Black Holes, Singularity Theorems, and All That

Edward Witten

NZMRI 2020

The goal of these lectures is to provide an introduction to global properties of General Relativity. We want to understand, for example

(1) why formation of a singularity is inevitable once a trapped surface forms

(2) why the area of a classical black hole can only increase

(3) why, classically, one cannot traverse a wormhole ("topological censorship")

(4) why the AdS/CFT correspondence is compatible with causality in the boundary CFT (the Gao-Wald theorem).

These are mostly statements about the causal structure of spacetime – where can one get, from a given starting point, along a worldline that is everywhere inside the local light cone?

The causal structure is what is new in Lorentz signature General Relativity relative to the Euclidean signature case which is more visible in everyday life.

To those of you not already familiar with our topics, I can offer some good news: there are a lot of interesting results, but they are all based on a few ideas – largely developed by Penrose in the 1960's, with important elaborations by Hawking and others. So one can become conversant with this material in a short span of time.

I will follow rather closely my lecture notes arXiv:1901.03928. Here are some other references:

Books:

R. Penrose, *Techniques of Differential Geometry in Relativity* (1972)

S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure Of Spacetime* (1973)

R. Wald, *General Relativity* (1984)

J. K. Beem, P. E. Ehrlich, and K. L. Easley *Global Lorentzian Geometry* (2nd edition, 1996)

Lecture notes:

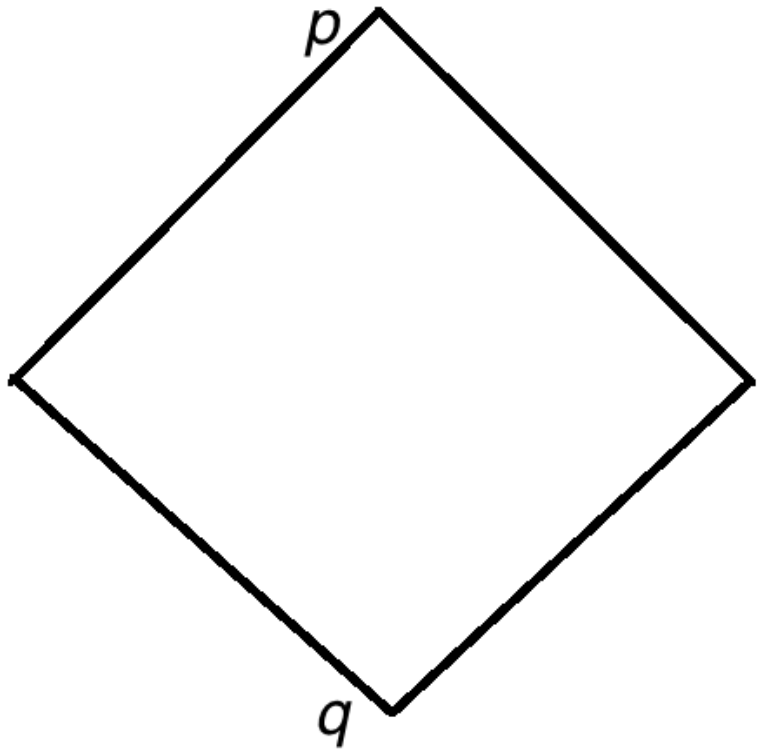
P. T. Chrusciel, "Elements of Causal Theory," arXiv:1110.6706

G. Galloway, "Notes on Lorentzian Geometry," available online

S. Aretakis, "Lecture Notes on General Relativity," available online

Since the questions we will be asking are about causal structure, we will be studying causal paths in spacetime. A causal path $x^\mu(\tau)$ is one whose tangent vector $\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}$ is everywhere timelike or null. Geodesics will play an important role, because a lot of the questions have to do with "what is the best one can do with a causal path?" For example, what is the closest one can come to escaping from the black hole or traversing the wormhole? The answer to such a question usually involves a geodesic with special properties.

We start by just considering causal paths from a point q in Minkowski spacetime to a point p in its future, i.e. inside the future light cone.



Drawn is the "causal diamond" D_q^p, consisting of all points in the causal future of q and causal past of p. A causal path from q to p will lie in this diamond.

In relativity theory, a "causal path" is the path in spacetime of a particle that never travels faster than light. Thus in the case of a parametrized path $x^\mu(s)$, where x^μ, $\mu = 1, \cdots, D$ are spacetime coordinates, the condition is that the tangent vector to the path

$$\frac{\mathrm{d}x^\mu(s)}{\mathrm{d}s}$$

is everywhere timelike or null. In more detail

$$g_{\mu\nu}\frac{\mathrm{d}x^\mu}{\mathrm{d}s}\frac{\mathrm{d}x^\nu}{\mathrm{d}s} \leq 0$$

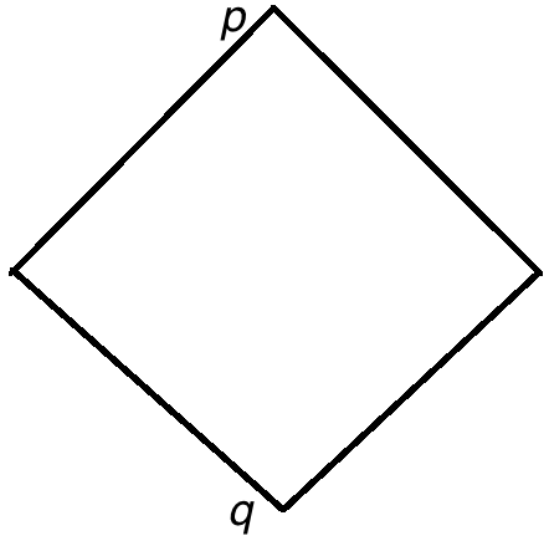
where $g_{\mu\nu}$ is the spacetime metric, which I take to be of signature $-+++\cdots+$. On a causal path we can pick a *future-directed* orientation, i.e. so that its tangent vector is a future-going timelike or lightlike vector, and normally we do that. (In any local Lorentz frame, $\mathrm{d}t/\mathrm{d}s > 0$ where t is the time.)

The first essential point is that the space of causal paths from q to p is in an appropriate sense compact. Causality is essential here. Without it, a sequence of paths could oscillate more and more wildly and have no convergent subsequence. For example, in two-dimensional Minkowski spacetime with metric $\mathrm{d}s^2 = -\mathrm{d}t^2 + \mathrm{d}x^2$, here is a sequence of non-causal paths from $q : (t, x) = (0, 0)$ to $p : (t, x) = (1, 0)$:

$$x = \sin(\pi n t).$$

These paths oscillate more and more wildly with no limit for $n \to \infty$. Taking a subsequence does not help, so the space of all paths from q to p is noncompact.

Causality changes this because in the picture



a causal path cannot have an angle of more than $\pi/4$ from the vertical. Although we are really interested in causal paths in the Lorentz signature metric $\mathrm{d}s^2 = -\mathrm{d}t^2 + \mathrm{d}x^2$, it is useful to compare to the Euclidean signature metric

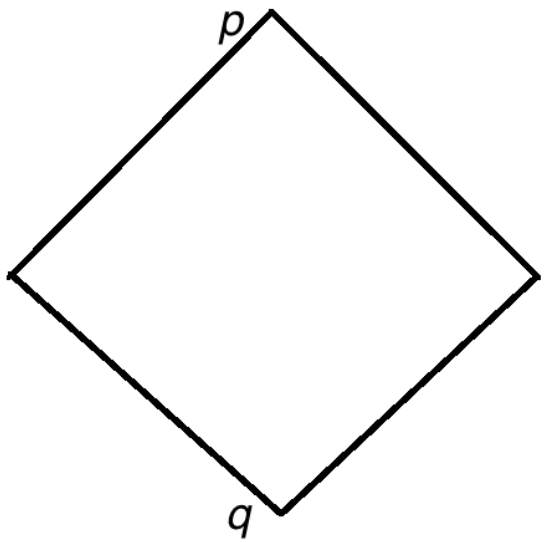
$$\mathrm{d}s_E^2 = \mathrm{d}t^2 + \mathrm{d}x^2.$$

A straight line from $q = (0, 0)$ to $p = (1, 0)$ has Euclidean length 1, and an arbitrary causal path has length no more than $\sqrt{2}$.

Once we have an upper bound on the Euclidean length, compactness follows. Parametrize a causal path of Euclidean length $\lambda \leq \sqrt{2}$ by a parameter s that measures λ times the arclength, so s runs from 0 to 1. Suppose we are given a sequence of such paths

$$x_n(s), \qquad n = 1, 2, 3 \cdots$$

Since the total Euclidean length is $\leq \sqrt{2}$, all these paths lie in a compact subset D of Minkowski space.



They all satisfy $x_n(0) = q$, $x_n(1) = p$.

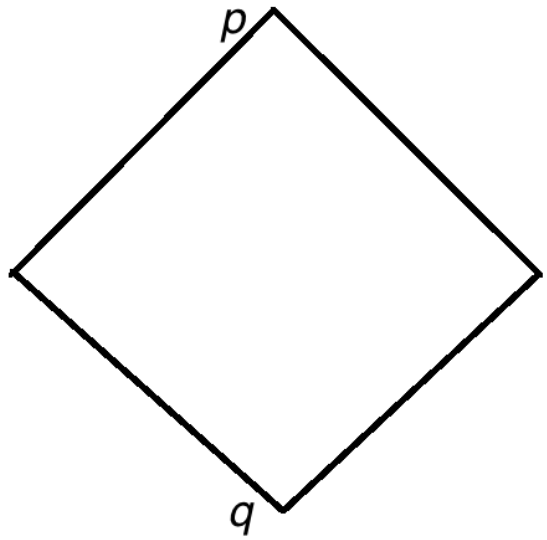
The existence of a convergent subsequence of the paths $x_n(s)$ follows by an argument that many of you will recognize: Since $x_n(s = 1/2)$ lies in a compact subset D_q^p of Minkowski space, there is a subsequence of the paths $x_n(s)$ such that $x_n(1/2)$ converges. Extracting a further subsequence, one ensures that $x_n(1/4)$ and $x_n(3/4)$ converges. Continuing in this way, we get a subsequence such that $x_n(a/2^k)$ converges for integers a, k. The constraint on the length ensures that wild fluctuations are not possible, so actually for this subsequence, $x_n(s)$ converges for all s. We have learned that any sequence of causal paths from q to p has a convergent subsequence, and thus the space of causal paths is compact.

A corollary is that there must be a causal path from q to p that *maximizes* the elapsed proper time, which is

$$\tau = \int_0^1 \mathrm{d}s \sqrt{\left(\frac{\mathrm{d}t}{\mathrm{d}s}\right)^2 - \left(\frac{\mathrm{d}x}{\mathrm{d}s}\right)^2}.$$

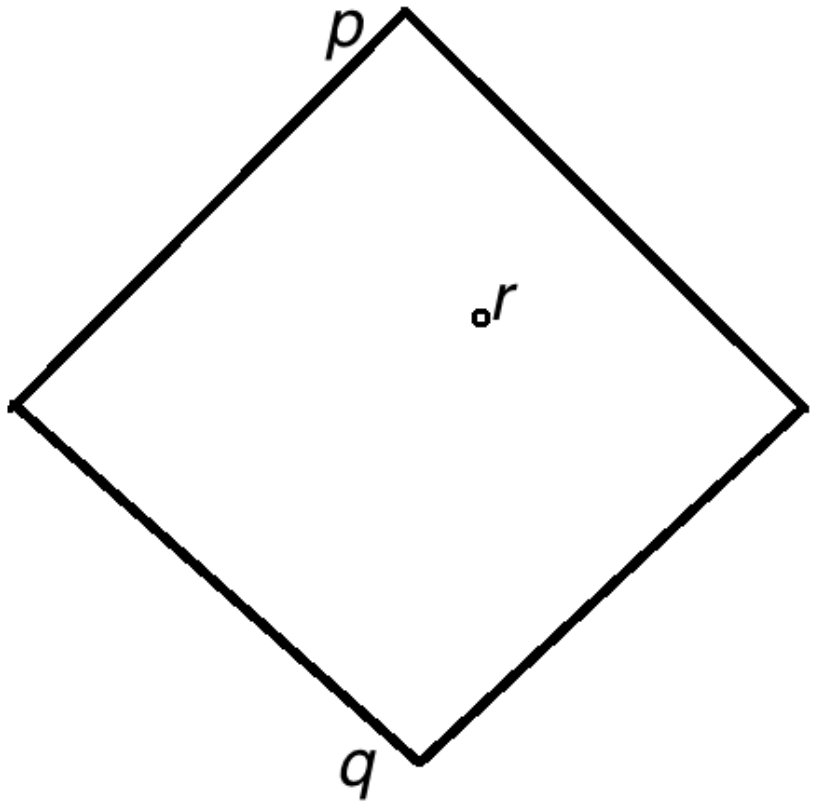
(There must be a finite upper bound on τ, because a sequence of paths whose proper time τ_n increases without limit could not have a convergent subsequence. If τ_0 is the least upper bound on τ, then a sequence of paths with $\tau_n \to \tau_0$ would have a convergent subsequence, which would be a causal path of proper time τ_0.) In the particular case of Minkowski spacetime, we can prove this more trivially: there is a unique geodesic from q to p, namely a straight line, and it is the path of greatest proper time.

However, the only fact that we really needed about Minkowski spacetime to establish the compactness of the space of causal paths from q to p was that the "causal diamond" D_q^p of points that can be visited by such a path is compact.



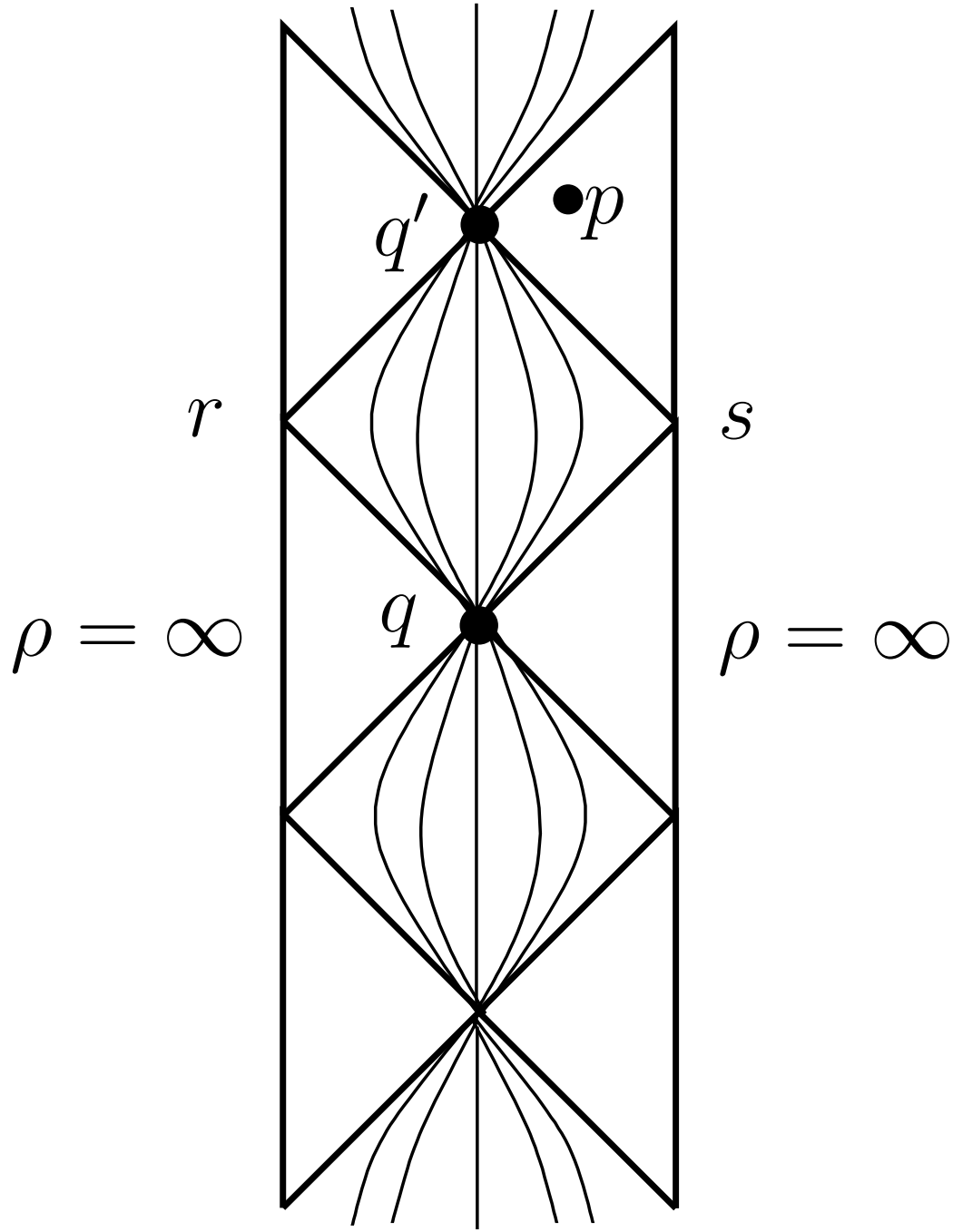
The space of causal paths is compact whenever that is true. D_q^p is the intersection of the causal future of q with the causal past of p. We call it a causal diamond.

If the causal diamond D_q^p is not compact, the space of causal paths is generally not compact and a causal path of greatest proper time may not exist. I will give a trivial example to illustrate this and also a less trivial example. Here is the trivial example:



A point r has been omitted, so a sequence of causal paths from q to p need no longer have a limit.

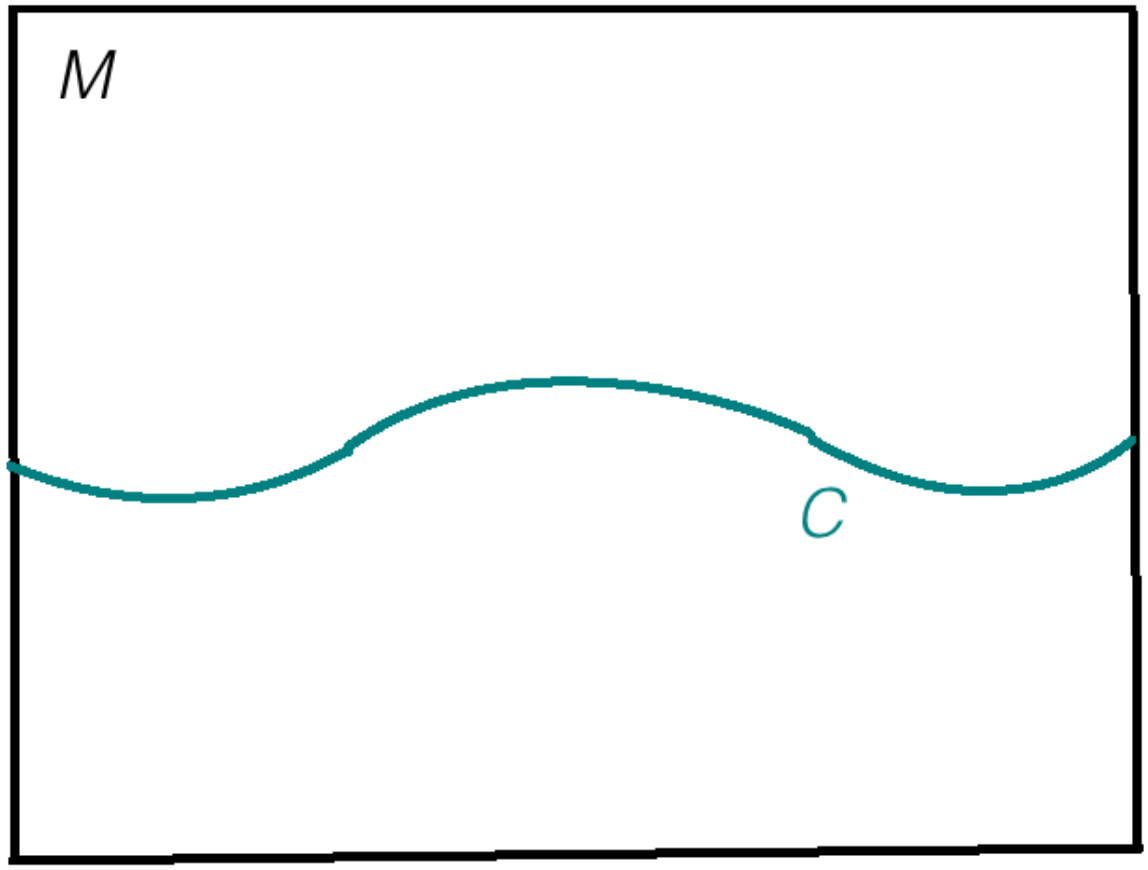
A more interesting example is provided by Anti de Sitter space, but unfortunately to describe it properly without assuming familiarity with AdS spacetime would require quite a digression. I will just give a short explanation for those who are familiar with the Penrose diagram of Anti de Sitter spacetime.


q'
p
r
s
$\rho = \infty$
q
$\rho = \infty$

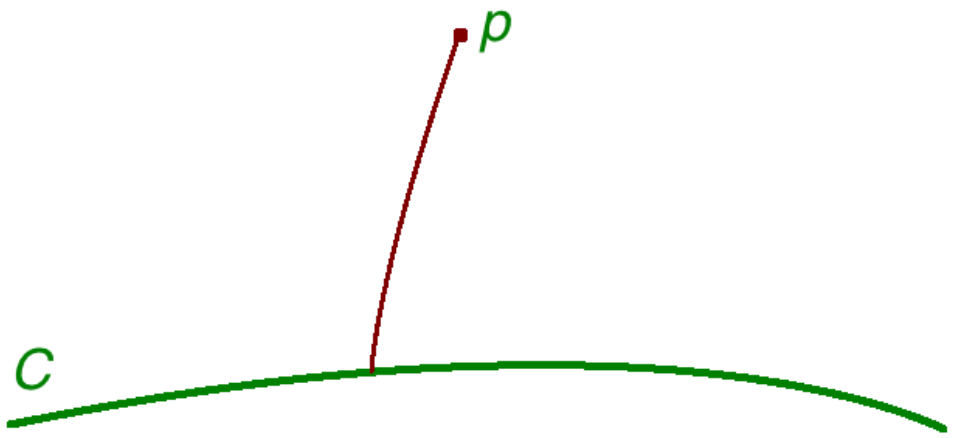
To recapitulate, we are studying causal paths because they are the key to understanding black holes, singularities, and all that, and we have learned that the space of causal paths from a point q to a point p in its future is compact as long as the causal diamond D_q^p is compact. We would like a useful criterion that ensures this.

Luckily there is a good criterion that is very well-motivated physically.

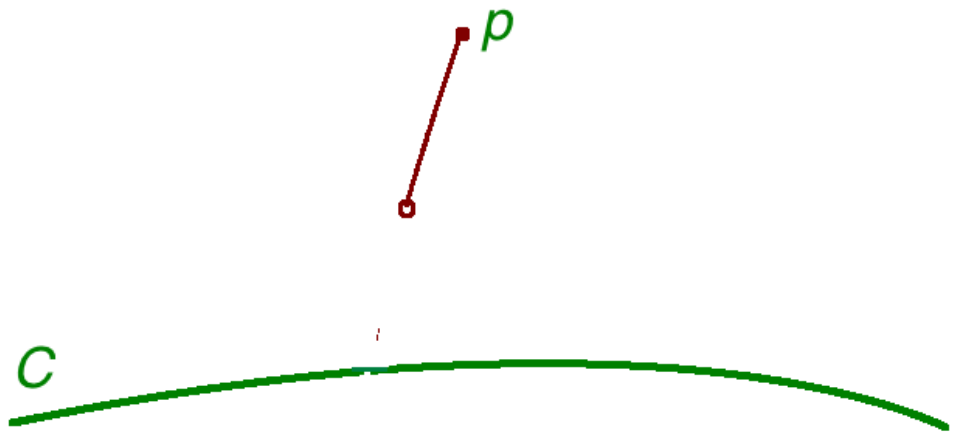
We ask for a spacetime M to be "globally hyperbolic." This means that it has an initial value surface C, and whatever happens in M can be determined from initial conditions on C.



Technically, C is spacelike, every point in M that is not on C is to the past or future of C (but not both), and if p is a point to the future of C, then any past-going causal path from p can be continued until it meets C (if q is to the past of C, any future-going causal path from q can be continued until it meets C). The intuition is that any signal that one observes at p must have arrived at p along some causal path, so it must have originated on C (or more precisely, it could be predicted if one knows the initial data on C).

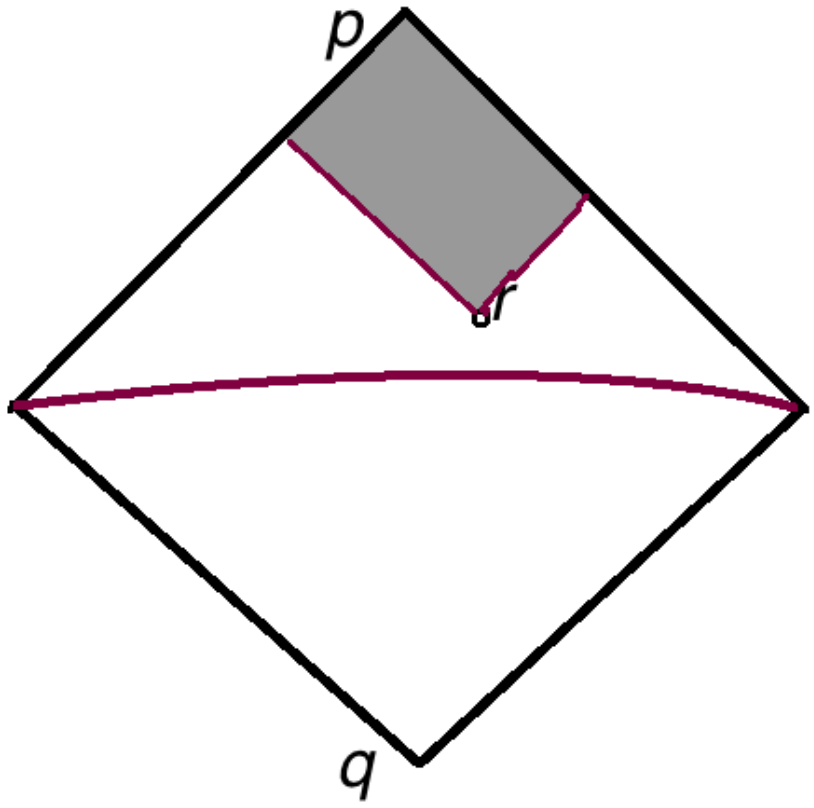


If there were a missing point between p and C, a past-going causal path from p could get "stuck" and could not be continued until it reaches C:



To predict the physics at p, initial data at C would not be enough; one would need to know what signal emerges from the missing point.

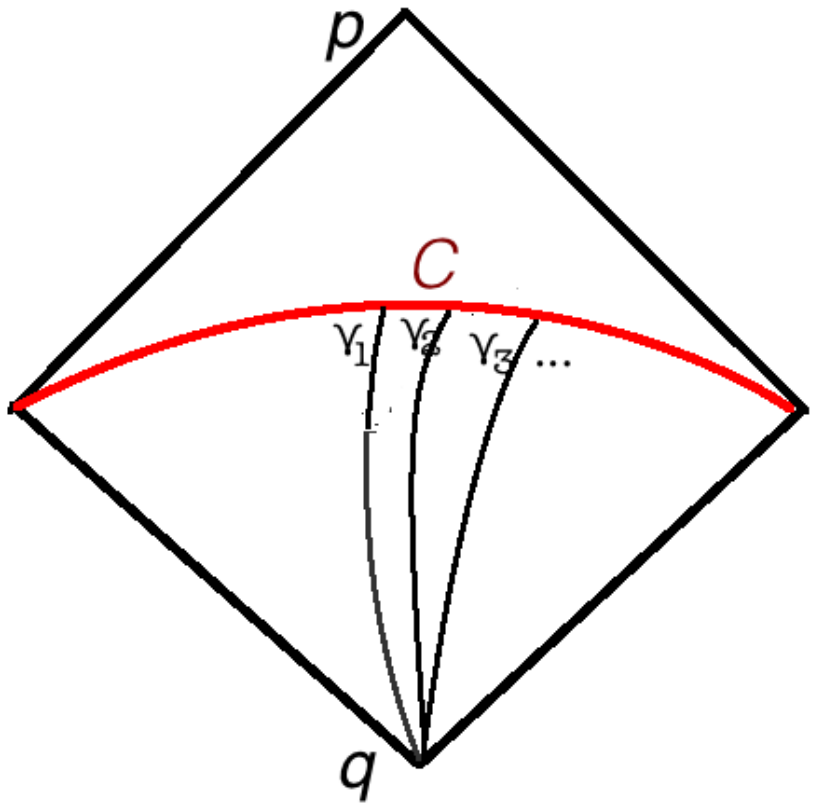
If there were a missing point, one could not predict what happens to its future



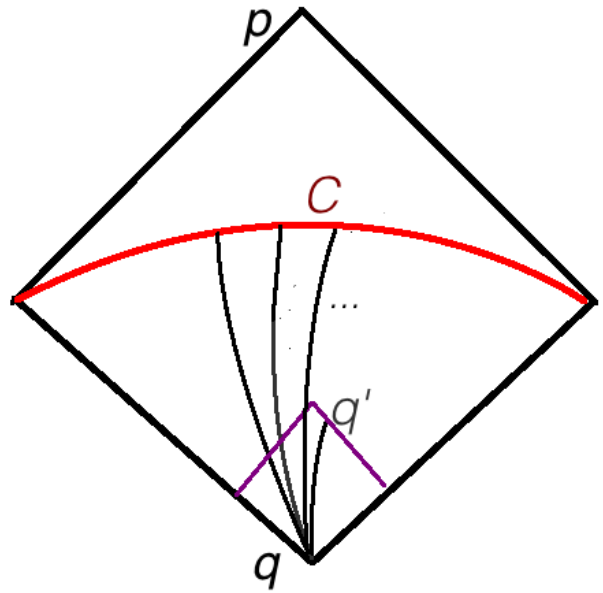
So conversely, a spacetime that develops from an initial value surface in a way that is determined by the equations does not have a missing point.

It is not totally clear that only globally hyperbolic spacetimes should be considered in General Relativity, but they are the ones that match the traditional idea of solving the equations to predict the present and the future in terms of the past. So alternatives are more exotic.

Globally hyperbolic spacetimes have the property that spaces of causal paths with suitable conditions on the endpoints are compact. For instance, let $\mathcal{C}_q^C$ be the space of causal paths that go from q – in the past of C – to C. The space of such paths is compact, as one can see by considering a sequence of paths $\gamma_n \in \mathcal{C}_q^C$:

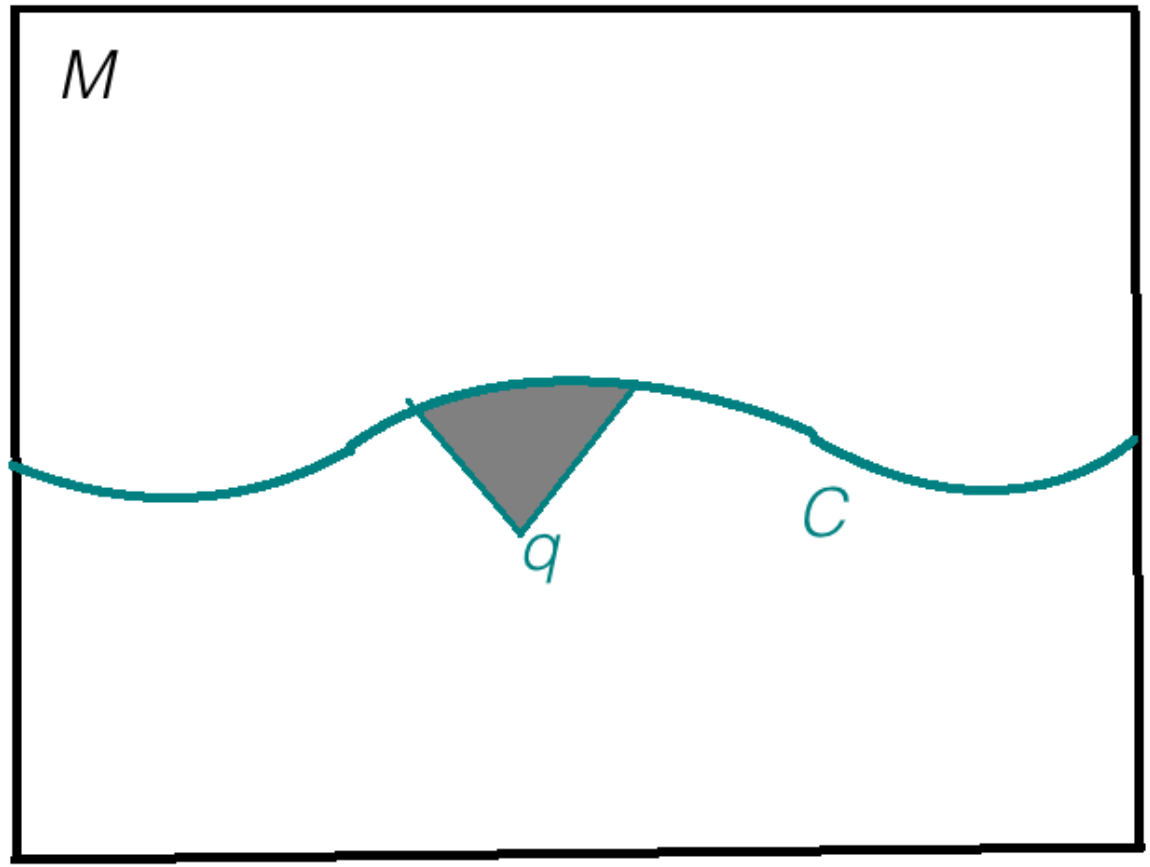


Inside the future of q, there is a small piece that looks like a causal diamond in Minkowski spacetime and in particular is compact. If we restrict the γ_n to that diamond, we can make the same argument as in Minkowski space, showing that a subsequence converges to some path γ_* from q to a point q' on the boundary of the diamond:



Now we start at q', and continue in the same way. A further subsequence converges past q'. We keep going and learn that a subsequence of the γ_n converges all the way up to C. Global hyperbolicity ensures that we can always continue.

In particular, this implies that the subspace D_q^C of C that can be reached from a point q in its past is itself compact:

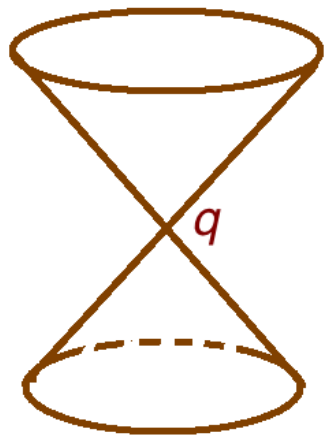


Of course, all this remains true if one exchanges "past" and "future"; in other words, similar remarks apply to a point p in the future of C.

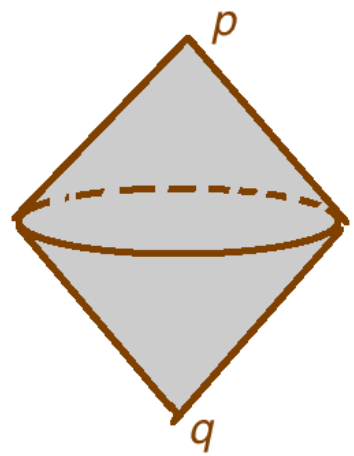
Exercise: Using what I've explained so far, show that in a globally hyperbolic spacetime M, if p is a point that can be reached by a causal path from q – for instance a future-directed causal path – then $\mathcal{C}_q^p$, the space of all causal paths from q to p, is compact. (Likewise the space of points that can be visited by such a path is compact. This space is the causal diamond D_q^p.) Hint: Since M is globally hyperbolic, it has a Cauchy hypersurface C. Consider separately the cases, for example, that q and p are both to the past of C or q is to the past and p to the future.

Compactness of the spaces of paths implies (just as in our discussion in Minkowski spacetime) that in a globally hyperbolic spacetime, there is a causal path of maximal proper time from any point q to any given initial value surface C, and also from any point q to any point p in its future.

To recapitulate a couple of points from yesterday, the points that can be reached by a causal path from a point q in Minkowski space make up its past and future light cones



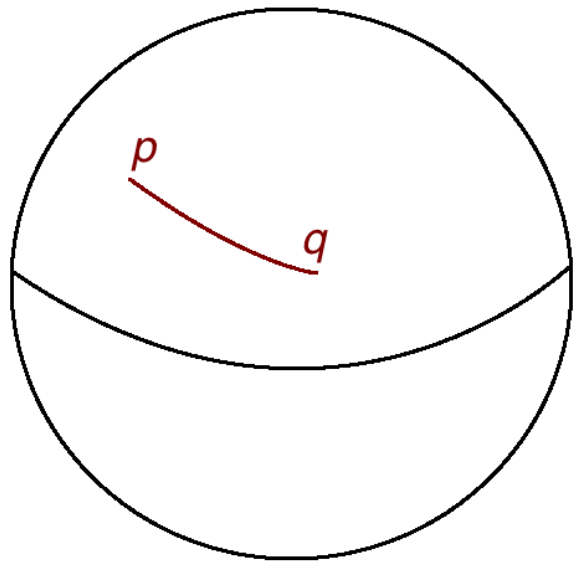
The points that can be reached by a future-going causal path from a point q to a point p in its future consist of the intersection of the future light cone of q (including its interior) with the past light cone of p (plus interior):



This is commonly called a causal diamond D_q^p. As I explained, causal diamonds are compact in any globally hyperbolic spacetime.

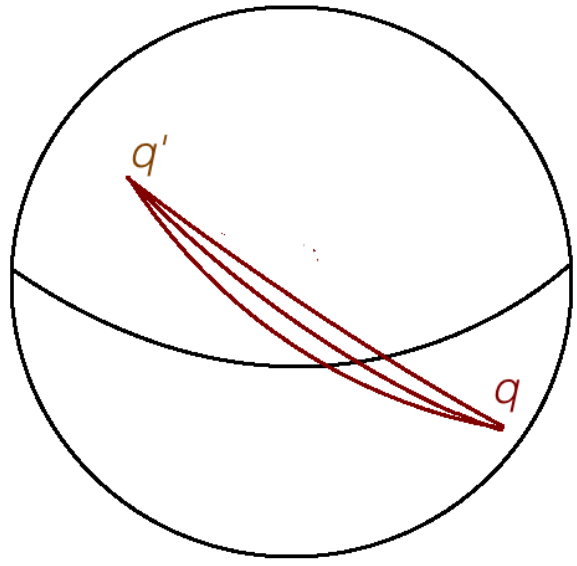
From here we will navigate to the easiest-to-explain non-trivial result about singularities. This means not following the historical order. The easiest result to explain is Hawking's theorem about the Big Bang singularity in traditional cosmology without inflation. It is easier to explain because it only involves timelike geodesics, while more or less all the other applications we will discuss involve the slightly subtler case of null geodesics.

We will start in ordinary Riemannian geometry, where we have more intuition, and then we will go over to the Lorentz signature case. Here is a question: In Riemannian geometry, is a geodesic the shortest distance between two points? Yes for a short enough geodesic, but in general no if one goes too far. A geodesic extremizes the length but may not minimize it. Here is a picture on a two-sphere



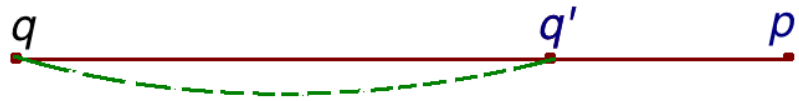
A geodesic from q to p that goes less than halfway around the sphere is the unique shortest path between those two points.

But any geodesic that leaves q and goes half-way around the sphere will arrive at a "focal point" q' (also called a caustic) on the other side of the sphere.



A geodesic that is continued past the focal point is no longer the shortest path to its destination, as one can do better by "slipping the path around the sphere."

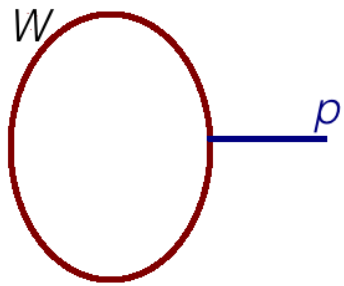
There is a general reason for this. If a geodesic qp contains a focal point q'



then the qq' part of the geodesic can be slightly displaced to a nearby geodesic also connecting the two points qq'. 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. So the original geodesic qp was not length minimizing.

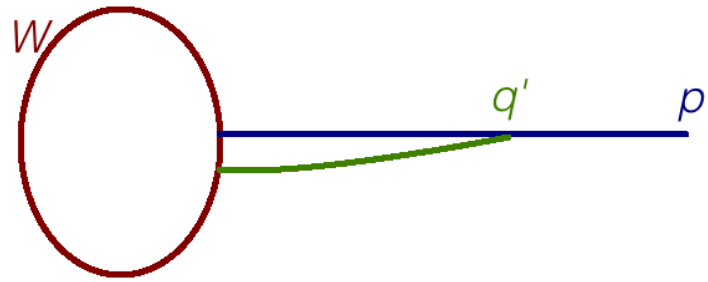
It is not important that *all* geodesics from q are focused to q' (as happens in the case of a sphere). To ensure that the geodesic $qq'p$ is not length minimizing, it is sufficient that there is *some* direction in which the qq' part can be displaced, not changing its endpoints. We do not even need to know that the qq' geodesic can be displaced *exactly* as a geodesic. We only need to know that it can be displaced while still solving the geodesic equation in first order. That ensures that the displacement does not change the length function in second order. Rounding off the kink in $qq'p$ does reduce the length in second order, so displacing the qq' segment and rounding off the kink will reduce the length if the displacement caused no increase in second order.

Often, we are interested in a length-minimizing path, not from a point q to a point p, but from some initial set W to a point p. (This will be the situation when we are proving Hawking's singularity theorem.) The simple case is that W is a submanifold. A path that *extremizes* the distance from W to p is now a geodesic that is orthogonal to W:



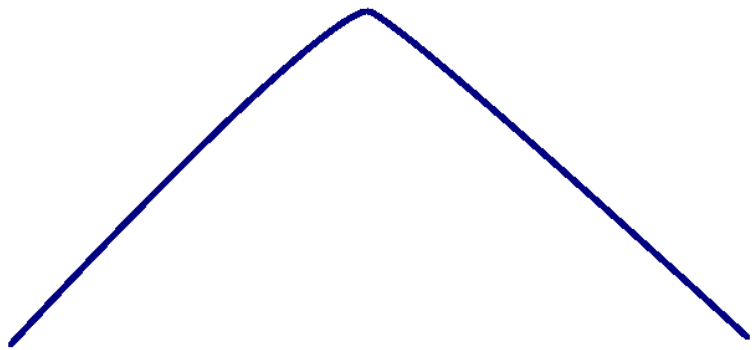
Just as before, if p is close enough to W, such a geodesic minimizes the length among paths from W to p.

But if we continue past a focal point, the geodesic ceases to be length-minimizing:



The appropriate definition of focal point is now a little different. A focal point q' on a geodesic ℓ from W to p (with ℓ orthogonal to W) is defined to be a point on ℓ that can be reached from W by another nearby geodesic ℓ', also orthogonal to W. The reasoning showing that a geodesic with a focal point is not length-minimizing is the same as before; displacing ℓ to ℓ' does not change the length, and rounding off the kink reduces the length.

Now we go over to Lorentz signature. What we have said has no good analog for *spacelike* geodesics. 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

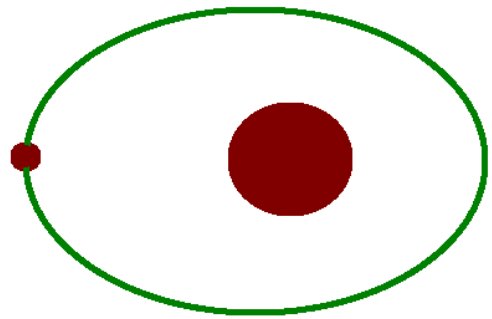


or arbitrarily long.

However, what we have said does have a close analog for *timelike* geodesics. Here we should discuss the elapsed proper time of a geodesic (not the length) and spatial fluctuations tend to reduce it. So a sufficiently short segment of any timelike geodesic *maximizes* the elapsed proper time. But if we continue a timelike geodesic past a focal point, it no longer maximizes the proper time

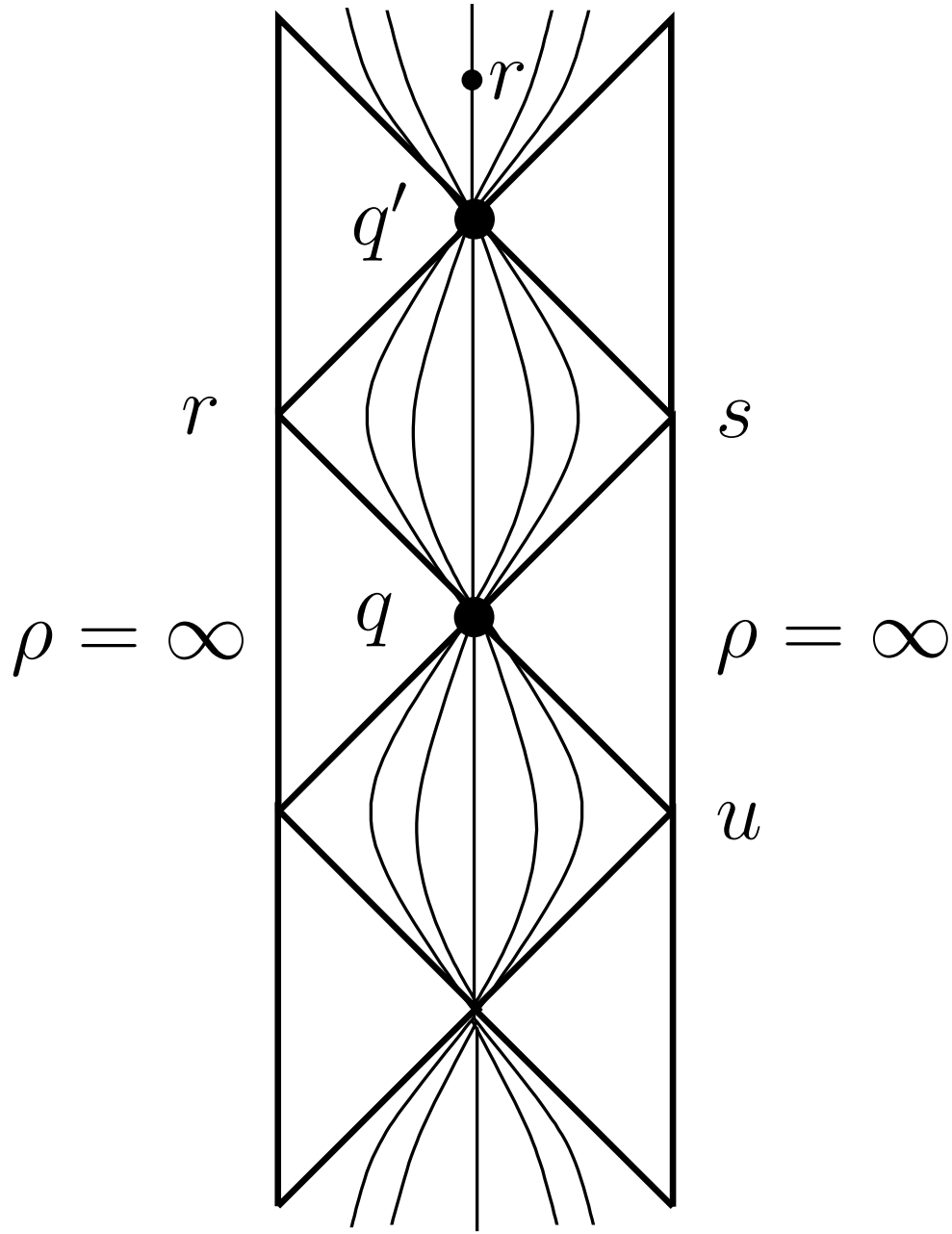
for much the same reason as in the Euclidean signature case. Rounding out the kink would increase the proper time.

I will give two examples of timelike geodesics that do not maximize the proper time between the initial and final points. First is in the motion of the Earth around the Sun.



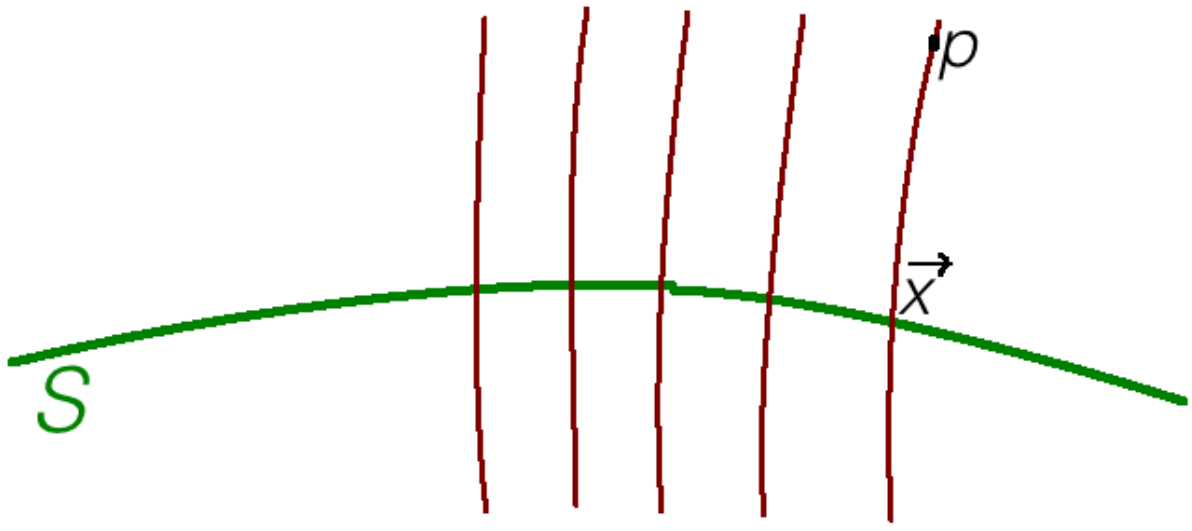
If you follow this motion over many orbits, you get a geodesic that does not maximize the proper time. One could do better by launching a spaceship into space with almost the escape velocity from the Solar System, so that after a very long time the rocket falls back to Earth. The elapsed proper time is greater for the rocket than for the Earth because it is less affected both by the gravitational redshift and by the Lorentz time dilation.

A second example is in Anti de Sitter spacetime.



In general, focal points are very easy to come by, because a very slight gravitational attraction can cause geodesics to eventually converge. To prove a singularity theorem, we need a good way to prove that timelike geodesics develop focal points. This is provided by the Raychaudhuri equation (actually the original equation due to Raychaudhuri (1955), not a more subtle variant with null geodesics that was introduced later). In D dimensions, we consider a spacetime with an initial value surface S with coordinates $\vec{x} = x^i$, $i = 1, \ldots, d = D - 1$.

By looking at timelike geodesics orthogonal to S, we construct a coordinate system on spacetime:



If a point p is on a geodesic that meets S orthogonally at $\vec{x}$, and p is a proper time t to the future of $\vec{x}$ (along the geodesic) then we assign to p the coordinates $t, \vec{x}$.

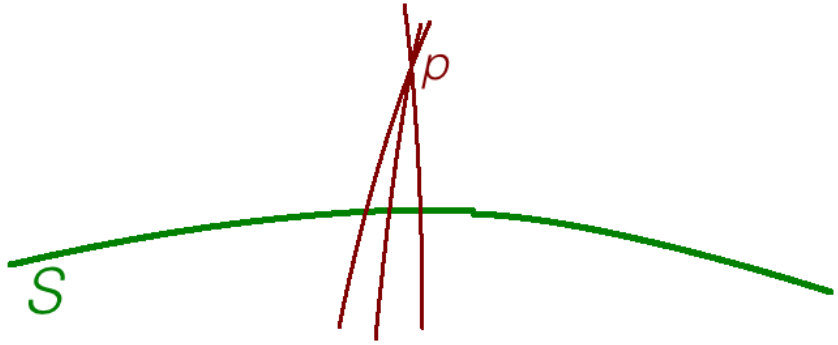
In this coordinate system, the metric is

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + \sum_{i,j=1}^{d} g_{ij}(t, \vec{x})\mathrm{d}x^i \mathrm{d}x^j.$$

Thus, this coordinate system could be described by imposing the gauge conditions $g_{00} = -1$, $g_{0i} = 0$.

The reason that we used the more geometric language of orthogonal geodesics is that this will help us understand how the coordinate system can break down.

Even if the spacetime does not become singular, our coordinate system breaks down at focal points



since a focal point cannot be labeled by which orthogonal geodesic it is on. Since $g_{ij}(t, \vec{x})$ measures the distance between neighboring geodesics, a sufficient criterion for a focal point is

$$\det g_{ij} \to 0.$$

(It can be shown that this condition is necessary as well as sufficient.)

The Raychaudhuri equation gives a sufficient condition to ensure that

$$\det g_{ij} \to 0,$$

and therefore that a focal point (or maybe a spacetime singularity) develops, within a known proper time. The Raychaudhuri equation is just the Einstein equation

$$R_{00} = 8\pi G \left(T_{00} - \frac{1}{2} g_{00} T^{\alpha}_{\alpha} \right)$$

in the coordinate system defined by the orthogonal geodesics.

A straightforward computation in the metric

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + \sum_{i,j=1}^{d} g_{ij}(t,\vec{x})\mathrm{d}x^i\mathrm{d}x^j$$

shows that

$$\begin{aligned} R_{00} &= -\frac{1}{2}\partial_t(g^{ik}\partial_t g_{ik}) - \frac{1}{4}\left(g^{ik}\partial_t g_{kj}\right)\left(g^{jm}\partial_t g_{mi}\right) \\ &= -\frac{1}{2}\partial_t \mathrm{Tr}\, g^{-1}\dot{g} - \frac{1}{4}\mathrm{Tr}(g^{-1}\dot{g})^2. \end{aligned}$$

It is customary to define

$$A = \sqrt{\det g}$$

which measures the area (or volume) occupied by a little bundle of geodesics. Then

$$\frac{\dot{A}}{A} = \frac{1}{2} \mathrm{Tr}\, g^{-1} \dot{g}.$$

The quantity $\dot{A}/A$ is called the expansion. (It is often denoted as θ.)

It is convenient to also define the traceless part of $g^{-1}\dot{g}$ ("the shear")

$$M^i_j = g^{ik}\dot{g}_{kj} - \frac{1}{d}\delta^i_j \mathrm{Tr}\, g^{-1}\dot{g}.$$

Then

$$R_{00} = -\partial_t\left(\frac{\dot{A}}{A}\right) - \frac{1}{d}\left(\frac{\dot{A}}{A}\right)^2 - \frac{1}{4}\mathrm{Tr}\, M^2.$$

If we define

$$\widehat{T}_{\mu\nu} = T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T^\alpha_\alpha,$$

then the Einstein-Raychaudhuri equation $R_{00} = 8\pi G\,\widehat{T}_{00}$ becomes

$$\partial_t\left(\frac{\dot{A}}{A}\right) + \frac{1}{d}\left(\frac{\dot{A}}{A}\right)^2 = -\frac{1}{4}\mathrm{Tr}\, M^2 - 8\pi G\,\widehat{T}_{00}.$$

The *strong energy condition* is the statement that

$$\widehat{T}_{00} \geq 0$$

at every point and in every local Lorentz frame. It is satisfied for nonrelativistic matter, for radiation, and for a negative cosmological constant. The outstanding example that does *not* satisfy the strong energy condition is a positive cosmological constant. If we assume the strong energy condition, then all the terms on the right hand side of the Einstein-Raychaudhuri equation are negative and so we get an inequality

$$\partial_t \left(\frac{\dot{A}}{A} \right) + \frac{1}{d} \left(\frac{\dot{A}}{A} \right)^2 \leq 0.$$

Equivalently,

$$\partial_t \left(\frac{1}{\dot{A}/A} \right) \geq \frac{1}{d}.$$

Now we can get a useful condition for the occurrence of focal points. Let us go back to our initial value surface

and assume that $\dot{A}/A < 0$ at some point on this surface, say $\dot{A}/A = -\lambda$, $\lambda > 0$. So the initial value of $1/(\dot{A}/A)$ is $-1/\lambda$ and the inequality

$$\partial_t \left(\frac{1}{\dot{A}/A} \right) \geq \frac{1}{d}$$

implies that $1/(\dot{A}/A)$ goes to 0 (from below) within a time d/λ.

Equivalently $\dot{A}/A \to -\infty$ within a time d/λ. In other words, the future-going orthogonal geodesic leaving S at any point at which $\dot{A}/A = -\lambda < 0$ either meets a singularity $\dot{A} \to \infty$ or a focal point $A \to 0$ within a time d/λ.

(Focal points are easy to come by because of gravitational lensing. In any situation in which the gravitational fields remain weak, we will get a focal point, not a singularity.)

Now we are ready for Hawking's theorem. Hawking assumed that the Universe is globally hyperbolic with initial surface S.

He also assumed the strong energy condition, i.e. he assumed that the stress tensor is made of ordinary matter and radiation. (The inflationary Universe, which gives a way to avoid Hawking's conclusion because a positive cosmological constant does not satisfy the strong energy condition, was still in the future.) If the Universe is perfectly homogeneous and isotropic, it is described by the Friedmann-Lemaître-Robertson-Walker solution and emerged from the Big Bang at a calculable time in the past.

Suppose, however, more realistically, that the initial surface is not perfectly homogeneous

but that the local Hubble constant is everywhere positive. Did such a Universe emerge from a Big Bang? One could imagine that following the equations back in time, the inhomogeneities become more severe, the FLRW solution is not a good approximation, and part or most (or maybe even all) of the Universe did not really come from an initial singularity.

Hawking, however, proved that assuming the strong energy condition and assuming that the Universe is globally hyperbolic, this is not the case. To be more exact, he showed that if the local Hubble constant has a positive minimum value $h_{\min}$ on an initial value surface S

then there is no point in spacetime that is a proper time more than $h_{\min}$ to the past of S.

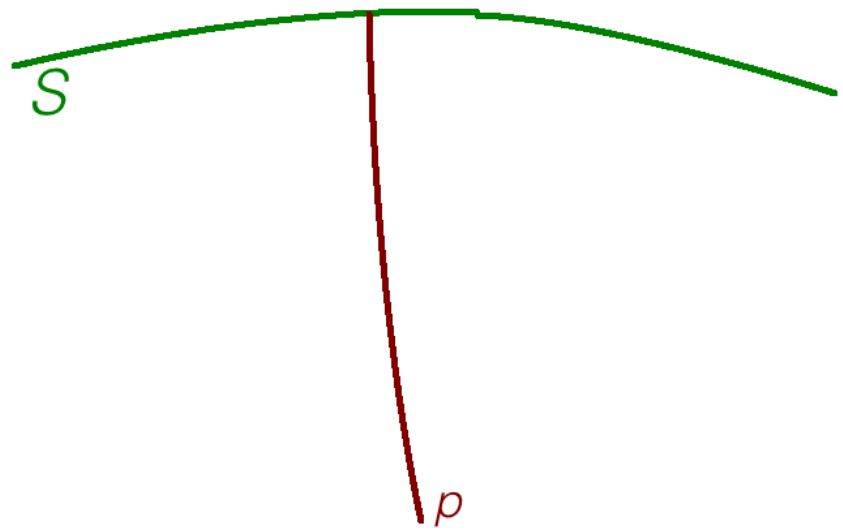
I should explain precisely what I mean by the local Hubble constant. For a homogeneous isotropic expansion such as

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + R^2(t)(\mathrm{d}\vec{x})^2$$

one usually defines the Hubble constant by $h = \dot{R}/R$. This is the same as $h = \dot{A}/dA$ (since $A = \sqrt{\det g}$ is the same as R^d in the homogeneous isotropic case). We will use that definition in general, so the assumption on the Hubble constant is $\dot{A}/A \geq dh_{\min}$, assuming time is measured towards the future. But we will measure time towards the past and so instead we write the assumption as

$$\frac{\dot{A}}{A} \leq -dh_{\min}.$$

Hawking's proof 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.



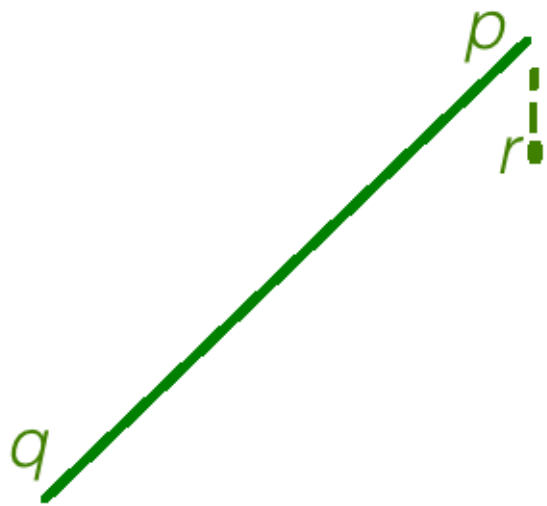
As we know, such a path is a timelike geodesic without focal points that is orthogonal to S, as shown.

(2) But the assumption that the initial value of $\dot{A}/A$ on the surface S is everywhere $\leq -dh_{\min}$ implies that any past-going timelike geodesic orthogonal to S develops a focal point within a proper time at most $1/h_{\min}$.

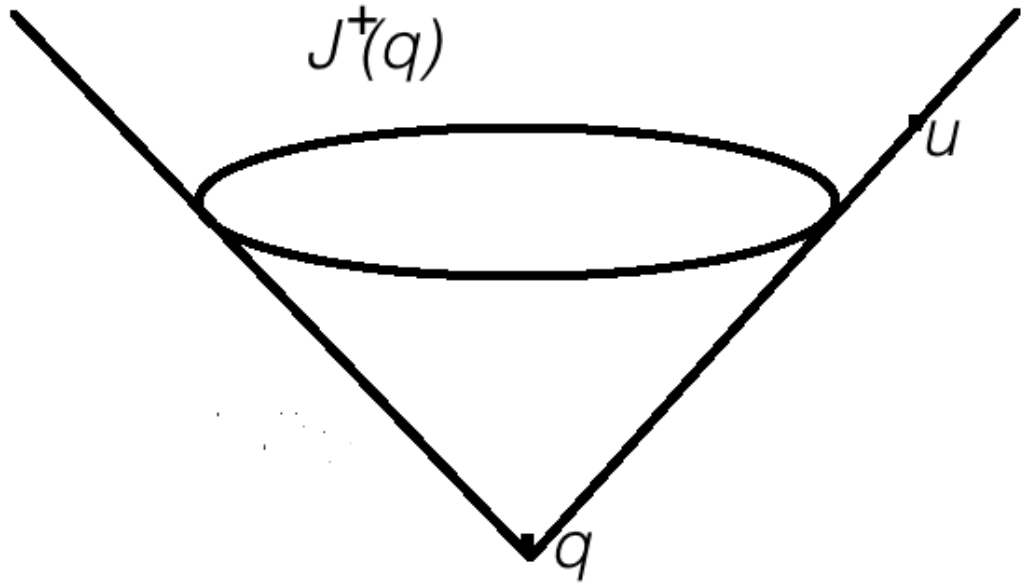
Combining the two statements, we see that there is no point in spacetime that is to the past of S by a proper time more than $1/h_{\min}$. Thus (still assuming the strong energy condition) the minimum value of the local Hubble constant gives an upper bound on how long anything in the Universe could have existed in the past.

This is Hawking's theorem about the Big Bang.

Hopefully, this was enough fun so that you are all anxious to know what we can learn by studying null geodesics. Any null geodesic has zero elapsed proper time. Nevertheless there is a good notion that has similar properties to "maximal proper time" for timelike paths. We will say that a causal path from q to p is "prompt" if there is no causal path from q to p that could have arrived sooner. To be precise, the path ℓ from q to p is prompt if there is no causal path ℓ' that would have arrived at a point r near p and to its past:



This is a very severe condition. More or less by definition, it means that p must be in the boundary of the causal future of q. By the causal future $J^+(q)$, we mean all points that can be reached from q by a future-going causal path. For instance, in Minkowski spacetime, $J^+(q)$ consists of all points in or on the future light cone of q:



In Minkowski spacetime, every null geodesic is prompt, and every point on the boundary $\partial J^+(q)$ is connected to q by a prompt null geodesic.

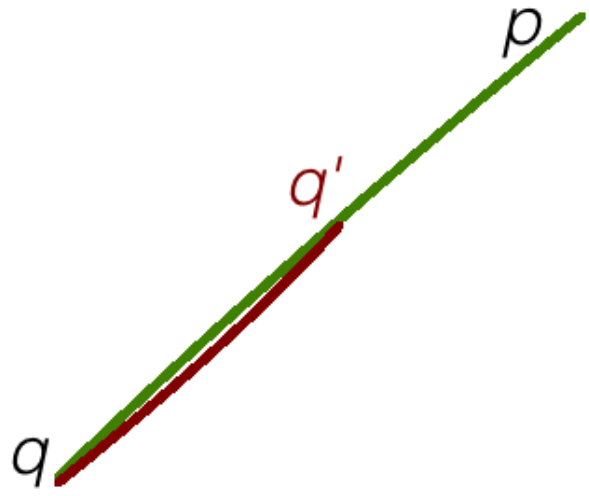
Even though prompt causal paths are very exceptional, they are the useful analogs of proper time maximizing timeline paths. Let us explore their properties.

A short enough segment of any null geodesic in any Lorentz signature spacetime is prompt (as a path between its endpoints), just as if one were in Minkowski spacetime. But if continued, a null geodesic may become non-prompt because of gravitational lensing. For example, when we see multiple images of the same supernova explosion, the images do not arrive at the same time and clearly the ones that do not arrive first are not prompt.

A causal path ℓ whose tangent vector is somewhere timelike (rather than null) cannot be prompt, because by modifying ℓ slightly to be everywhere null, we could find a nearby causal path that would arrive in the past of p. Actually, ℓ has to be a null geodesic since if it bends anywhere

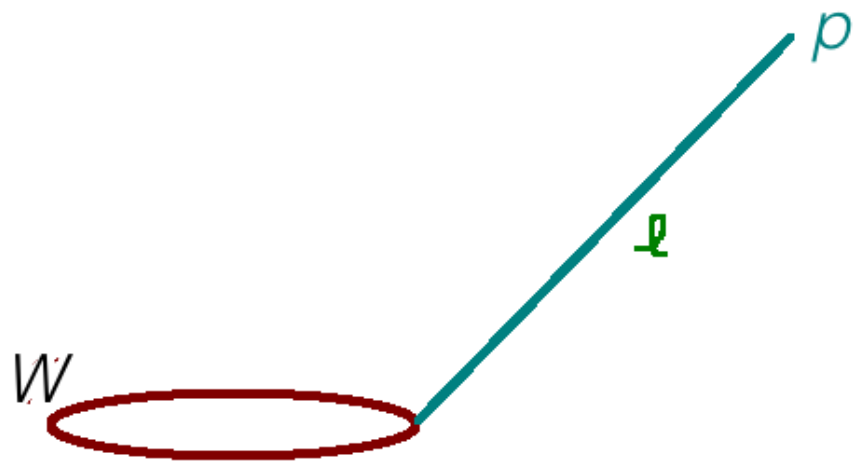
it is again not prompt.

Finally, to be prompt, a null geodesic from q to p must contain no focal point. Here q' is a focal point if the qq' segment of the null geodesic qp can be slightly displaced to a nearby null geodesic from q to q':



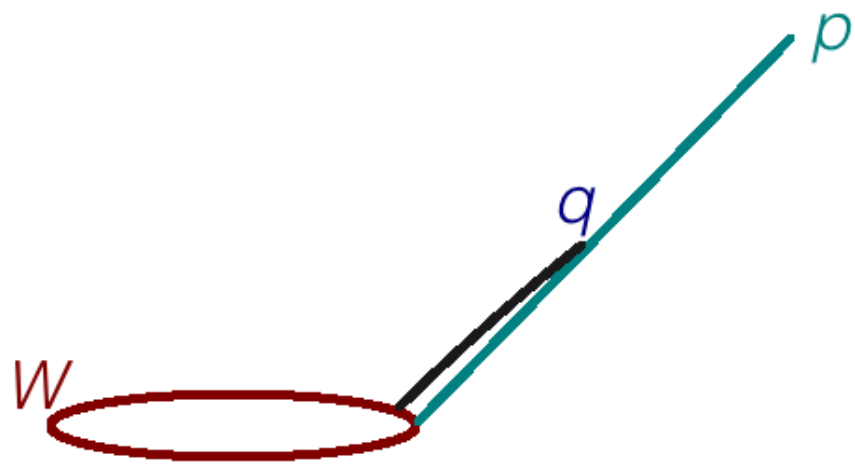
If this is the case, by rounding the corner, one can get a nearby causal path from q to p that is timelike near the corner, and the original qp null geodesic is not prompt.

Likewise, we say that a causal path from a set W to a point p is prompt if it arrives as soon as possible in the same sense. To be prompt, a causal path ℓ from W to p has to be a null geodesic, just as before. If W is a submanifold of spacetime, then in addition ℓ has to be orthogonal to W:



Otherwise, by changing the initial point of ℓ a little, one can get a causal path to p that is not everywhere null and which can be deformed to arrive a little sooner. Orthogonality between a spacelike set W and a null curve that intersects it is only possible if W has real codimension at least 2.

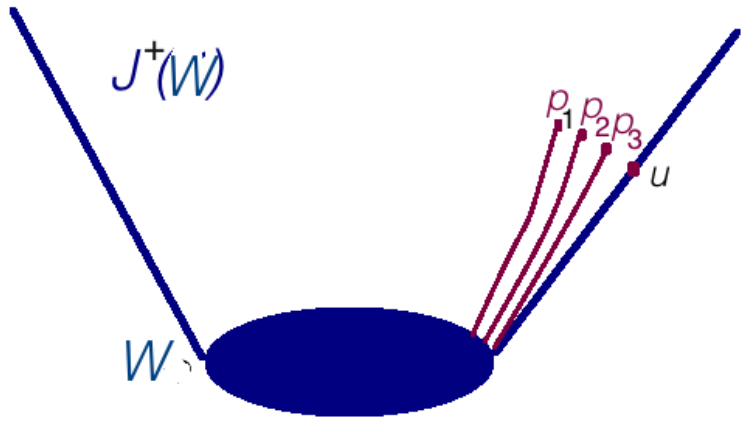
A prompt path ℓ from W to p must again have no focal point, where now q is a focal point if the segment of ℓ that connects W to q can be deformed slightly to a nearby null geodesic, also connecting W to q and orthogonal to W:



The reasoning is the same as before.

In a general globally hyperbolic spacetime – such as, maybe, the real Unverse – a null geodesic can develop a focal point because of gravitational lensing, so not every null geodesic is prompt. But it is true, just as in Minkowski space, that every point in $\partial J^+(q)$ – the boundary of the causal future of q – is connected to q by a prompt null geodesic. This is actually true if q is replaced by any closed spacelike set W: writing $J^+(W)$ for the causal future of W, any point p in the boundary $\partial J^+(W)$ is connected to W by a prompt null geodesic.

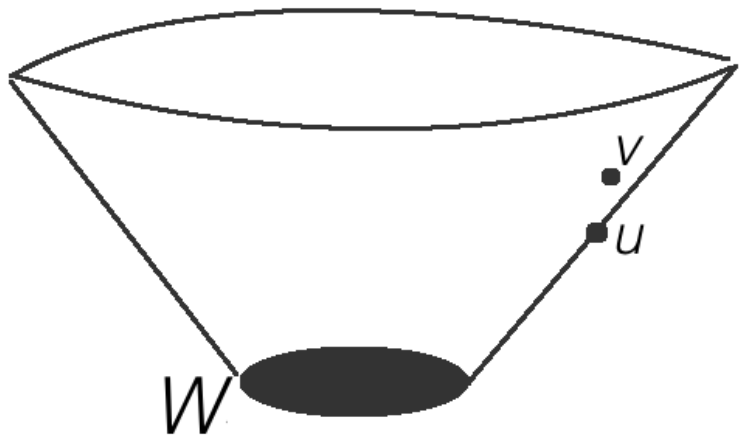
For the proof: Let u be a point in $\partial J^+(W)$. It is the limit of a sequence of points p_i in the interior of $J^+(W)$:



Each p_i can be reached from W by a causal path γ_i.

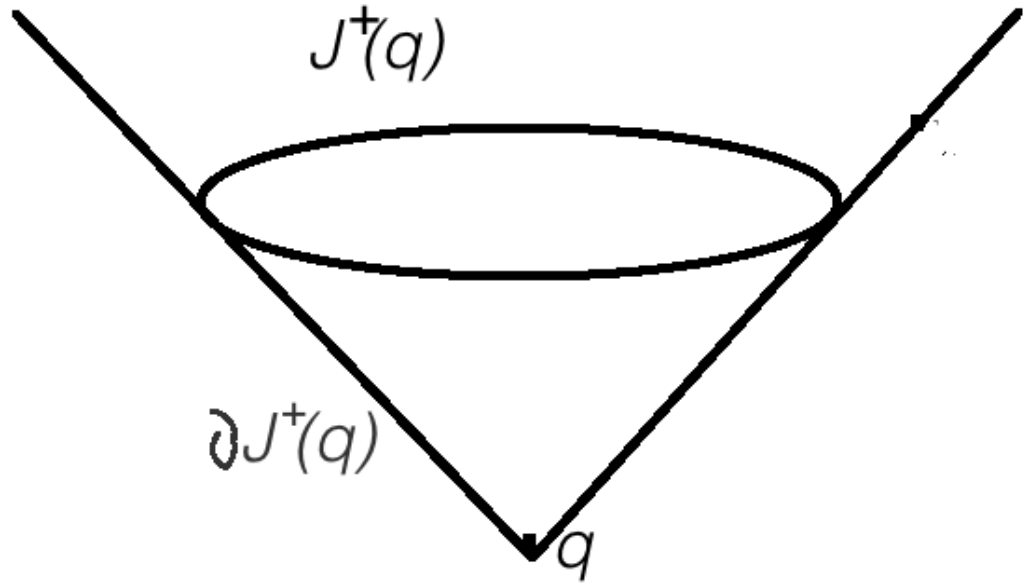
In a globally hyperbolic spacetime, by choosing a convergent subsequence of the γ_i, one can find a causal path γ from W to u. If u is really in the boundary of $J^+(W)$, the path γ is automatically prompt. After all, if some nearby path arrives in the past of u, then u is not in the boundary of $J^+(W)$.

In general, a set C is called *achronal* if there are no pairs of points $u, v \in C$ with v connected to u by a timelike path. The boundary of the future of any set W is always achronal



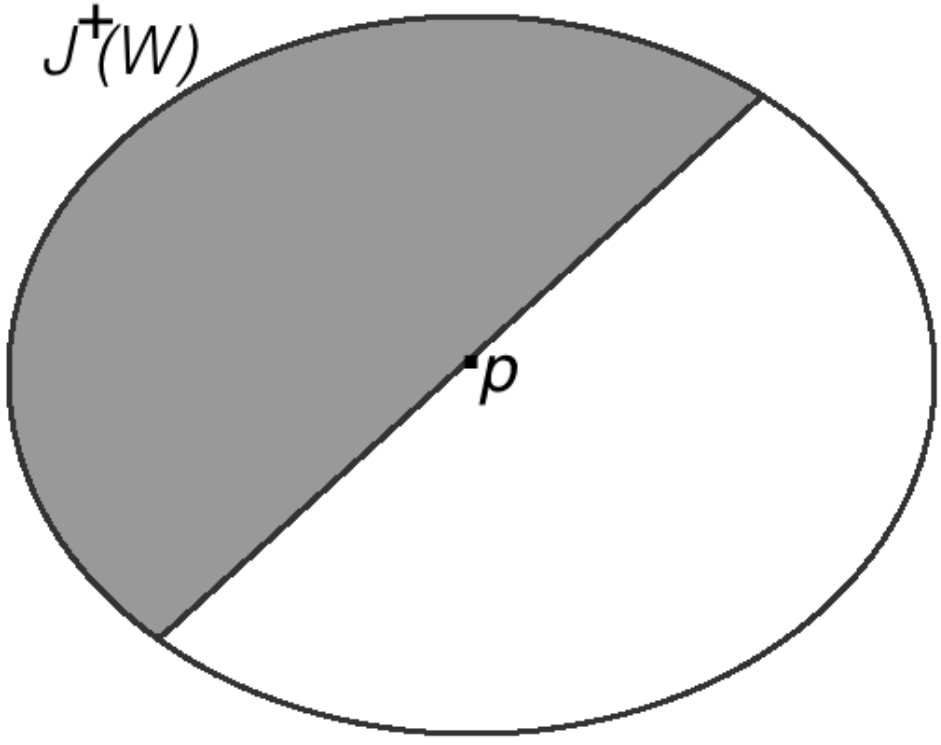
since if $u \in \partial J^+(W)$ and v is in the future of u, then v is in the interior of $J^+(W)$, not on its boundary.

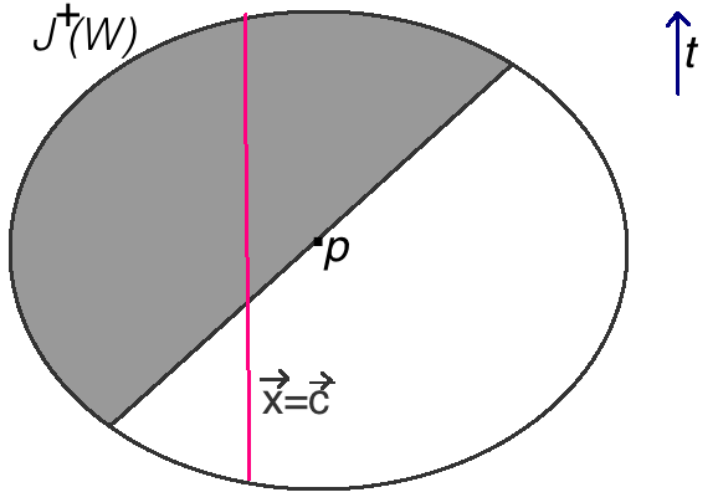
The boundary of the future of a (closed) set has one more property, which is crucial in understanding black holes: $\partial J^+(W)$ is always a $d = D - 1$-dimensional manifold without boundary, though often not smooth. In a globally hyperbolic spacetime, it is a closed submanifold. An example is the future light cone of a point q, which is a closed but not smooth submanifold of spacetime:



(Note that q itself is considered to be in $J^+(q)$ and in its boundary; in other words, here and later we allow a prompt null geodesic that consists of only one point.)

For the proof, we pick an arbitrary point $p \in \partial J^+(W)$. We want to show that $\partial J^+(p)$ is a manifold near p. We pick a small ball near p in which M can be approximated by Minkowski spacetime and pick coordinates $t, \vec{x}$ centered at p. Here is a drawing in two dimensions:



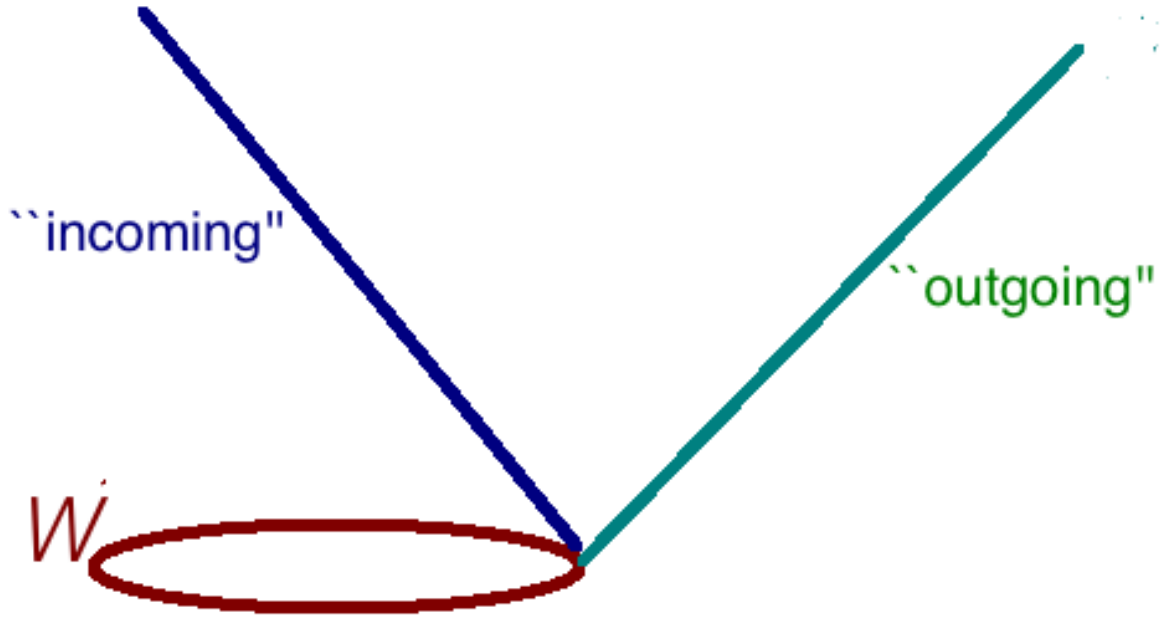


Look at the timelike path $\gamma_{\vec{c}}$ parametrized by t, with fixed $\vec{x} = \vec{c}$. This path is in the future of p and hence in the interior of $J^+(W)$ is t is positive enough (keeping $\vec{c}$ fixed); it is in the past of p and so not in $J^+(W)$ is t is negative enough. As t increases, the point $(t, \vec{c})$ can only pass once from outside $J^+(W)$ to inside and this happens at a unique t at which $(t, \vec{c}) \in \partial J^+(W)$. Hence we can take $\vec{c}$ as local coordinates for $\partial J^+(W)$ and $\partial J^+(W)$ is a manifold.

Exercise: Let W be a circle or a two-sphere, embedded in Minkowski four-space in the obvious way. Describe $\partial J^+(W)$, showing that it is a manifold, though not smoothly embedded in spacetime. What is the topology of $\partial J^+(W)$? Same question with an ellipse or an ellipsoid instead of a circle or sphere. (You should find that $\partial J^+(W)$ is equivalent topologically in each case to the initial value surface $t = 0$, where t is the time. This is consistent with a general constraint on $\partial J^+(W)$ that we will explain presently.)

Just as is in the timelike case, to learn something we need a reasonable way to be able to predict that a null geodesic will develop a focal point if continued far enough. This is provided by the (null) Raychaudhuri equation, which formally is very similar to what we already described in the timelike case.

Given a spacelike surface W of codimension 2, there are two families of future-going null geodesics that are orthogonal to W, namely the "outgoing" and "ingoing" ones.



Pick one of those families, say the outgoing one.

The outgoing (future-directed) orthogonal null geodesics from W together form a $(D-1)$-manifold U that is "null" in the sense that its metric is degenerate. We pick any coordinates $\vec{x} = x^i$, $i = 1, \ldots, D-2$ on W. We recall that a null geodesic does not have a "proper time," but there exists an affine parameter u, well-defined up to $u \to au + b$ (a, b constants) in which the geodesic equation is just

$$\frac{D^2 x^\lambda}{Du^2} = 0.$$

We normalize u to be zero along W and then it is well defined up to multiplication by a function of the x^i. So x^i, u are a coordinate system for U.

The metric of U in these coordinates is $\mathrm{d}s^2 = g_{ij}(\vec{x}, u)\mathrm{d}x^i\mathrm{d}x^j$. (This is degenerate as $\mathrm{d}u$ does not appear.) The null Raychaudhuri equation is just the Einstein equation

$$R_{uu} = 8\pi G T_{uu}.$$

It is almost the same as it was in the timelike case, except that $d = D - 1$ is replaced by $D - 2$ because there are now only $D - 2$ normal spacelike coordinates. As before, we define

$$A = \sqrt{\det g}$$

and the null expansion is defined by

$$\theta = \frac{\dot{A}}{A}, \qquad \dot{A} = \partial_u A.$$

Equivalently

$$\theta = \frac{1}{2} g^{ik} \partial_u g_{ik} = \frac{1}{2} \mathrm{Tr}\, g^{-1} \dot{g}.$$

One also defines M^i_j to be the tracefree part of $g^{-1}\dot{g}$ (the "shear"). Then the Einstein-Raychaudhuri equation is

$$\partial_u \left(\frac{\dot{A}}{A}\right) + \frac{1}{D-2}\left(\frac{\dot{A}}{A}\right)^2 = -\frac{1}{4}\mathrm{Tr}\, M^2 - 8\pi G T_{uu}.$$

The *null energy condition* is the statement that at each point and in each local Lorentz frame,

$$T_{uu} \geq 0.$$

It is not affected by a cosmological constant and is satisfied by any of the usual relativistic classical fields.

Assuming the null energy condition, the Einstein-Raychaudhuri equation gives

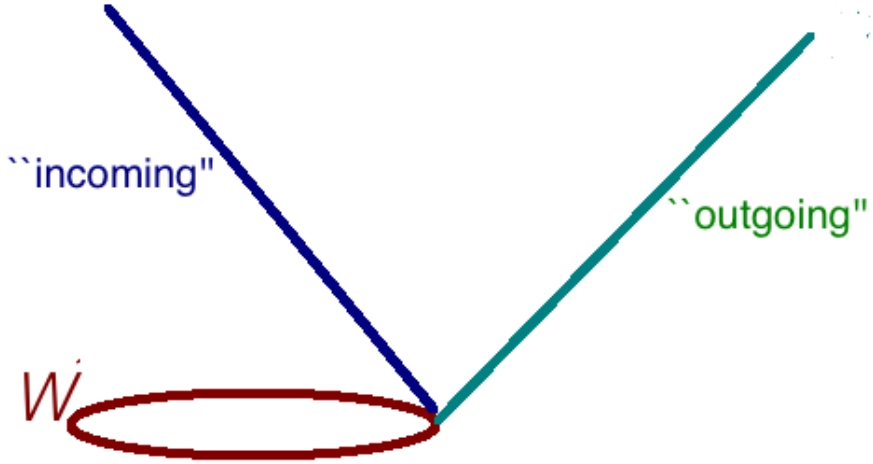
$$\partial_u \left(\frac{\dot{A}}{A} \right) + \frac{1}{D-2} \left(\frac{\dot{A}}{A} \right)^2 \leq 0.$$

By exactly the same steps as in the timelike case, we deduce from this that if, at a given point on W, the initial value of the null expansion is $\dot{A}/A = -\lambda$, $\lambda > 0$, then that geodesic will reach a focal point $A = 0$ (or possibly a singularity $\dot{A} = \infty$) at a value of the affine parameter

$$u \leq \frac{D-2}{\lambda}.$$

This knowledge about focal points of null geodesics ultimately leads to singularity theorems, just as we described in the timelike case.

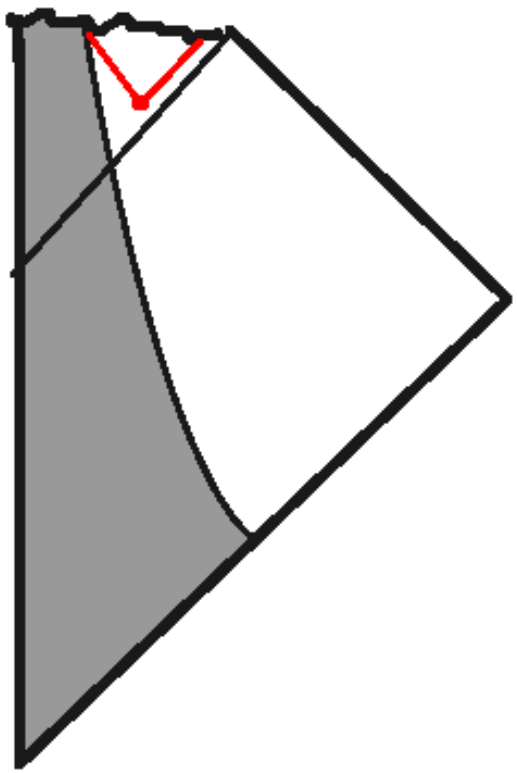
Recalling that, for each W, there are two families of future-going orthogonal null geodesics, "incoming" and "outgoing"



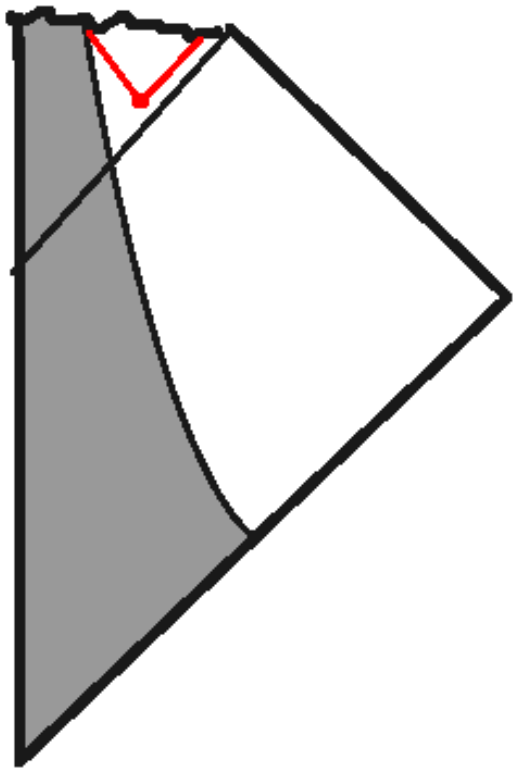
we see that there are two null expansions. If W is, for example, a sphere, embedded in the obvious way in Minkowski spacetime, then one null expansion is positive and one is negative.

Exercise: Describe an embedding of a sphere in Minkowski spacetime such that each of the null expansions is positive in one region and negative in another.

Penrose defined a "trapped surface" to be a *compact* spacelike codimension two surface W such that both null expansions are everywhere negative. The motivating example is a surface behind the horizon of a Schwarzschild black hole, represented by a point in the Penrose diagram:



Remember that a point on the Penrose diagram represents a spacelike sphere S (of dimension $D-2$) whose area, for $D=4$ for example, is $4\pi r^2$. Behind the horizon, r becomes a "timelike" coordinate and decreases along every causal path. So a point behind the horizon corresponds to a sphere whose expansions are both negative

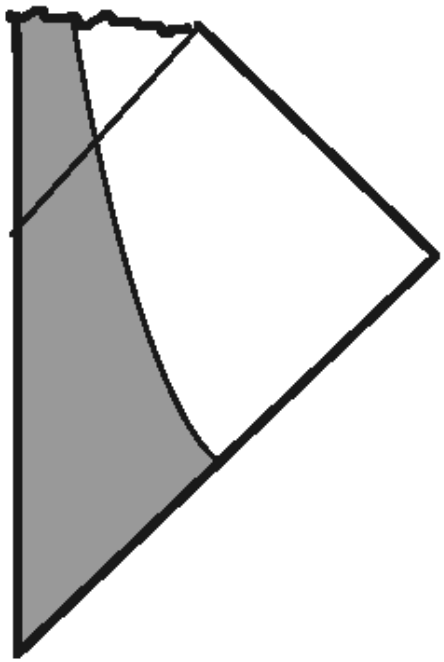


and this is the motivating example of a trapped surface.

Before explaining what Penrose proved about trapped surfaces, we need to know a couple more facts about a globally hyperbolic spacetime M with initial value surface S: If C is any achronal set in M, it is topologically equivalent to a subset of S. Roughly speaking that is because if one picks a time direction, then the flow in the time direction maps C to M. Here it is important for C to be achronal:

The special case of this that Penrose used is the following: In a globally hyperbolic spacetime M of dimension D, with *noncompact* (connected) initial value surface S, for any subset C, the boundary of the future of C, namely $\partial J^+(C)$, cannot be compact. For $\partial J^+(C)$ is an achronal manifold with the same dimension $d = D - 1$ as S. As such it will be topologically equivalent to a submanifold of S. But a noncompact (connected) manifold S doesn't have any compact submanifold of the same dimension. So in such a spacetime, for any C, $\partial J^+(C)$ is not compact.

Now let me explain what Penrose did. Spherically symmetric collapse to a black hole is described by the familar picture



Behind the horizon a singularity forms.

For the spherically symmetric case, one can just solve the equations and demonstrate the formation of a singularity. But what happens if the geometry is not quite spherically symmetric? Does infalling matter still collapse to a singularity, or does it "miss"? Penrose wanted a robust criterion for formation of a singularity that would not depend on precise spherical symmetry. For this, he introduced the idea of a trapped surface.

He showed that a singularity – or at least a failure of predictivity – must occur once a trapped surface forms. Obviously, the condition for a spacetime to contain a compact trapped surface is stable against small perturbations of the geometry, so Penrose's result shows that singularities form generically in gravitational collapse.

To be precise, Penrose's singularity theorem says that if M is a globally hyperbolic spacetime with a noncompact initial value surface S, and C is a compact trapped surface in M, then M is geodesically incomplete: at least one orthogonal null geodesic from C cannot be continued indefinitely into the future (technically, to an infinite value of its affine parameter). This is commonly called a "singularity theorem" because of a presumption that the reason that the geodesic cannot be continued is the same as in Schwarzschild: it ends at a singularity. However, this goes beyond what is proved.

If C is a compact trapped surface, its null expansions $\dot{A}/A$, being everywhere negative, satisfy a bound

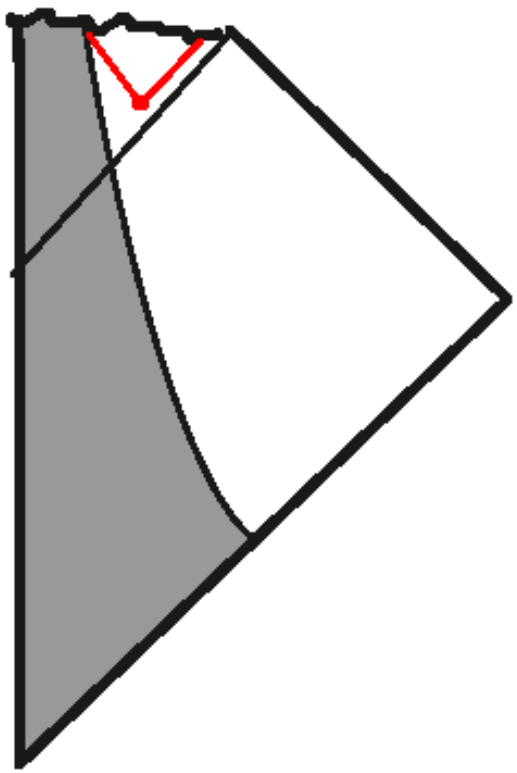
$$\frac{\dot{A}}{A} < -\lambda$$

for some constant $\lambda > 0$. Then Penrose proves, to be precise, that at least one of the future-going null geodesics orthogonal to C cannot be extended to a value of its affine parameter greater than $(D-2)/\lambda$.

Suppose on the contrary that every one of the future-going null geodesics orthogonal to C can be extended beyond an affine distance $(D-2)/\lambda$. This means, according to Raychaudhuri's equation, that they can be extended beyond their first focal point. Now think about $\partial J^+(C)$, the boundary of the future of C. It consists of points on the future-going orthogonal null geodesics from C that are not beyond focal points. If a given geodesic ℓ can be extended beyond its first focal point, then the part of ℓ that is in $\partial J^+(C)$ is compact. Since C itself is compact, this implies that $\partial J^+(C)$ is compact.

We proved before that it is impossible for $\partial J^+(C)$ to be compact, so the conclusion is that at least one of the orthogonal null geodesics in question cannot be extended in the way that was assumed.

Notice that if a future-going null geodesic ends at a singularity before reaching a focal point, then it is not compact and may be entirely contained in $\partial J^+(C)$, which would then also not be compact. This is what actually happens for the motivating example:



Here are some exercises to understand the fine print in Penrose's theorem:

(1) De Sitter space in dimension D can be described with the metric

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + R^2 \cosh^2(t/R)\mathrm{d}\Omega^2$$

whre $\mathrm{d}\Omega^2$ is the metric of a $D-1$-sphere. Convince yourself that this is a globally hyperbolic spacetime with a *compact* initial value surface (so Penrose's theorem doesn't apply). Also convince yourself that it is geodesically complete; all geodesics can be continued to infinite affine parameter in both directions. Find a compact trapped surface C. Can you describe the boundary of the future of C? You should find that $\partial J^+(C)$ is topologically equivalent to the initial value surface, as the above arguments imply.

(2) Consider the following metric, which describes just a portion of de Sitter space:

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + R^2 \exp(-2t/R)(\mathrm{d}\vec{x})^2, \qquad \vec{x} \in \mathbb{R}^{D-1}.$$

Show that it is globally hyperbolic with noncompact initial value surface. Thus Penrose's theorem applies. Find a compact trapped surface. Do you see a singularity? What does Penrose's theorem mean for this spacetime? What is the boundary of the future of the trapped surface?

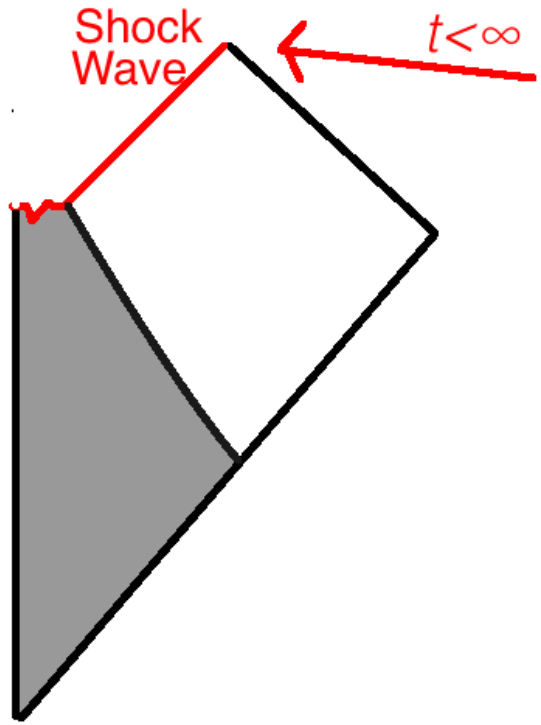
Hint: Look at null geodesics. You should find that every null geodesic intersects the initial value surface at (say) $t = 0$, as required by global hyperbolicity. But are geodesics complete in the sense of extending to infinite affine parameter in both directions?

What Penrose's theorem actually tells us about the region inside a black hole is very limited, and therefore it comes as a nice surprise that the ideas we've developed lead relatively easily to an understanding of important properties of black holes. (This is largely due to Hawking.)

To be more exact, I should say that these ideas *plus one more assumption* lead to such an understanding.

It is not possible to get a good theory of black holes without knowing, or assuming, that something worse does not happen. "Something worse" would be the formation of a naked singularity, visible to an outside observer, and possibly bringing spacetime, or at least the predictive power of classical General Relativity, to an end.

For example, imagine that gravitational collapse creates an outgoing shock wave singularity



If the singularity is bad enough that the classical Einstein equations break down, and the classical spacetime cannot be continued beyond it (based only on information provided by Einstein's theory), then we cannot use classical General Relativity to get a general theory of gravitational collapse. We would need a better theory, maybe quantum gravity or string theory.

Penrose introduced the hypothesis of "cosmic censorship," which (in its simplest form) says that this does not happen: in gravitational collapse, or any localized process in an asymptotically Minkowskian spacetime, the region in the far distance and far future continues to exist, just as in Minkowski spacetime.

Moreover, the hypothesis says that there is no "naked singularity" visible to a distant observer. Any singularity – where the classical Einstein equations break down – is supposed to be hidden behind a horizon.

If true, this is a quite remarkable and genuinely surprising fact and possibly a little disappointing. It is genuinely surprising because the classical Einstein equations have no obvious stability properties, and disappointing because we lose our chance to get observational evidence concerning a hypothetical better theory.

Personally as of 15 years ago, I thought there was very little evidence for cosmic censorship. But by now, the fact that computer simulations of black hole collisions have not generated any naked singularities has given reasonably strong evidence for cosmic censorship. Of course, recent observations of gravitational waves have greatly enhanced the interest of these simulations.

Whether cosmic censorship is true – and how exactly to formulate it; I have here omitted numerous important technical issues – is regarded by many as the outstanding unanswered question about classical General Relativity.

If cosmic censorship is assumed, one can make a nice theory of black holes. First of all, the *black hole region* in spacetime is the region B this is not visible to an outside observer. To be somewhat more exact, let $\mathcal{I}$ be the worldline of a timelike observer who remains more or less at rest at a great distance, in the asymptotically flat region observing whatever happens. We denote as $J^-(\mathcal{I})$ the causal past of this observer, i.e. all the points from which the observer can receive a signal. If M is the full spacetime, then the black hole region B is the complement of $J^-(\mathcal{I})$ in M:

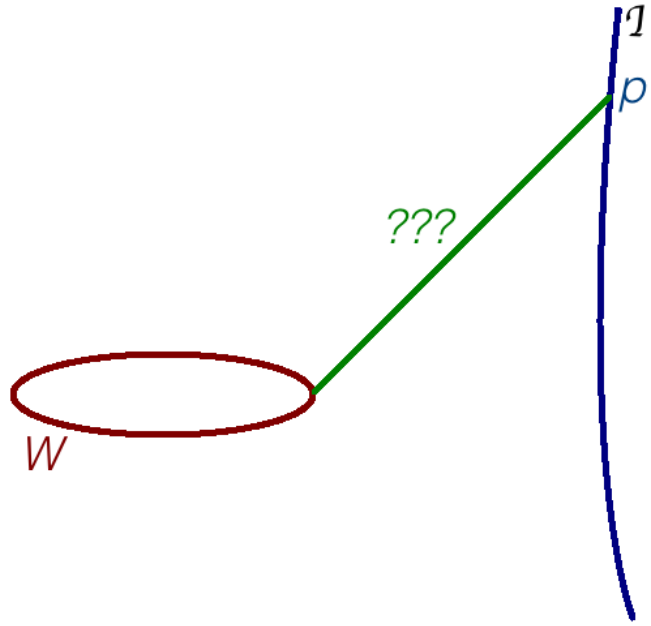
$$B = M \backslash J^-(\mathcal{I}).$$

The *black hole horizon* H is defined to be the boundary of B:

$$H = \partial B.$$

Let us first prove that this definition is sensible by showing that the existence of a black hole region is a generic property of gravitational collapse. We will show that *any trapped surface W is in the black hole region B.* In other words, we will show that a signal from a trapped surface cannot reach the outside observer.

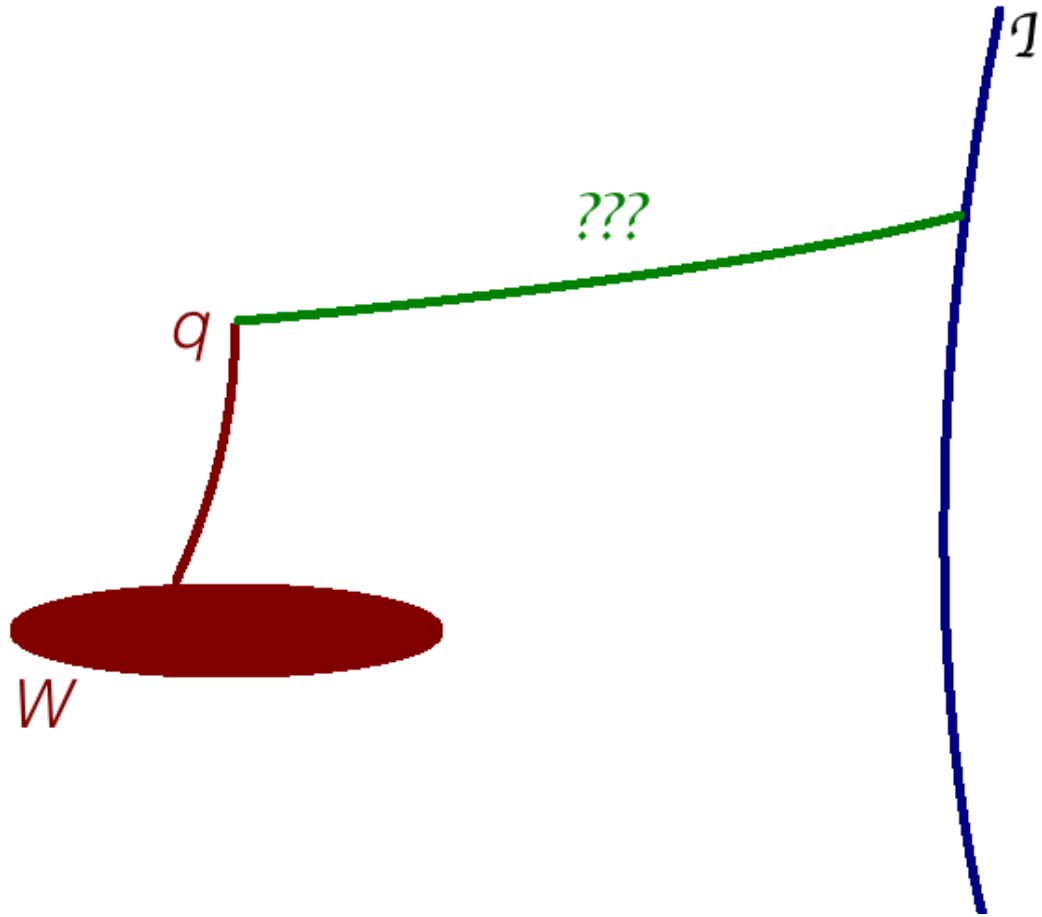
If a causal signal from the trapped surface W can reach the worldline $\mathcal{I}$ of the distant observer, there is a first point p at which this can occur:



The causal path from W to p would be prompt, so it would be a future-going null geodesic ℓ from W to p, orthogonal to W, and without focal points. Since W is a trapped surface, there is a focal point on ℓ within a known, bounded affine distance from W. But $\mathcal{I}$, the worldline of the outside observer, can be arbitrarily far away. So this is a contradiction and there is no causal signal from W to $\mathcal{I}$.

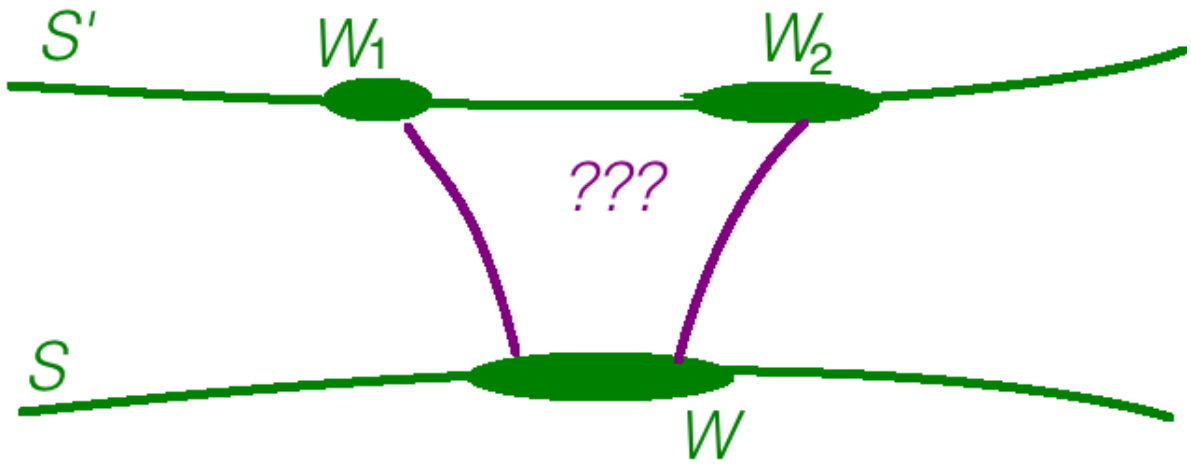
So assuming cosmic censorship, a black hole forms in any asymptotically flat spacetime that contains a trapped surface – and thus in any spacetime that is close enough to the explicit Schwarzschild and Kerr solutions.

An obvious fact is that if a set W is contained in the black hole region B, then its future $J^+(W)$ is also in B. For if q is in the future of W and an event at q can be seen by the distant observer, then that observer can also receive a signal from W:



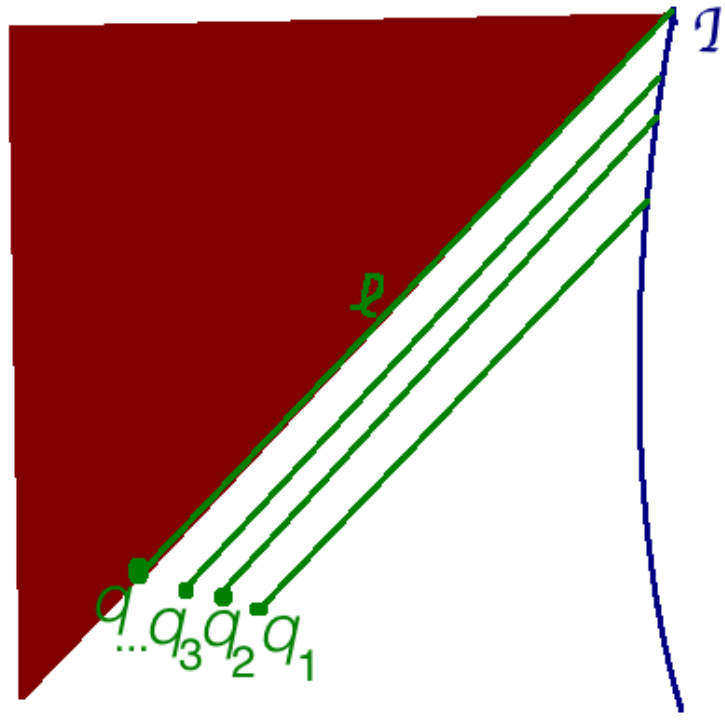
There might be any number of black holes in spacetime, so on a given initial value surface S, the black hole region might have several connected components W_i:

Black holes can merge, but a black hole cannot split, in the sense that if W is a connected component of the black hole region on a given initial value surface S, then the future of W must intersect any initial value surface S' that is to the future of S in a connected component. In other words the following is impossible:

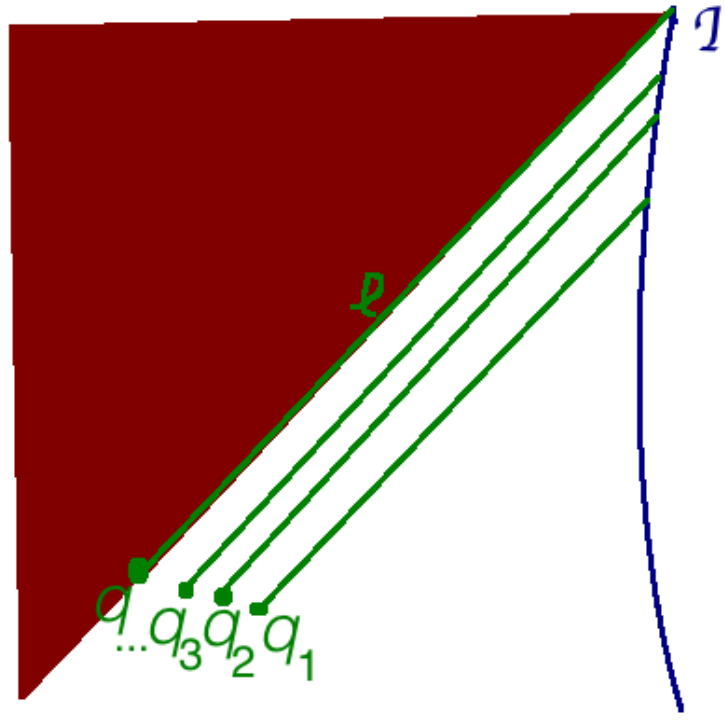


(If there is any causal path from W to W_1 or W_2, then by maximizing the proper time, we learn that there is a causal (timelike or null) geodesic with that property. The space of causal geodesics starting at W is connected, and cannot be continuously divided into two disjoint sets, consisting of geodesics that connect W to W_1 or to W_2.)

Now we want to discuss the "horizon generators." Let q be a point on the black hole horizon, and let $\mathcal{I}$ be the timelike worldline of a more or less stationary observer at infinity.



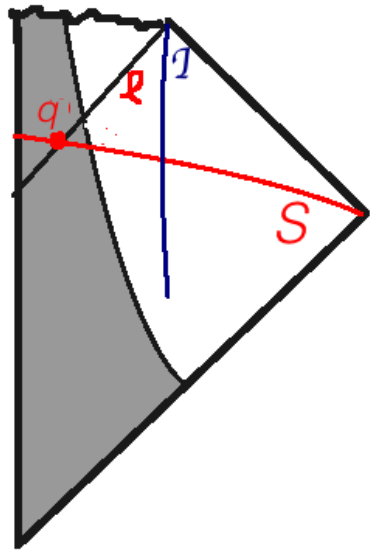
A point q on the horizon is the limit of a sequence of points $q_1, q_2, q_3, \cdots$ outside the horizon. Each of the q_i is connected to the worldline $\mathcal{I}$ by a future-going prompt null geodesic ℓ_i. As $q_i \to q$, the ℓ_i approach a future-going null geodesic ℓ from q.



ℓ is everywhere on the horizon H, which we recall is the boundary of the black hole region B: (1) ℓ can never go outside B, since a causal curve starting at $q \in H = \partial B$ can never reach outside B; (2) and ℓ cannot be in the interior of B, because it is the limit of the prompt null geodesics from q_i that are strictly outside B.

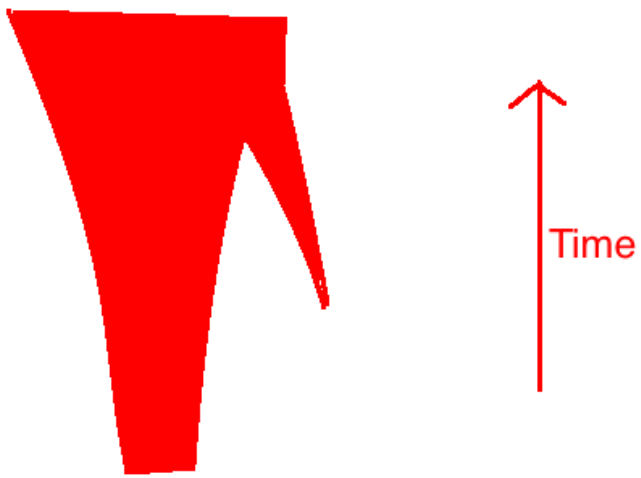
So ℓ is everywhere in $H = \partial B$.

Pick an initial value surface S that contains q and let $W = S \cap H$.

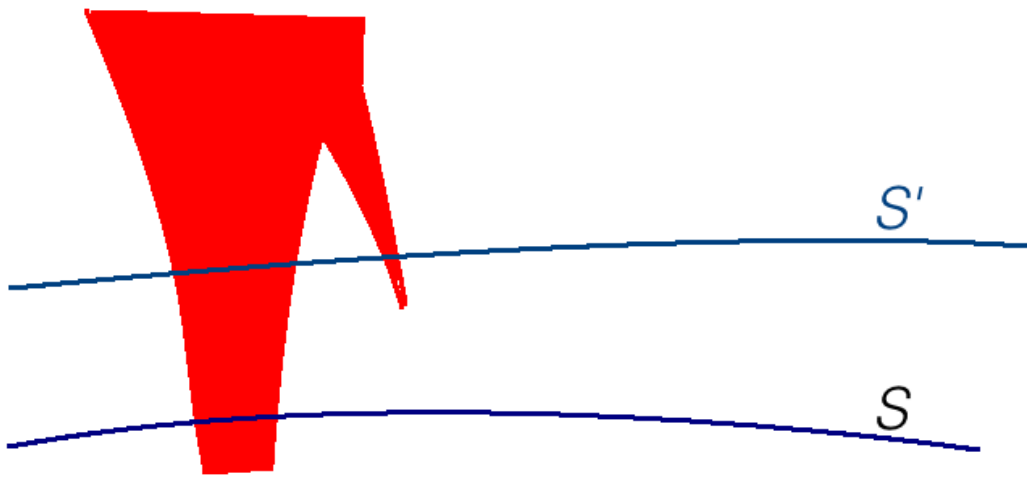


ℓ must be orthogonal to W, or else by leaving W from a nearby point to q, one could find a causal path that would go to the past of ℓ, that is, it would go outside B, contradicting the fact that by definition $W \subset H \subset B$. For similar reasons ℓ has no focal points. So ℓ is a prompt causal path from W.

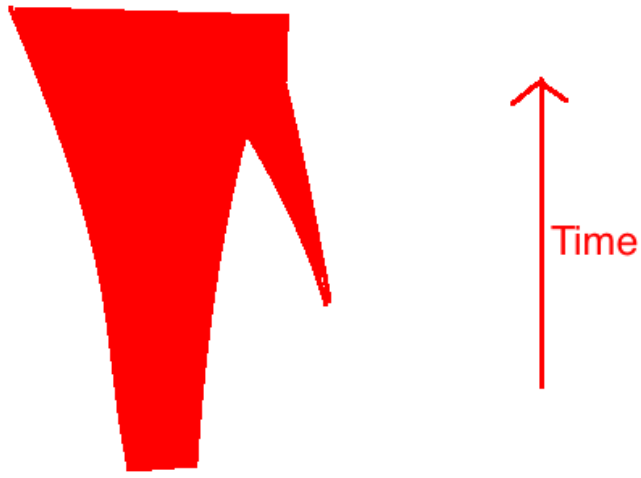
The orthogonal null geodesics from W that stay in the horizon are called horizon generators. Every point in $W = H \cap S$ is contained in one of these horizon generators, and together they sweep out a three-manifold H'. From what we've seen, H' is contained in the horizon H, and near W, the two are the same. But if we continue into the future, H' may 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:



The final result that I will explain about classical horizons is probably the most important: the Hawking area theorem. It says that the area of a black hole horizon can only increase, meaning that the area measured on an initial value surface S' to the future of S is at least as big as the area measured on S:

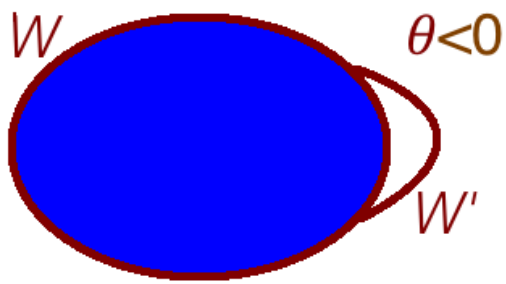


It suffices to show that the null expansion $\theta = \dot{A}/A$ of the horizon generators that connect W to i^+ is everywhere nonnegative. This being true for any choice of S (or $W = S \cap H$) means that the horizon area is everywhere locally nondecreasing:

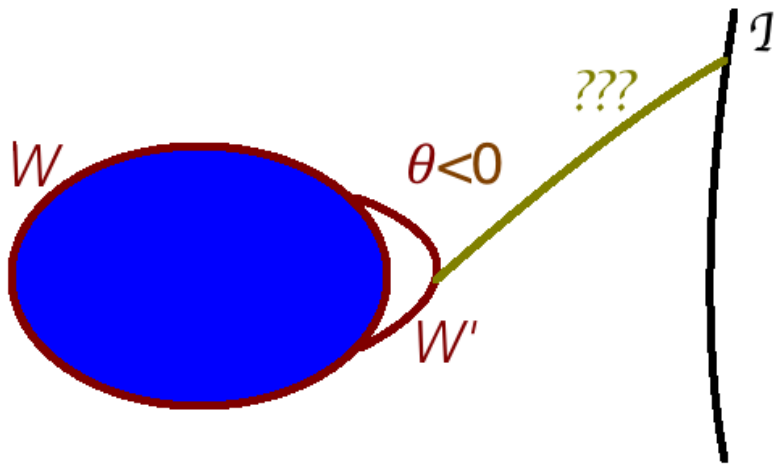


(To fully determine the growth of the horizon area, one has to take into account that new horizon generators can come into existence in the future of S. But that can only give a further increase in the horizon area.)

To show that the null expansion of the horizon generators is everywhere ≥ 0, imagine that at some point of $W = H \cap S$, one has $\theta < 0$. Then we can push W out to a surface W' that is partly outside of B, such that $\theta < 0$ in the portion outside of B:



Since it is not entirely contained in B, W' is connected to the worldline $\mathcal{I}$ of an observer at infinity by a causal path ℓ, which we can choose to be prompt:



ℓ must connect $\mathcal{I}$ to a point in W' that is outside B, but we chose 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 $\theta < 0$.