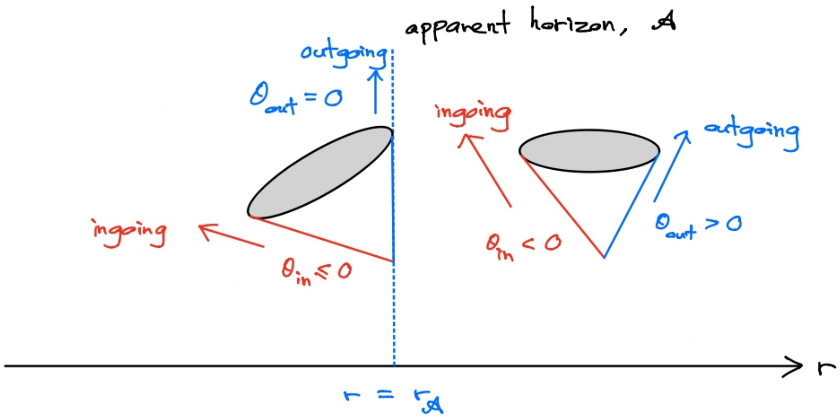

Figure 23: The apparent horizon, $\mathcal{A}$, is an outer marginally trapped surface with vanishing expansion, $\theta = 0$, if the total trapped region $\mathcal{T}$ on the Cauchy surface Σ is a manifold with boundary.

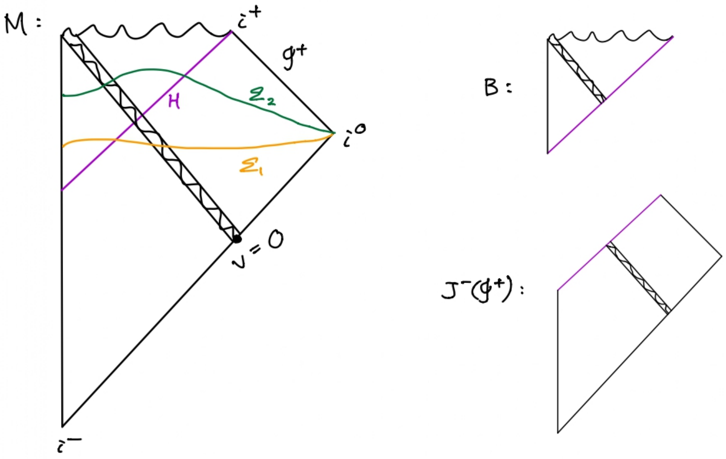

Figure 24: Penrose diagram of Vaidya spacetime. The event horizon $H \cap \Sigma_1$ forms on Σ_1 before the formation of a black hole. After the black hole is formed, the event horizon $H \cap \Sigma_2$ on Σ_2 is the apparent horizon $\mathcal{A}$ on Σ_2.

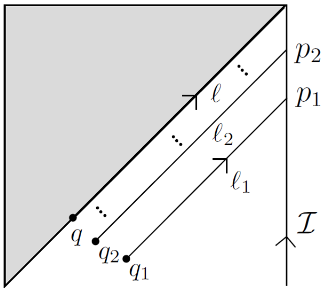

Figure 25: In this Penrose diagram, the black hole region is shaded, and q is a point in its boundary. (The part of the spacetime to the right of $\mathcal{I}$ is not drawn.) The upper corner of this diagram, at which ℓ and $\mathcal{I}$ appear to meet, is actually at future infinity. It is not a point in the spacetime.

ℓ remains everywhere in the horizon. One can see this as follows. ℓ can never go outside B, since in general a causal curve starting at $q \in H = \partial B$ can never reach outside B. On the other hand, ℓ can nowhere be in the interior of B, because it is the limit of the prompt null geodesics from q_i that are strictly outside B. So ℓ is everywhere in $H = \partial B$. ℓ is called a **horizon generator**.

ℓ is a *prompt null geodesic, without focal point.* Indeed, if ℓ fails to be prompt, there is a causal path from $q \in \ell$ to some point $r \in M\backslash B$ that is strictly to the past of some $q' \in \ell$. Such an r is outside the black hole region, and the existence of a causal path from q to r would show that q is also outside B.

Now pick an initial value surface S that contains q and define $W = S \cap H$. ℓ must be orthogonal to W, since a prompt causal path from any submanifold (to any destination) is always orthogonal to that submanifold. Together, the horizon generators that pass through W sweep out a codimension 1 submanifold $H' \subset H$. Near W, H' and H are the same (assuming that W is differentiable). But if we continue into the future, H' might not coincide with H, since, for example, new black holes may form, as a result of which the horizon (even its connected component that contains W) may not be swept out entirely by the horizon generators that come from W.

Since it is swept out locally by orthogonal null geodesics, H must be a *null hypersurface*, that is a hypersurface with a degenerate metric of signature $++\cdots+0$.

Properties of horizon generator:

1. A horizon generator can enter the horizon at some point.
2. Once enter, it cannot leave the horizon.
3. Two generators cannot intersect.
4. Through each point on the event horizon, except where new generators enter, there passes one and only one generator.

Hawking area theorem

Hawking area theorem says that the area of the black hole horizon can only increase, in the sense that the area measured on an initial value surface S' that is to the future of S is equal to or greater than the area measured on S. The theorem applies separately to each component of the black hole region; if two or more black holes merge, the theorem says that the merger produces a black hole with a horizon area at least equal to the sum of the horizon areas of the original black holes.

It suffices to show that the null expansion $\theta = \dot{A}/A$ of the horizon generators is everywhere nonnegative. This being so along every horizon generator means that the horizon area is everywhere nondecreasing.

If we know that the horizon generators are complete (which roughly says that the horizon is nonsingular), and assume that θ is negative for one of the horizon generators ℓ, then from Raychaudhuri equation, there will be a caustic or focal point along ℓ at some bounded value of its affine parameter. But we have already shown that the horizon generators are prompt null geodesics that have no focal points.

More generally, it is possible to prove the area theorem without assuming the completeness of horizon generators. Imagine that in some portion of $W = H \cap S$, one has $\theta < 0$. Then we define a new surface W' by pushing W out a little in the region with $\theta < 0$, leaving W unchanged wherever $\theta \geq 0$. θ varies continuously as W is moved in M, so if we do not push too far, W' has $\theta < 0$ in the portion outside of B (Fig. 26). Since W' is outside B, there exists a prompt null geodesic ℓ connecting $\mathcal{I}$ to a point in W'. But we have chosen W' so that at such points, $\theta < 0$. Hence ℓ must have a focal point within a bounded affine distance of W'. This contradicts the fact that $\mathcal{I}$ can be arbitrarily far away. So in fact there was nowhere on W with $\theta < 0$.

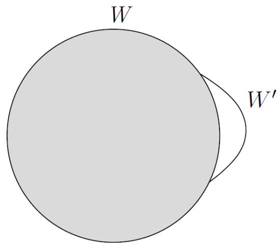

Figure 26: W is the boundary of a component of the black hole set (shaded) on some Cauchy surface S. Such a W is a codimension 2 spacelike submanifold of spacetime. If there is a portion of W on which the null expansion θ of the horizon generators is negative, then by pushing W outward slightly in that region, while remaining in S, we get a new submanifold W', still spacelike and of codimension 2, that is partly outside of the black hole set, and such that the part of W' that is outside the black hole set still has $\theta < 0$. In the figure, the region between W and W' is unshaded, as it is not part of the black hole set.

Let us make a comparison between the apparent horizon $\mathcal{A}$ and the event horizon $H \cap \Sigma$.

Apparent horizon $\mathcal{A}$	**Event horizon $H \cap \Sigma$**
Defined locally	Defined globally
Codimension 2	Codimension 1
$\delta A(\mathcal{A}) \geq 0$	$\delta A(H \cap \Sigma) \geq 0$
Signals the formation of black hole	Cannot predict the black hole formation unless we know the future!

13 Topological Censorship and ANEC

Wormholes

A **"wormhole"** is a geometrical connection between two different asymptotically flat worlds (Fig. 27(a)), or a shortcut between two distant regions of a single asymptotically flat world (Fig. 27(b)). (A similar discussion applies in a spacetime that is asymptotic to AdS space.) The two cases are closely related. The existence of a wormhole of the second type means that spacetime is not simply-connected; by taking a cover of spacetime, one can pass to a situation of the first type.

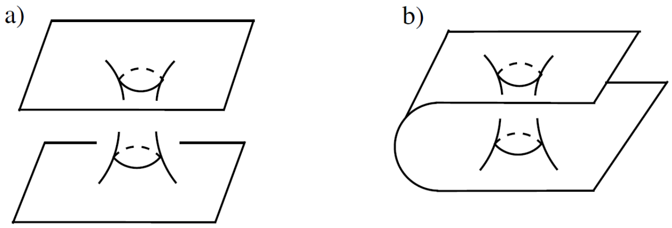

Figure 27: Two types of wormhole: (a) a connection between two different worlds; (b) a shortcut between distant regions of one world. In each case, what is depicted is a Cauchy surface S.

We recall that, topologically, a globally hyperbolic spacetime M is simply $M = S \times \mathbb{R}$ where S is an initial value surface and $\mathbb{R}$ parameterises the "time." So for M to have a wormhole means simply that S has a wormhole.

Example (Wormhole of Schwarzschild black hole)**.** *The motivating example of topological censorship is the analytically continued Schwarzschild solution (Fig. 28). This is a globally hyperbolic spacetime with two asymptotically flat ends, labeled I and II in the figure. The left and right ends of the initial value surface S are respectively in the asymptotically flat regions I and II; the interior part of S passes through the wormhole. But as one can see from the Penrose diagram, a causal signal from one asymptotically flat region cannot travel through the wormhole to the other region. If one enters the wormhole from the left, hoping to traverse it and to come out on the right, one will instead end up at the black hole singularity.*

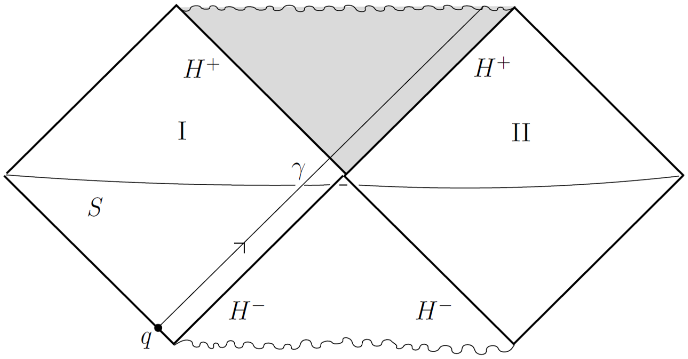

Figure 28: This is the Penrose diagram of the maximal analytic extension of a Schwarzschild black hole. It contains two asymptotically flat regions, here labeled I and II. They are spacelike separated and causal communication between them is not possible. The spacetime is globally hyperbolic, with initial value surface S. It connects the two asymptotically flat regions, and has a "wormhole" topology, similar to that of Fig. 27(a). Labeled H^+ are the future horizons of observers who remain at infinity in regions I or II; beyond H^+ and to its future is the black hole region (shaded). From the point of view of the outside observer, there also are past horizons H^-; beyond H^- and to its past is the white hole region (unshaded). Shown in the figure is a radial null geodesic γ that originates at the point q at past null infinity in the asymptotically flat region on the left. It crosses S in the "wormhole" region, but it does not "traverse the wormhole" and enter the asymptotically flat region on the right; rather it enters the black hole region and terminates on the future singularity.

Example (Wormholes in AdS space)**.** *The $(d+1)$-dimensional Euclidean AdS space is topologically equivalent to a $(d+1)$-dimensional hyperbolic space, i.e. $AdS_{d+1} = \mathbb{H}_{d+1}$. Consider a hyperbolic slicing by*

$$ds^2_{\mathbb{H}_{d+1}} = d\rho^2 + \cosh^2\rho\, ds^2_{\mathbb{H}_d}. \tag{13.1}$$

The wormholes can be obtained by taking the quotient of $\mathbb{H}_d$ by a discrete group Γ which generates a compact group. When $\rho \to \infty$, we have the boundary of $\mathbb{H}_{d+1}$, which we denote as ∂M_1; when $\rho \to -\infty$, we have the other boundary denoted as ∂M_2 (Fig. 29).

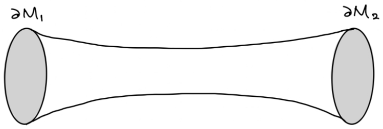

Figure 29: A wormhole in AdS space with two boundaries ∂M_1 and ∂M_2.

Topological censorship

In a spacetime that satisfies the assumptions:

1. NEC,
2. classical Einstein equation, and
3. globally hyperbolic,

the **topological censorship** says that, there may be a wormhole in space, but it is not traversable, in the sense that it is not possible for a causal signal to go through the wormhole and come out the other side. In proving this, it suffices to consider a wormhole that connects two distinct asymptotically flat worlds, as in Fig. 27(a). The case of Fig. 27(b) can be reduced to this, without affecting the classical null energy condition, by taking a cover of spacetime.

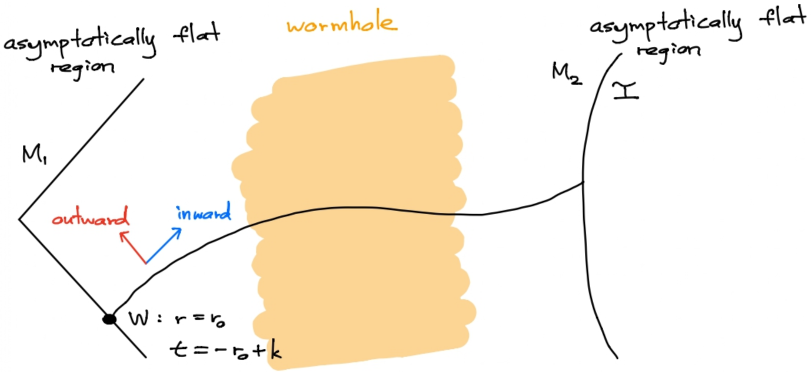

Figure 30: M_1 and M_2 are the two asymptotically flat "ends" of spacetime. W is a large sphere embedded in M_1 in the far past. $\mathcal{I}$ in region M_2 is the worldline of a stationary observer at a great distance from the wormhole. By the arguments in the text, there can be no prompt causal path from W to $\mathcal{I}$.

Let M_1 and M_2 be two asymptotically flat "ends" of spacetime (Fig.30). The line element of M_1 is asymptotic at infinity (far region) to the Minkowski line element:

$$ds^2 = -dt^2 + dr^2 + r^2 d\Omega^2, \tag{13.2}$$

where $d\Omega^2$ is the line element of a $(d-1)$-sphere. Let W be a large sphere embedded in M_1 by, say, $r = r_0$, $t = -r_0 + k$, where k is a constant and r_0 is taken to be very large. Thus W is embedded in M_1 at a very great distance from the wormhole and in the far past, in such a way that the "advanced time" $t + r = k$ remains fixed. The purpose of this is to ensure that a signal can be sent in from W to arrive at the wormhole at a time of order 1, independent of r_0. By varying k, we can adjust when the signal will arrive at the wormhole. We will see that regardless of the choice of k, a signal from W cannot travel through the wormhole.

In region M_2, we take $\mathcal{I}$ to be the worldline of a stationary observer at a great distance from the wormhole. If it is possible for a signal to propagate from M_1 through the wormhole and to emerge in M_2, then there is a causal path that propagates from W, travels into the interior of M_1, passes through the wormhole, and eventually reaches $\mathcal{I}$. Then there is a prompt causal path – a future directed null geodesic, orthogonal to W and without focal points, which arrives on $\mathcal{I}$ as soon as possible.

Given how W was defined, the future-going null geodesics orthogonal to W propagate either "outward," to larger r, or "inward," to smaller r. The outward-going null geodesics simply propagate outward to the boundary of M_1 ($r = \infty$ in the asymptotically flat region). It is the inward-going null geodesics that might propagate into the wormhole. However, the inward-going null geodesics have an initially negative value of the null expansion $\theta = \dot{A}/A = -2/r < 0$. Assuming the NEC, Raychaudhuri equation implies that these inward-going null geodesics reach focal points within a bounded value of their affine parameters. As $\mathcal{I}$ can be arbitrarily far away in M_2, it follows that all of these inward-going null geodesics reach focal points before reaching $\mathcal{I}$. This shows that there can be no prompt causal path from W to $\mathcal{I}$, and hence that there can be no such causal path at all. Therefore, the wormhole is not traversable.

The averaged null energy condition (ANEC)

In our discussion of topological censorship, we have assumed the null energy condition. This is a pointwise condition on the stress tensor and is satisfied by reasonable classical matter, but in quantum field theory, it is not satisfied by the expectation value of the quantum stress tensor. So the question arises of whether topological censorship is valid in a quantum universe.

It turns out that topological censorship is valid under a weaker hypothesis known as the **"averaged null energy condition" (ANEC)**. For a complete null geodesic ℓ, with affine parameter U that runs to infinity at both ends, the ANEC asserts that

$$\int_\ell dU\ T_{UU} \geq 0. \tag{13.3}$$

Here

$$T_{UU} = \frac{dx^\alpha}{dU}\frac{dx^\beta}{dU}T_{\alpha\beta}. \tag{13.4}$$

In the same notation, the NEC gives

$$T_{UU} \geq 0. \tag{13.5}$$

Clearly, the ANEC is a strictly weaker condition.

The ANEC is taken to mean that the operator $\int_\ell dU\ T_{UU}$ is nonnegative, in the sense that its expectation value in any quantum state Ψ is nonnegative. The ANEC is, however, not true in general, even if the null geodesic ℓ is complete (meaning that its affine parameter extends to infinity in both directions). For a simple counterexample, consider the cylindrical spacetime with flat metric $ds^2 = -dt^2 + d\phi^2$, as in Fig. 3. Consider also a conformally invariant quantum field theory in the cylindrical spacetime. The ground state of a conformal field theory in this spacetime has a negative Casimir energy. Moreover, by translation invariance, the energy density in this state is a constant in spacetime. The integrand in the ANEC integral (for any null geodesic ℓ) is therefore negative-definite in this example, and the ANEC is not satisfied.

It is believed that the ANEC may hold under two additional conditions:

1. the null geodesic ℓ should be achronal,
2. the spacetime should be self-consistent, meaning roughly that the Einstein equations are obeyed with a source given by the expectation of the stress tensor.

This has not been proved. One basic result is that the ANEC holds for a geodesic in Minkowski space. There are partial results for more general spacetimes.

For our purposes, it suffices that ANEC can be applied in some classical results, including topological censorship. For a null geodesic ℓ that extends in both directions to infinite values of its affine parameter (i.e. from $\lambda \to -\infty$ to $\lambda \to \infty$), and if the ANEC is satisfied for ℓ as a strict inequality,

$$\int_\ell dU\ T_{UU} > 0, \tag{13.6}$$

then a segment of ℓ that is sufficiently extended in both directions is not prompt.

The Gao-Wald theorem

Let M be an asymptotically AdS spacetime (i.e. a spacetime that is asymptotic at spatial infinity to AdS space). By adding some points at spatial infinity, one can construct a partial conformal compactification of M. The points at infinity make up a Lorentz signature manifold N, whose dimension is one less than the dimension of M, and **AdS/CFT duality** says that a gravitational theory on M is equivalent to some conformal field theory on N. The picture is schematically indicated in Fig. 31(a).

The **Gao-Wald theorem** says that, assuming the ANEC, the AdS/CFT correspondence is compatible with causality. This means that the bulk causality is compatible with the boundary causality in AdS/CFT duality.

In the boundary CFT, an event q can influence an event p only if there is a future-going causal path from q to p in N. But in the bulk gravitational theory, q can influence p if there is a future-going causal path between them in M. AdS/CFT duality obeys causality if there is "no shortcut through the bulk," that is, no causal path from q through M can arrive at a boundary point that could not be reached by a causal path in the boundary N (Fig. 31(b)).

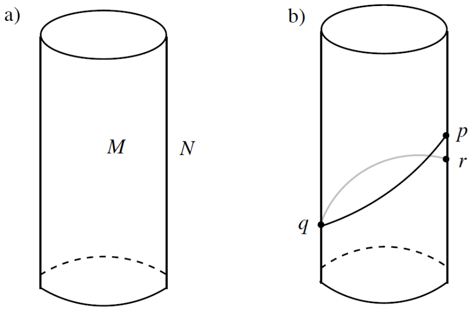

Figure 31: (a) According to the AdS/CFT correspondence, a quantum gravity theory in the AdS spacetime M is equivalent to an ordinary quantum field theory on a spacetime N of one dimension less which is the conformal boundary of M. (b) In the boundary theory, an event at a point $q \in N$ can influence an event at $p \in N$ if a future-going causal curve in N can propagate from q to p. Causality will be violated if it is possible to take a shortcut through the bulk, in the sense that a future-going causal curve in M can propagate from q to a point $r \in N$ that is strictly to the past of p. In the figure, the curve qp is a causal curve in the boundary while qr is a hypothetical causality-violating shortcut in the bulk. The Gao-Wald theorem asserts that such causality violation does not occur.

Causality condition is precisely satisfied in empty AdS spacetime in which every null geodesic is prompt, and a bulk null geodesic from q can arrive on the boundary precisely at p, but no earlier. What happens if one perturbs away from empty AdS, by adding some matter or gravitational perturbations? The Gao-Wald theorem says that, assuming the ANEC, there is never a causality-violating shortcut through the bulk.

To show that this situation cannot occur, we need to know that the affine parameter of a null geodesic diverges as it reaches the conformal boundary of an asymptotically AdS spacetime. This happens even though the time, as measured on the boundary, does not diverge. In suitable coordinates (or in Fefferman-Graham gauge with $z \to 0$), the line element of an asymptotically AdS spacetime looks near the conformal boundary like

$$ds^2 = \frac{1}{z^2}\left(dz^2 - dt^2 + d\vec{x}^2\right). \tag{13.7}$$

The spacetime is the region $z > 0$; the conformal boundary is at $z = 0$. A typical null geodesic is

$$z = -t, \quad \vec{x} = \vec{x}_0, \quad t \leq 0, \tag{13.8}$$

and clearly reaches the boundary at $z = 0$ at a finite value of t, namely $t = 0$. However, the affine parameter of this geodesic diverges as $t \to 0^-$. To see this, one observes that the equation $D^2x^\mu/d\lambda^2 = 0$ for a geodesic with affine parameter λ gives in this case

$$\frac{d}{d\lambda}\left(\frac{1}{z^2}\frac{dt}{d\lambda}\right) = 0, \tag{13.9}$$

so

$$\frac{1}{z^2}\frac{dt}{d\lambda} = q_0 \tag{13.10}$$

with a constant q_0. This constant cannot be zero, since t is not constant in the geodesic. Setting $z = -t$, we get $dt/t^2 = q_0 d\lambda$, so that up to an additive constant, $\lambda = -1/q_0 t$. Thus λ diverges as $t \to 0^-$.

A similar argument can be obtained for a null geodesic with

$$z = t, \quad \vec{x} = \vec{x}_0, \quad t \geq 0. \tag{13.11}$$

The result is the same, i.e. $\lambda \to -\infty$ as $t \to 0^+$.

Now let us return to our hypothetical causality-violating prompt geodesic ℓ, a shortcut through the bulk from q to a point r that is strictly to the past of p.

- If the ANEC is satisfied along ℓ with a strict inequality, $\int_\ell dU\ T_{UU} > 0$, then, since the affine parameter of ℓ diverges at both ends, ℓ cannot be prompt and therefore the hypothetical causality-violating shortcut ℓ cannot exist.
- For the saturated case, $\int_\ell dU\ T_{UU} = 0$, if a prompt shortcut exists, a suitably small perturbation (by adding an infinitesimal amount of matter or slightly changing the quantum state) will make the ANEC a strict inequality, $\int_\ell dU\ T_{UU} > 0$. The prompt shortcut ℓ still exists and will arrive on the conformal boundary so close to r as to be still strictly to the past of p. This is a contradiction.

Of course, as a special case of this, we could have deduced the Gao-Wald theorem from the classical NEC, $T_{UU} \geq 0$, rather than the ANEC.

References

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

[2] E. Witten, "Light Rays, Singularities, and All That," *Rev. Mod. Phys.* **92** no. 4, (2020) 045004, arXiv:1901.03928v5 [hep-th].