Advanced General Relativity: Geometry, Topology and Causality of Space-Time

---

Sunil Mukhi

Preliminary version - author not responsible for typos, errors or misconceptions!

Contents

Contents

1 Introduction

The General Theory of Relativity [1, 2] is one of the greatest achievements in the history of Physics. It is at the same time utterly simple and incredibly complex. The simplicity arises from the fact that the gravitational field is identified with a geometrical quantity, the *metric* of space-time. The equations that this field must satisfy, in classical Physics, are the Einstein equations – which express the curvature of space-time in terms of the matter present. They are simple to write down:

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R + \Lambda g_{\mu\nu} = 8\pi G\, T_{\mu\nu} \tag{1.1}$$

and can be derived from a variational principle, as the stationary conditions for the action:

$$S = \frac{1}{16\pi G} \int d^D x \sqrt{g}(R - 2\Lambda) + S_{\text{matter}}[g_{\mu\nu}, \phi^I] \tag{1.2}$$

with appropriate boundary conditions and/or boundary terms. All the relevant notation and conventions can be found in Appendix A.

Despite their conceptual simplicity, unpacking these equations for their mathematical and physical content is extremely complicated. There are two broad reasons for this, one mathematical and the other physical. On the mathematical side, Eq.(1.1) is a set of second-order partial differential equations for 10 independent variables $g_{\mu\nu}$. While this makes things tedious, one might have thought it would be within reach of modern computing methods. Yet, it is not possible to find the most general solutions to the Einstein equations. To solve differential equations involving space, we need boundary conditions. To solve differential equations in time, we need initial conditions. Now

these are equations in both space and time, so we need boundary as well as initial conditions. But space-time mixes space and time, so it is tricky to separate the two conditions.

And there's another problem – general coordinate invariance of the equations is a *gauge invariance* or *redundancy*, so 4 out of 10 variables $g_{\mu\nu}$ are redundant and can be removed by a *gauge choice*. Another 4 correspond to *constraints* on the initial values and therefore are also effectively removed. That leaves two independent degrees of freedom of the metric, corresponding to the two independent polarisations of a gravitational wave. However, there are many (ultimately equivalent, but technically distinct) ways to go from the metric to the independent degrees of freedom.

Those were the mathematical problems. On the physical side, one can find space-times that do a lot of strange things. For example, time can be a circle – then as we go forward in time, we will keep cyclically returning to the past. Such space-times are considered *unphysical*. But who, or what, decides what is physical and what isn't? In 1926, Arthur Eddington claimed that black holes were unphysical solutions of the Einstein equations. Later on, S. Chandrasekhar also doubted they could form in nature. Yet today we know they are among the most common objects in the universe! So there is now a consensus that black holes are physical objects, but the status of other types of space-times, such as those with observable singularities, is still unclear. Current research deals with energy conditions, causality conditions, singularity avoidance and other physical requirements, not all of which fall within the scope of these notes.

Before proceeding, let us mention that there has to be more to gravitation than the Einstein equations. The universe obeys quantum mechanics, so logically the laws of gravity should also be quantum laws. This means we should think of the metric $g_{\mu\nu}$ as a set of operators, find their canonical momenta and quantise them as we do any other quantum field. Or we should write a path integral over the space of metrics and compute it. Today it is not known how to do either of these in any complete sense, and the attempts made so far have encountered formidable obstacles – essentially due to the non-linearity of the system as well as general coordinate invariance. There have been partial successes, but the general problem of quantising gravity remains under active investigation.

These notes are intended to help students and other researchers wishing to participate in this investigation by providing a deeper understanding of classical gravity than the standard courses. Here, gravity will be treated as classical – and things will

be complicated enough despite this limitation. Our focus will be causality, geometry and topology for classical space-times, with particular emphasis on singularities. Our exposition will follow and review the pioneering works of A.K. Raychaudhuri, Roger Penrose and Stephen Hawking [3–10]. All this is still too much material for a short course, so even within this apparently narrow remit, many topics will unfortunately have to be excluded.

There are two pre-requisites for these notes: (i) a basic understanding of General Relativity, typically corresponding to a first graduate course in the subject at the level of [11] or [12], (ii) an understanding of basic topology and differential geometry, at the level of the first half of [13] or the more informal level, tailored to physicists, of the exposition in [14].

These notes closely follow the ideas expounded/reviewed in References [3, 15–18]. Even when discussing the same results, the first three references often take different approaches. I have generally picked the one that I found most understandable at each point. The order of presentation of results also varies, and again I have chosen one that made sense to me. In particular, following [17] I described the work of Hawking, which involves time-like singularities, before that of Penrose which involves null singularities, despite the historical order being somewhat opposite. The reason is that the time-like case is considerably simpler to describe than the null one.

The above sources are far too long and dense for a short course of 12 lectures, for which the present notes are designed, so I have condensed some of the expositions considerably. Moreover the books [16] and [18], and to a lesser extent the review article [17], assume considerable familiarity with topology and differential geometry. Some familiarity is assumed here too, but an Appendix is also provided to help the reader with some of the details.

2 Causality

2.1 First look at causality

In this section we will study causal properties of generic space-times, as well as space-times satisfying certain properties that we will specify. At this stage we will not assume they are solutions of the Einstein equations. We start with the basic definition of space-time itself.

Definition 2.1. *A **space-time** M is a D-dimensional differentiable manifold with a smooth metric that, in the coordinate system x^μ, is denoted $g_{\mu\nu}(x)$. The metric is taken to be of Lorentzian signature, or "pseudo-Riemannian". M is also taken to be **time-orientable**, namely there is a notion of "future" and "past" with respect to the time direction, that varies smoothly over M and is preserved under coordinate transformations* [1].

The key feature of space-times, as compared to metric spaces of Euclidean signature, is the presence of a single negative eigenvalue of the metric. In terms of coordinates, there is one coordinate distinct from all the others that we call *time*. When physical laws are formulated on space-times, we understand time to be flowing in the forward direction. It is conventional to depict this in a *space-time diagram* where time is on the vertical axis while the horizontal axis represents all $D-1$ spatial directions. This is straightforward to draw for $D=2$, and not too hard for $D=3$ where we imagine the horizontal direction to be a plane, as sketched in Fig.1. Beyond that, explicit visualisation becomes more difficult. Another common approach is to plot the radial direction against time and "imagine" a $D-2$-dimensional sphere at each radial point.

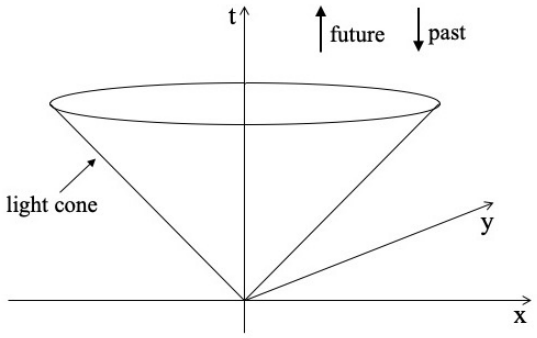


Figure 1: Light cone in Minkowski. space-time

In a space-time diagram, the meaning of "future" and "past" directions in a small neighbourhood of any given point is determined by the positive and negative parts of the time axis relative to that point. The light-cone at a point gives us the instantaneous trajectory of a light-ray arriving at, or emanating from, that point. In Minkowski space-time, in $D=2$ this is just a pair of straight lines at 45°, but in $D=3$ it is a 2-dimensional cone. In a general (non-Minkowski) space-time M we cannot think of

[1] Technically M is also required to be *Hausdorff* and *paracompact*. Details can be found in [16], Appendix A.

the light cone as living in M, rather at each point $P \in M$ it lives in the tangent space $T_P(M)$.

Now we can discuss causality. Consider space-time paths (also called world-lines) $P(s)$ that assign a point $P \in M$ to each value of the parameter s, an arbitrary parameter that labels points on the path. In a coordinate system x^μ, a path would be denoted $x^\mu(s)$ [2]. Two paths are treated as equivalent if they differ by a non-singular reparametrisation of s: $s \to \tilde{s}(s)$, and also if they differ by a non-singular reparametrisation of the space-time coordinate, $x^\mu \to x'^\mu(x)$. At this stage we do not assume geodesic paths, so the paths we have been discussing are not necessarily the world-lines of a freely moving particle.

Definition 2.2. *A path*[3] $x^\mu(s)$ *is* ***time-like*** *if the tangent vector* $\frac{dx^\mu}{ds}$ *is everywhere time-like, and* ***causal*** *if the tangent vector is everywhere time-like or null:*

$$\begin{aligned} \textbf{\textit{Time-like path}}: &\quad g_{\mu\nu}\big(x(s)\big)\frac{dx^\mu}{ds}\frac{dx^\nu}{ds} < 0, \quad \textit{all } s \\ \textbf{\textit{Causal path}}: &\quad g_{\mu\nu}\big(x(s)\big)\frac{dx^\mu}{ds}\frac{dx^\nu}{ds} \leq 0, \quad \textit{all } s \end{aligned} \tag{2.1}$$

The first condition says the tangent vector lies "inside" the light cone while the second allows it also to be partly or entirely on the light cone. Note that a path which becomes space-like even for a very short segment fails both of the above conditions.

We have used a coordinate system to write the definition, but a curve is a geometric entity that can be described in any coordinate system. For this reason we initially denoted it $P(s)$ and this is followed by some authors, e.g. [16]. This permits abstract discussions of paths without assigning a cooordinate system. Also, sometimes one adds "future directed" or "past directed" to the definition above. This means the tangent vector lies in the future half or past half of the light cone.

Next we consider an arbitrary point P in M and discuss the future and past of that point. We first define the future and past as they would be experienced by a massive (possibly accelerated) particle.

Definition 2.3. *The* ***chronological future*** $I^+(P)$ *or* ***chronological past*** $I^-(P)$ [4] *is the set of the points reachable in the future/past by a time-like path starting at* P.

[2] We typically consider continuous paths, most often they are also smooth but not always.

[3] We use the words "path" and "curve" interchangeably, though "path" is more common when there are end-points, as in "causal path from P to Q".

[4] Recall that "Chronos" was the Greek god of time.

We can define the chronological future of a set $S \subset M$ as the union of chronological futures of each point in S: $I^+(S) \equiv \cup_{P\in S} I^+(P)$ and similarly for $I^-(S)$.

This is illustrated in Fig. 2. The chronological future is an open set in the given topology, because a point in the region can be completely enclosed in an open set (a time-like curve remains time-like under an infinitesimal deformation). A related fact is that P is in its own chronological future/past only if the space-time has closed time-like curves, and not otherwise.

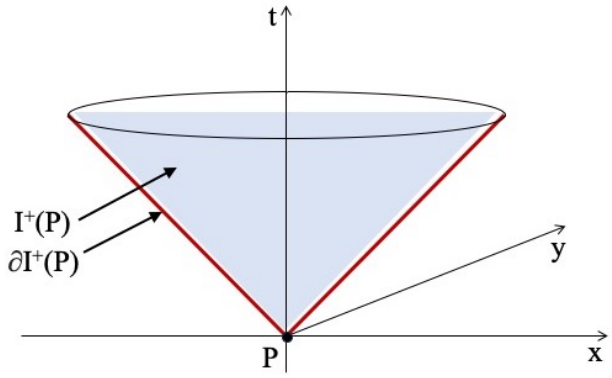


Figure 2: Chronological future of P in Minkowski space-time.

Exercise 2.1. *Prove both the statements above.*

Since massless particles travel (in vacuum) at the speed of light, the above definition would exclude points that can be accessed in the future by a light signal, or points in our past that can access us through a light signal (every instant on a star is in the latter category). So we generalise the concept of "chronological" to "causal". The mathematical properties of the resulting spaces will turn out somewhat different.

Definition 2.4. *The* ***causal future*** $J^+(P)$ *or* ***causal past*** *of a point P is the set of points reachable in the future/past by a causal path starting at P. Again we define the causal future of a set $S \subset M$ as the union of causal futures of each point in S: $J^+(S) \equiv \cup_{P\in S} J^+(P)$ and similarly for $J^-(S)$.*

Unlike $I^\pm$, the spaces $J^\pm$ can include part or all of their own boundary – the set of points that can be reached by/from an everywhere-null causal path. By this definition, the *constant* path from P to itself is causal, so P lies in its own causal future (and past).

Exercise 2.2. *Verify that any point P lies in its own causal future and past.*

The boundary of $I^+(P)$ is denoted $\partial I^+(P)$ [5]. One may guess that the union $I^+(P) \cup \partial I^+(P)$ is the same as $J^+(P)$, but this is not true in general. An example is Minkowski space-time minus a point Q on the future light-cone of P. Then $J^+(P)$ lacks the deleted point and all points to its future along the light cone, while $I^+(P) \cup \partial I^+(P)$ only lacks the deleted point but includes all points to its future, see Fig. 3.

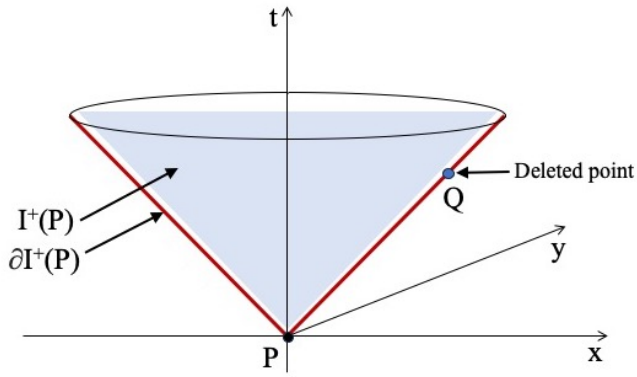


Figure 3: The causal future $J^+(P)$ does not contain the future of the deleted point Q along the light-cone, but $I^+(P) \cup \partial I^+(P)$ does contain it.

Exercise 2.3. *To get some practice, consider Minkowski space-time with a point in $I^+(P)$ deleted. Now examine carefully the sets $I^+(P), \overline{I^+(P)}, J^+(P), \partial I^+(P), \partial J^+(P)$ (here a bar denotes closure of a set).*

Exercise 2.4. *Show that (i) $I^+(S)$ is an open set, (ii) $I^+(I^+(S)) = I^+(S)$, (iii) $I^+(\bar{S}) = I^+(S)$.*

The example of Minkowski space-time with a point deleted also shows us that $J^+(P)$ need not be closed. In this case, the closure of $J^+(P)$ will include the deleted point but $J^+(P)$ itself does not.

Next we move on to consider two distinct points P and Q with Q in the future of P. We will be interested in paths that start at P and end at Q.

Definition 2.5. *The **causal diamond** of P and Q is $D_P^Q = J^+(P) \cap J^-(Q)$, namely the intersection of the causal future of P with the causal past of Q.*

[5] Some references denote it by $\dot{I}^+(P)$.

Fig.4 illustrates a causal diamond, while Fig.5 shows a situation where the causal diamond is empty (P and Q are space-like separated). A causal path from P to Q must lie in the causal diamond of these points (but clearly not every path that lies in the causal diamond is a causal path).

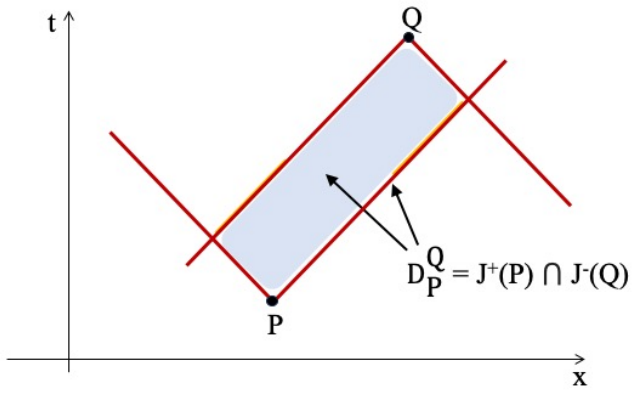


Figure 4: Illustration of causal diamond.

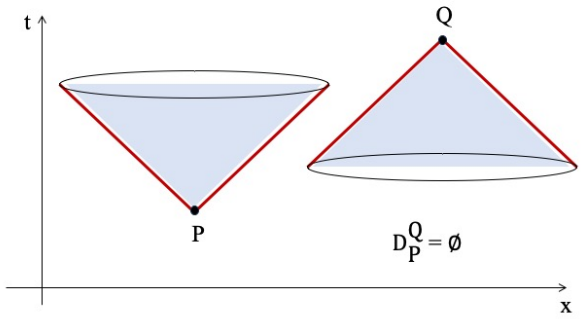


Figure 5: Situation with empty causal diamond.

Exercise 2.5. *Draw paths from P to Q that are (i) in the causal diamond, and causal, (ii) in the causal diamond, but not causal, (iii) not in the causal diamond.*

So far we did not invoke geodesics. In Minkowski space-time it is evident that the causal future and causal past of P are the regions connected to P by time-like or null geodesics. These statements are not true in general, as one can see by removing a

point from Minkowski space-time, as a result of which some points may no longer be reachable from P by a geodesic. However they are true *locally* in any space-time M. This is embodied in the following definition and theorem.

Theorem 2.1. *In a space-time M there always exists an open neighbourhood of P, call it U, such that any $Q, R \in U$ are connected by a unique geodesic lying within U. Such a U is called a* ***convex normal neighbourhood (CNN)*** *of P (see Fig.6). If we restrict the causal future of P to the part inside U, then it is true that all of it can be reached by time-like or null geodesics from P and similarly for the causal past* [6].

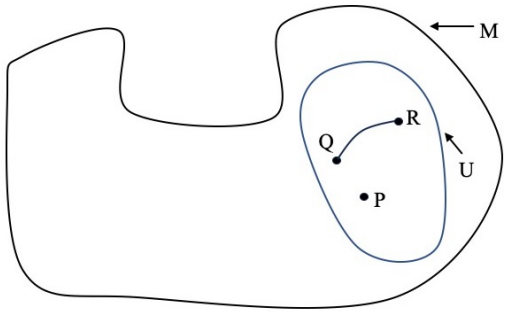


Figure 6: Illustration of a convex normal neighbourhood U.

2.2 Globally hyperbolic space-times

We would like to specify the conditions that a space-time must satisfy for it to be a physically acceptable one. To this end, we start with a few definitions that will help us define physical initial-value hypersurfaces.

First we define subsets of M in which no two points can be connected by a time-like curve.

Definition 2.6. *A subset $S \subset M$ is said to be* ***achronal*** *if the intersection $I^+(P) \cap I^-(Q)$ is empty for all pairs of points $P, Q \in S$.*

[6] This is theorem 8.1.2 of [16], but – as pointed out there – the proof is to be found elsewhere. The first part, the existence of a CNN, is a result in regular Riemannian geometry and its proof can be found in standard texts on the subject such as Section 9.4 of [19]. It is also discussed in Section 2.5 of [18]. The second part is of course specific to Lorentzian-signature geometry, and its proof is provided in [18], proposition 4.5.1.

Recall that $I^{\pm}(P)$ is the *chronological* future/past of P. So in an achronal set S there are no pairs of time-like-separated points. That means all pairs of points in S are either spacelike-separated or null-separated.

Note, however, that surface that is locally space-like or null at each point may still fail to be achronal. As an example, consider a hypersurface corresponding to a spatial circle periodically swept out in time. In a space-time diagram this looks like a helix and can easily be taken to be locally space-like everywhere. We easily see (Fig.7) that there are time-like paths that take us from a point on the helix at one time to a nearby point at a future time. Hence the helix, though space-like, is not achronal.

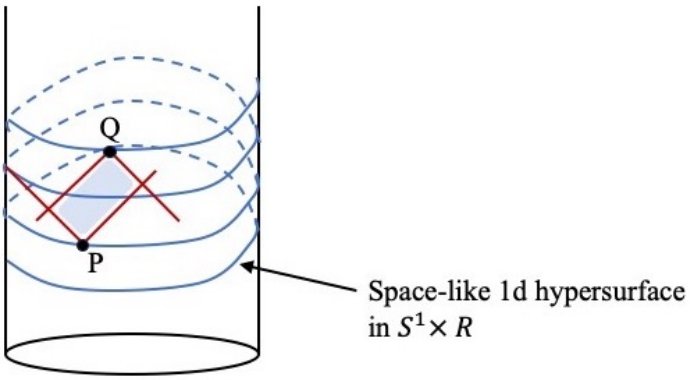


Figure 7: The helix is space-like but not achronal.

Theorem 2.2. *The boundary $\partial I^+(S)$ of any set $S \subset M$ is a codimension-1 achronal submanifold of M (for a more formal statement and a proof see [16], Theorem 8.1.3).*

The proof uses the fact that $I^{\pm}$ of any point or set is open. First assume the contrary of the assertion, namely that $\partial I^+(S)$ is not achronal. Then we can take two points $P, Q \in \partial I^+(S)$ such that $Q \in I^+(P)$. Next, we prove that $I^+(P) \subset I^+(S)$. It follows that $Q \in I^+(S)$ which, being open, is disjoint from $\partial I^+(S)$ and we have a contradiction.

The next concept we need to understand is whether a curve in space-time can extend forever or must come to an end. Accordingly we define the concept of an inextendible curve, one that cannot be extended further.

Definition 2.7. *Consider a future-directed causal curve $x^\mu(s)$. A point $P \in M$ is called a* ***future end-point*** *of this curve if every open neighbourhood $O \subset M$ containing P*

completely contains the curve beyond some value s_0 of the parameter s, i.e. if $x^\mu(s) \in O$ for all $s > s_0$. A curve that has such an end-point is called ***future extendible****, while a curve without any future end-point is called* ***future inextendible****.*

Note that the end-point P does not need to actually lie on the curve! In other words, there does not have to be a value $s = \bar{s}$ such that $x^\mu(\bar{s})$ is the point P. The above definition requires the curve to come arbitrarily close to P and to stay close forever.

Exercise 2.6. *Construct a curve in Minkowski space-time that approaches a point P but never passes through it. Such a curve is extendible by our definition. Construct a new curve that extends it to infinity.*

We can distinguish various typical cases. A curve that goes on to infinity is, by definition, inextendible. However a curve that terminates arbitrarily close to a point that is missing from the manifold, is also inextendible. The reasoning is that if we remove a point P from M then a causal curve terminating at P, that was previously extendible, will become inextendible. After all P can no longer be an end-point of the curve (it's not in M in the first place). On the other hand a curve that either terminates at $P \in M$, or for some $s > s_0$ comes arbitrarily close to a point $P \in M$, is extendible.

Alternatively suppose we have an inextendible causal curve passing through a point P in a space-time M. Now if we remove the point P from M, our curve breaks up into two inextendible causal curves, one in the past of P and the other in the future. Conversely if we add back P, these two curves become extendible in the new space-time and of course they can each be extended to be the original curve.

Now there is an important lemma. It tells us that if we have an arbitrary past inextendible causal curve, then we can always find a past inextendible time-like curve in its chronological future.

Theorem 2.3. *Let $x^\mu(s)$ be a past inextendible causal path through P. Then through any point $Q \in I^+(P)$, there exists a past inextendible time-like path $\tilde{x}^\mu(\tilde{s}))$ that lies entirely in $I^+(x^\mu(s))$ (the proof, as well as a method to construct $\tilde{x}^\mu(\tilde{s})$ given $x^\mu(s)$, can be found in [16] Lemma 8.1.4).*

This theorem comes from the existence of convex normal neighbourhoods, discussed above. The idea is to first pick a Euclidean signature metric on our manifold M. Then, construct a curve $x'^\mu(s)$ in the chronological future of $x^\mu(s)$ such that the largest

separation between points on the two curves is decreasing as s goes towards negative time. By construction, $x'^{\mu}(s)$ is time-like. Because it comes closer and closer to $x^{\mu}(s)$ as we go to the past, an end-point of $x'^{\mu}(s)$ would also be an end-point of $x^{\mu}(s)$ but the latter does not exist since we assumed $x^{\mu}(s)$ to be past inextendible.

Having discussed inextendible causal curves, we turn to the important concept of domains of dependence. This will set the stage for us to specify what kind of initial-value surface we wish to use to solve Einstein's equations.

Definition 2.8. *The* ***future domain of dependence*** $D^+(S)$ *of an achronal space-like set* S *is the set of all points* $P \in M$ *such that all past inextendible causal paths through* P *pass through* S*. The* ***past domain of dependence*** $D^-(S)$ *is defined similarly, and the* ***domain of dependence*** $D(S)$ *is just* $D^+(S) \cup D^-(S)$*.*

Several illuminating examples that illustrate the notion of domains of dependence can be found in [15]. The first of these examples shows that, for there to be a non-trivial domain of dependence at all, the achronal set S must be a hypersurface, i.e. a co-dimension-1 subspace of M – otherwise $D(S)$ is just S itself. Accordingly, below we restrict our attention to hypersurfaces.

Exercise 2.7. *If* S *is an achronal space-like set and* H *is an achronal subset of* $D^+(S)$*, show that* $D^{\pm}(H) \subset D^{\pm}(S)$*. An example is discussed in [15].*

The future and past domains of dependence $D^{\pm}(S)$ need not be closed. We may consider the closure of $D^+(S)$, denoted $\overline{D^+(S)}$. Then we have a theorem for points $P \in \overline{D^+(S)}$:

Theorem 2.4. *A point* P *lies in* $\overline{D^+(S)}$ *if and only if every past inextendible time-like curve time-like curve through* P *intersects* S*. (The proof is straightforward – see [16], Proposition 8.3.2.)*

Finally we define the central object in terms of which an initial-value problem can be set up. This is the Cauchy hypersurface.

Definition 2.9. *A* ***Cauchy hypersurface*** *in* M *is an achronal space-like* [7] *surface* Σ *such that every inextendible causal path through a point* $P \in M, P \notin \Sigma$*, passes through* Σ*. A space-time* M *with a Cauchy hypersurface is said to be* ***globally hyperbolic*** [8]*.*

[7] Ref. [17] requires it to be space-like, while [16] and others do not. For now, we will stick with the space-like definition and generalise to the null case as needed later on.

[8] As we will see in sub-section 2.4, there are essentially three different definitions of global hyperbolicity in the literature, all of which are equivalent. The one we have used here is the easiest to understand physically, and – probably for this reason – the most widely used.

This definition basically says that every causal path in M is connected to a Cauchy hypersurface, in the past or future (see Fig.8). This sets the stage for us to specify initial conditions on the Cauchy hypersurface and be confident that they will determine the system for all future times.

Ref [16] requires Σ to be closed as part of the definition. However in [17] it is shown that it must be closed, thus closure does not need to be part of the definition. For this we will use the property that topologically, a space-time with a Cauchy hypersurface is $\Sigma \times R$ (this will be proved shortly). Now Σ is the same as $\Sigma \times \{0\}$ which is a closed subset of the full space-time $\Sigma \times R$. We will also give a more formal proof of the closure of Σ shortly.

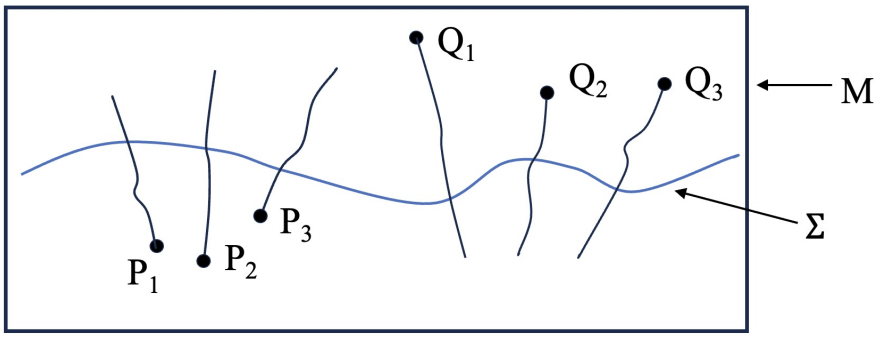


Figure 8: A Cauchy hypersurface: all causal paths intersect it in the future or past.

A Cauchy hypersurface Σ divides the whole space-time into a future and a past. To quote [16], "in a globally hyperbolic space-time, the entire future and past history of the universe can be predicted (or retrodicted) from conditions at the instant of time represented by Σ". Clearly a Cauchy hypersurface Σ is a surface for which $D(\Sigma) = M$, i.e. its domain of dependence is the whole space-time. If S is achronal and space-like but not a Cauchy hypersurface, then its domain of dependence $D(S)$ is not the whole space-time M. But $D(S)$ on its own is a globally hyperbolic space-time, essentially by construction, and as expected, S divides $D(S)$ into a future and a past.

Note that an inextendible time-like path through P must actually intersect the Cauchy surface exactly once. For, if it intersected multiple times then the segment between two intersections would be a time-like path from the Cauchy surface to itself (yellow path in Fig. 9), contradicting achronality.

The standard example of a non-globally hyperbolic space-time is Minkowski space-

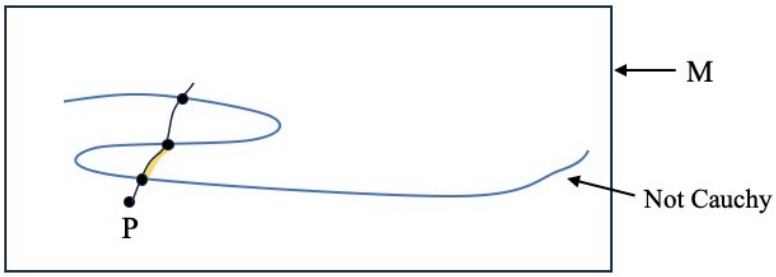


Figure 9: If a time-like path intersects a surface multiple times, that surface is not achronal (due to the yellow segment) and hence not Cauchy.

time with one point P deleted. Take a space-like surface S for which P lies in its causal future. Deletion of P does not change the fact that other points in its future are in the causal future of S, since we can always get to such points by going around P. But now consider a point Q in the causal future of P and a past-directed time-like path from Q that lands on P. This path is inextendible yet it does not intersect S. This shows that any S in the past of P cannot be Cauchy. The same argument applies to P in the past of S by reversing all time directions. Hence $M - \{P\}$ is not globally hyperbolic.

There are situations where there is an achronal space-like set that fails to be a Cauchy hypersurface. Let us consider an example of such a hypersurface [15]. Let S be the past hyperboloid in Minkowski space-time, defined by $\vec{x}^2 + R^2 = t^2$ (red line in Fig.10). It is everywhere space-like since:

$$\left|\frac{d\vec{x}}{dt}\right| = \frac{t}{\sqrt{t^2 - R^2}} > 1 \tag{2.2}$$

By drawing it we can see that it is also achronal – there are no time-like curves connecting two points on it. However we can draw a time-like curve that asymptotes to the time axis in the future and to the light cone in the past (blue line in Fig.10). This curve fails to intersect the set, which therefore cannot be Cauchy.

Thus such sets do exist, and we cannot use them as initial-value hypersurfaces for the whole of M. Nevertheless they will be valid initial-value hypersurfaces for some *subset* of M, namely their own domain of dependence.

Next we define a surface that describes, in some sense, the failure of S to be a Cauchy hypersurface.

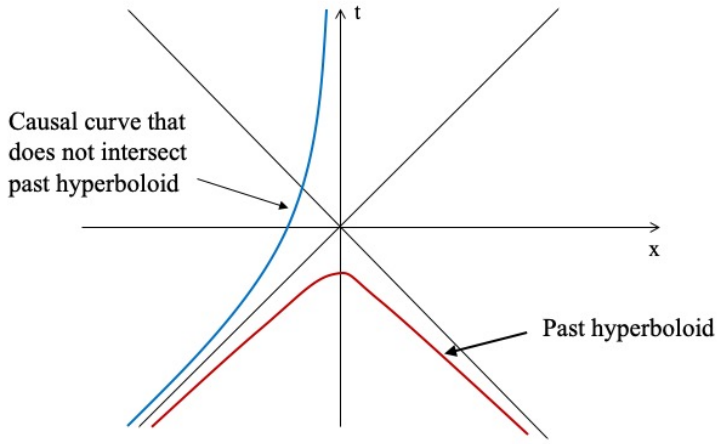


Figure 10: The past hyperboloid (red) is not Cauchy because the time-like curve (blue) does not intersect it.

Definition 2.10. *The boundary of the closure of the domain of dependence of an achronal space-like hypersurface S, namely $\partial\overline{D(S)}$, is called a* ***Cauchy horizon*** *$H(S)$. This in turn splits into $H^+(S)$ and $H^-(S)$, the future and past Cauchy horizons, which are defined similarly in terms of $\partial\overline{D^\pm(S)}$.*

A Cauchy horizon gives us a boundary between points for which S acts as a Cauchy hypersurface and points for which it does not. In the example of Fig.10 above, it is easy to see that the Cauchy horizon of the past hyperboloid is the past light-cone.

Before moving on let us state a possible property of space-time that is weaker than the globally hyperbolic property. This will come in useful if for any reason we need to relax global hyperbolicity. We definitely want to prevent closed time-like curves, so we will take that as a minimal requirement. However we also want to avoid space-times where there are *nearly closed* time-like curves, namely causal curves which are arbitrarily close to being closed. If such curves are allowed then a tiny perturbation of the metric $g_{\mu\nu}$ could make them closed. These observations are illustrated in Fig.11.

Definition 2.11. *A space-time is said to be* ***strongly causal*** *if, for every point P and an arbitrary open neighbourhood O of P, there exists a sufficiently small sub-neighbourhood $V \subset P$ through which no causal curve passes more than once.*

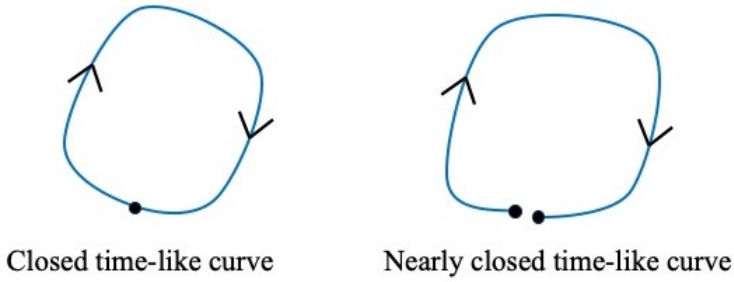


Figure 11: The nearly closed time-like curve at right could become closed under a metric perturbation. Note that the above curves do not "look" time-like in our diagram, but we are assuming that they are time-like in some non-trivial geometry.

2.3 Properties of globally hyperbolic space-times

Globally hyperbolic space-times have several good properties. First, they cannot contain closed time-like paths. This follows from the fact that an inextendible time-like path which is closed will necessarily intersect the Cauchy surface multiple times (in fact infinitely many times) but we saw above that this contradicts achronality of the surface.

Another important property is that two different Cauchy hypersurfaces Σ, Σ' in a globally hyperbolic space-time can be continuously deformed into each other, so they are topologically equivalent. The intuitive idea is that every inextendible causal curve from one Cauchy surface must intersect the other Cauchy surface and vice versa. This can be used to set up a smooth 1-1 map between them via time-like curves that are integral curves of a vector field (a more precise proof is in [17], Section 3.2). There is also a statement in case Σ is Cauchy, while S' is only achronal but not Cauchy. In this case it can be shown that S' is topologically equivalent to a portion of Σ. As a corollary, if Σ is noncompact – but connected – then any achronal hypersurface S' *cannot* be compact. These two points are illustrated in Fig.12.

Three more corollaries are:

(i) In a globally hyperbolic space-time, $J^+(P)$ is closed for every point P. This will be proved in the next subsection after we prove a theorem on the compactness of the space of curves between two points.

(ii) the full manifold M is topologically $\Sigma \times R$ (of course this does not mean the metric factorises). This follows from the integral curves discussed above that map one

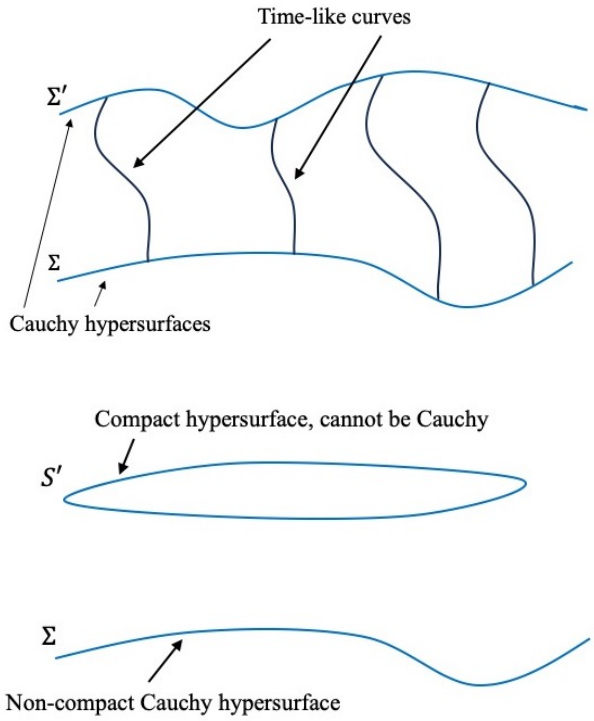


Figure 12: Above: Smooth map between Cauchy surfaces. Below: If Σ is noncompact and connected, then an achronal surface S' cannot be compact.

Cauchy hypersurface to another. This type of "translation" sweeps out a line transverse to the Cauchy hypersurface Σ, hence the result.

(iii) this finally proves that a Cauchy hypersurface must be closed in a globally hyperbolic space-time M, as we mentioned earlier. The result follows from the previous corollary: a Cauchy hypersurface Σ is the same as $\Sigma \times \{0\}$ where 0 is the location of Σ on the transverse curve R. Since Σ is closed in itself, and a single point $\{0\}$ is closed in R, Σ is closed in $\Sigma \times R$. A more detailed proof ([17], page 18) is as follows. Assume the contrary, namely that Σ is not closed. Then there exist points P such that $P \in \overline{\Sigma}$ but $P \notin \Sigma$, i.e. P is a limit point of Σ that does not lie in Σ. Now consider a causal curve passing through such a point. Since $P \notin \Sigma$ and M is globally hyperbolic, the curve must continue until it intersects Σ in a different point P'. Thus there is a causal curve between P and P'. Now although $P \notin \Sigma$, any open set enclosing P overlaps with Σ (as P is a limit point) so by moving infinitesimally away from P to P'' we get a causal curve between P' and P''. This contradicts achronality. The situation is illustrated in Fig.13.

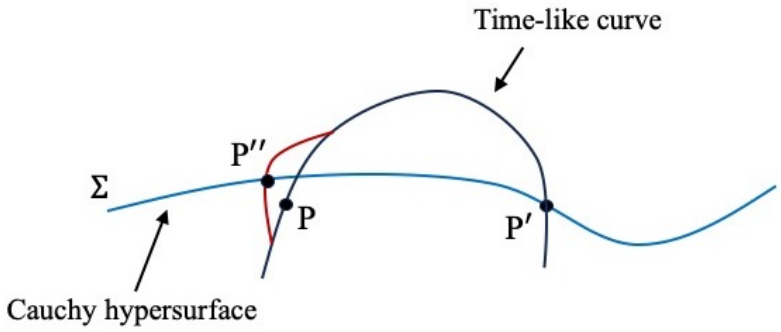


Figure 13: If Σ is not closed, it cannot be achronal.

There is one more property that will be relevant for us. A globally hyperbolic space-time M can often be extended to M' while remaining globally hyperbolic. For example if we start with part of Minkowski space-time lying between t_1 and t_2, this is globally hyperbolic and can also be extended to the full Minkowski space-time while remaining so. Sometimes there will be possible further extensions of M' to space-times M'', M''' etc, but at some point the extended space-time will no longer be globally hyperbolic. Thus we are led to the following concept: a *maximally extended space-time* is one that can no longer be extended while remaining globally hyperbolic. We will often assume this property to avoid artificially reducing the available manifold, for example due to a bad choice of coordinates.

2.4 Compactness of the space of paths

We now consider the topology, not of space-times, but of the *space of causal paths*. Before embarking on our discussion in detail, it will be useful to outline the picture that will emerge. Historically, there have been three different definitions of the defining property of globally hyperbolic space-times and we summarise them as follows:

(D1) The space C_P^Q of causal paths from P to Q is compact for every P, Q (plus strong causality). This is the "Leray definition" [20].

(D2) There exists a Cauchy hypersurface through which all inextendible curves pass. This is [9] the "Geroch definition" [15] and the one we have been using.

[9] Actually Geroch [15] accepts the Leray definition, but then shows it is equivalent to this one.

(D3) The causal diamond D_P^Q is compact for every P, Q (plus strong causality). This is the "Hawking-Ellis definition" [18].

It turns out these are all equivalent, i.e. any one of them implies the other two. In these notes, following [16] and [17], we have used Geroch's definition (D2), so we need to prove its equivalence to (D1) and (D3). We will now sketch how this is done.

Consider a globally hyperbolic space-time M and the set of all continuous [10] causal paths $x^\mu(s)$ starting at P and ending at $Q \in J^+(P)$. We may choose the parameter s such that $x^\mu(0)$ is the point P and $x^\mu(1)$ is the point Q. As always, we identify all $x^\mu(s)$ that differ only by a reparametrisation as representing the same path. Let the set of all such causal paths be denoted C_P^Q. Some causal paths are shown in Fig.14, along with a path (in yellow) that is non-causal and hence not in C_P^Q.

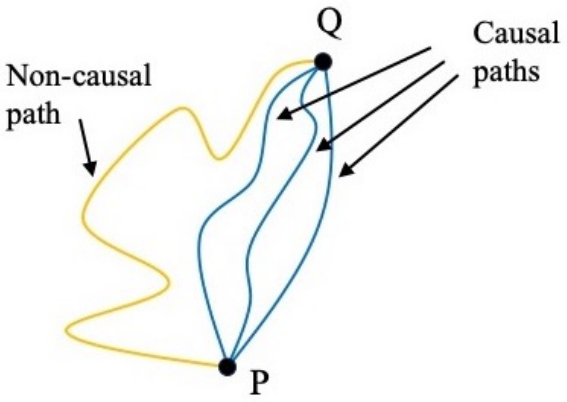


Figure 14: Causal and non-causal paths from P to Q

We define a topology on C_P^Q in terms of the topology on M. Take an open set $U \subset M$ that contains both P and Q. Then $O(U)$ is defined to be the set of all paths in C_P^Q that are fully contained in U (see Fig.15). We take the $O(U)$, with U running over all open sets of M, to be a basis of the open sets on C_P^Q (so arbitrary unions and finite intersections of the $O(U)$ are also open). One can show that this defines a topology on C_P^Q.

Theorem 2.5. *With the topology defined above, the space C_P^Q of causal paths from P to Q in a globally hyperbolic space-time is* ***compact*** *(Theorem 10 of [15], Theorem 8.3.9 of [16]).*

[10] Note that continuity is a significantly weaker condition than smoothness – a continuous path can be jagged, with ill-defined derivatives at some points, as long as it has no gaps.

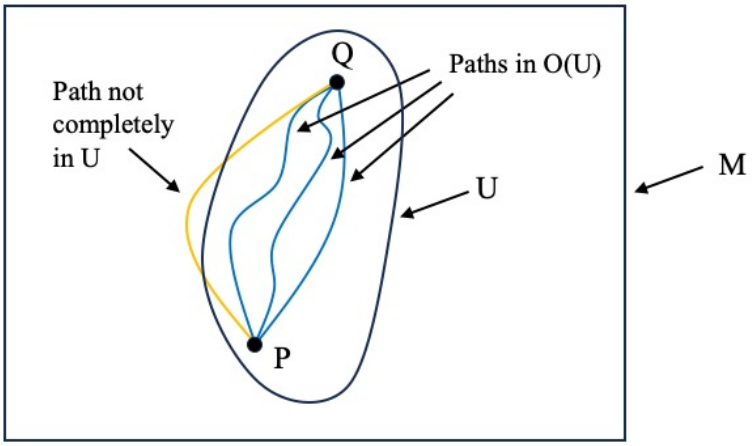


Figure 15: Paths that are in, and not in, $O(U)$.

To quote Geroch [15], the compactness of C_P^Q "requires, intuitively, that there be no "asymptotic regions," "holes," or "singularities" in the region between the points [P and Q]". While this may seem obvious, the importance of proving it is underscored by the importance of the word "causal" in the theorem, without which the corresponding statement would not even be true.

To see this, recall that compactness is closely related to convergence of infinite sequences in a space. And it is easy to show that sequences of general (non-causal) paths need not converge to a definite limiting path. As an example [17], consider a sequence whose nth member is the path $(x = \sin n\pi t, y = 0, z = 0)$ in Minkowski space-time. These paths connect $(0, 0, 0, 0)$ to $(1, 0, 0, 0)$ as t goes from 0 to 1. However there is no limiting path as $n \to \infty$, instead the paths just oscillate more and more wildly as n increases. None of these paths is causal – this is easiest to see if you sketch a few members of the collection. This shows that sequences of *non-causal* paths can have bad convergence properties. The theorem above says that when we consider the space of *causal* paths, the convergence properties are much better.

Since formal definitions of compactness are important in what follows, some essential features are explained in Appendix C. The primary definition C.1, which says every open cover has a finite sub-cover, is hard to apply for C_P^Q or even to much simpler spaces. Hence proofs of compactness often rely on the Heine Borel theorem which equates compactness to the more easily verifiable properties of being closed and bounded. Unfortunately this theorem is only applicable able to metric spaces that can

be embedded in $\mathbf{R}^n$ for some finite n, but C_P^Q is not a metric space, nor a subset of any $\mathbf{R}^n$ (the space of curves is, in a formal sense, infinite-dimensional). Hence we now use a different concept called sequential compactness, defined in C.2, which is basically the property that every sequence of curves has a sub-sequence that converges to a limit curve. It is known that this, together with a property called second-countability (the space has a countable basis of open sets), is equivalent to compactness.

Thus we need to verify that C_P^Q is both second-countable and sequentially compact. Now a globally hyperbolic manifold M does have a countable basis, and the basis of open sets for the space of curves C_P^Q is inherited from this basis, as explained above. Then it follows that C_P^Q is second-countable. Thus it only remains to show sequential compactness.

For this we need a few definitions involving convergence of curves.

Definition 2.12. *A point P in a manifold M is a **convergence point** of a sequence of curves $\{x_n^\mu(s)\}$ if every open neighbourhood O of P intersects $\{x_n^\mu(s)\}$ for all $n > N$ where N is some positive integer.*

The concept of convergence point is illustrated in Fig.16.

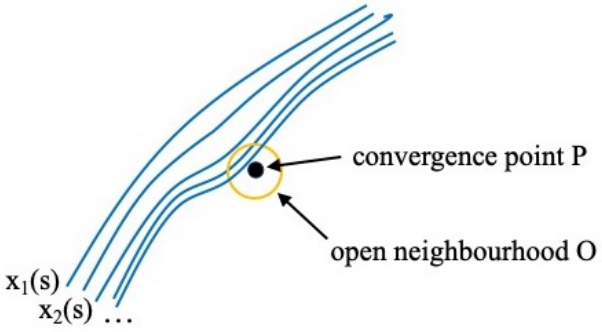


Figure 16: Illustration of a convergence point. Any open neighbourhood intersects the nth curve in the sequence for all $n > N$.

Next we use convergence points to define convergence curves:

Definition 2.13. *A curve $x^\mu(s)$ is a **convergence curve** of a sequence $\{x_n^\mu(s)\}$ of curves if each point of $x^\mu(s)$ is a convergence point of the sequence of curves.*

We can now state the following theorem:

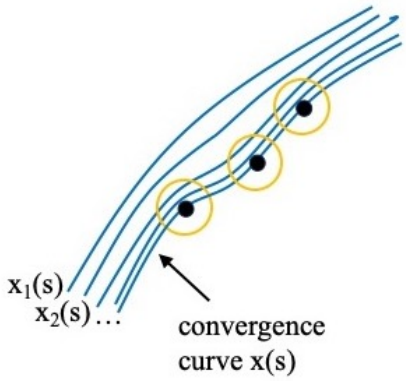


Figure 17: Illustration of a convergence curve. Each point of it has to be a convergence point for infinitely many curves in the given sequence.

Theorem 2.6. *If $\{x_n^\mu(s)\}$ is a sequence of future inextendible causal curves having a convergence point P, then there exists a future inextendible causal curve $x^\mu(s)$ passing through P which is a convergence curve of this sequence ([16], Lemma 8.1.5* [11] *).*

Let us sketch how this is proved. We are given that a sequence of inextendible curves $x_n^\mu(s)$ has a convergence point P. Thus every open neighbourhood $\mathcal{O}$ of P intersects these curves for all $n > N$. Now let us set up Riemann normal coordinates around P – these are coordinates such that the metric at the point P is $\eta_{\mu\nu}$ and first derivatives of the metric vanish. Next, take a ball around P of *coordinate radius* r in these coordinates as the open neighbourhood O (this is really a Euclidean ball, despite our being in Lorentzian signature). This ball intersects all the curves x_n^μ for $n > N$. These curves therefore pass through the spherical boundary of the ball, at a Euclidean distance r (in the chosen coordinates) from P. Since the sphere is compact, there must be a sub-sequence of these intersection points that converge to a point on the sphere. Now we take families of nested spheres of smaller and smaller radius $< r$, and taking a limit of these, we define a segment of a causal curve passing through P and going up to a coordinate distance of order r. Then we repeat the process for a new point P at the end of this curve and keep going until we get an inextendible curve which is the convergence curve of the given sequence.

Now we can return to our main theorem: that in a globally hyperbolic space-time,

[11] The proof in [16] is stated for limit points and limit curves (which we did not define or use here), rather than convergence points and convergence curves. However the latter are stronger conditions: a convergence point is also a limit point, while a convergence curve is also a limit curve.

C_P^Q is compact. We start by taking two points P and Q in $D^-(S)$ for some Cauchy hypersurface S, with Q in the causal future of P. Now we take an infinite family of curves starting at P and ending at Q. Clearly P is a convergence point (in fact all curves pass through it) so the previous theorem applies. If now we remove the point Q from the space-time M, the curves starting at P become inextendible. Then the previous theorem tells us they have an inextendible convergence curve that also starts at P. Now we put back the point Q. Then this convergence curve either remains inextendible or ends at Q. If the former is true then, since the original curves do not pass through the Cauchy hypersurface, the convergence curve also cannot pass through the Cauchy hypersurface – contradicting global hyperbolicity. Therefore the convergence curve lies in C_P^Q. This shows that C_P^Q is sequentially compact, which – together with the second countable property – means the space is compact. The proof now needs to be repeated for $P, Q \in D^+(S)$, which is straightforward, and for $P \in D^-(S), Q \in D^+(S)$ which is a little more subtle. We leave these to the reader.

The next equivalence we need to prove is between the existence of a Cauchy hypersurface and compactness of the causal diamond D_P^Q. In one direction, we have:

Theorem 2.7. *In a globally hyperbolic space-time, the causal diamond $D_P^Q = J^+(P) \cap J^-(Q)$ is compact* [12]*. In a Hausdorff space, this in turn implies that D_P^Q is closed.*

To prove this, we show that the space D_P^Q is second-countable and sequentially compact. The first property is true of any differentiable manifold M. For the second, we need to show that any infinite sequence of points $R_n \in D_P^Q$ has a sub-sequence that converges to a point $R \in D_P^Q$. Let's consider a sequence $x_n^\mu(s)$ of causal curves in C_P^Q, i.e. causal curves from P to Q, such that each one passes through the corresponding point R_n. By definition these lie entirely in D_P^Q. By compactness of C_P^Q, any infinite sequence $x_n^\mu(s)$ of curves in C_P^Q has a sub-sequence that converges to a curve $x^\mu(s)$. As a subset of M, the points on this curve make up a compact set, since the curve contains its end-points P and Q. One can now use this to argue that the points R_n in the sub-sequence converge to a point R on the convergence curve. Since the whole convergence curve is in D_P^Q, the point R must also lie in D_P^Q. Thus D_P^Q is compact ([16], Theorem 8.3.10). Using Theorem A.2 of [16], one immediately sees that D_P^Q is also closed.

[12] Recall that this was the "Hawking-Ellis definition" of globally hyperbolic space-times, but here we have assumed the Geroch definition and are deriving the other definitions as consequences.

It remains to show the converses of the above results, namely that compactness of C_P^Q or of D_P^Q implies global hyperbolicity in the sense of existence of a Cauchy hypersurface. These are left as exercises for the reader. Once this is done, we would have justified the statement that the three different definitions of global hyperbolicity listed at the beginning of sub-section 2.4 are equivalent. Hence from this point on, we may use any of them as convenient.

Theorem 2.8. *In a globally hyperbolic space-time, the causal future $J^+(P)$ of a point P is closed.*

Intuitively this can be shown by considering D_P^Q, the causal diamond of a pair of points P, Q, and taking Q off to future infinity. Then D_P^Q reduces to $J_+(P)$. A more formal proof is as follows. Suppose the contrary: $J^+(P)$ is not closed. That means there is a limit point P' of $J^+(P)$ that is not in $J^+(P)$. Now choose a point Q in the future of this point: $Q \in J^+(P')$. Then P' is in $J^-(Q)$. But P' is not in $J^+(P) \cap J^-(Q)$ since it is not in the first factor. However, P' is in the closure $\overline{J^+(P) \cap J^-(Q)}$ since it is in the closure of both factors. This is a contradiction, since we just proved that $D_P^Q = J^+(P) \cap J^-(Q)$ is closed, therefore it must contain all its limit points. Thus we have proved that $J^+(P)$ is closed.

Theorem 2.9. *In a globally hyperbolic space-time, there is a unique time-like geodesic, of maximal proper time, between P and Q.*

Recall that we only mentioned geodesics once so far, when talking about convex normal neighbourhoods (which are small neighbourhoods of a point). We will discuss them in some detail below. For now, it suffices to say that a time-like path from one point to another in Lorentzian signature has no minimum invariant proper time – it can be made up of null segments and thereby have a proper time as small as we like. However, it makes sense to maximise the proper time.

Now we can make the analogous statement about any pair of points in M. The proof of this statement follows from compactness of C_P^Q. Suppose we try to make a family of causal curves $x_n^\mu(s) \in C_P^Q$ of increasing proper time τ_n that grows without limit for large n. Since the proper time has no limit, this family cannot tend to a convergence curve. This contradicts compactness of C_P^Q. So there must be an upper bound on the proper time of a curve from P to Q for any given family of curves. Let the smallest value of this upper bound, taken over all families of curves, be $\bar{\tau}$. Now take a sequence of curves whose proper times τ_n converge to $\bar{\tau}$. Then by compactness this

must have a convergent sub-sequence converging to a path whose proper time is exactly $\bar{\tau}$. The fact that this limiting path maximises the proper time makes it a geodesic.

3 Geodesics and focal points

Now that we have established the causal properties of underlying space-time, we turn to how geodesics behave in a globally hyperbolic space-time. Geodesics are the paths followed by a free particle, so this study will lead to physical insights about how actual space-times behave. We will finally be using Einstein's equations! A brief survey of geodesics is provided in Appendix A.

3.1 Focal points and path shortening: Euclidean case

We start by studying geodesics in a space of Euclidean signature with a standard Riemannian metric. These are curves $x^i(t)$ satisfying:

$$\ddot{x}^i + \Gamma^i_{jk}\dot{x}^j\dot{x}^k = 0 \tag{3.1}$$

where t is a parameter along the curve.

At any given point, there is a unique geodesic going outward from it in the direction of any given tangent vector. Thus the geodesics from a point initially diverge in all possible directions. However, as we follow them along the manifold, depending on the metric they may converge again.

Definition 3.1. *If a family of geodesics converges at a particular point, that point is said to be a* ***focal point*** *of the family.*

A simple example is the set of geodesics going out from the north pole P of S^2 with the standard round metric. They initially diverge but then re-converge at the south pole P', which is therefore a *focal point*. In the literature, the pair of points P, P' where two geodesics intersect are also called *conjugate points*.

A focal point has the important property that when we go beyond it, the geodesic may not minimise the path length. To see this in our S^2 example, pick one of the geodesics from P to P' and continue it past P' to a point Q on the other side of the sphere. The geodesic $PP'Q$ (yellow curve in Fig. 18) does not minimise the distance from P to Q. What has happened is that, after crossing a focal point, the geodesic

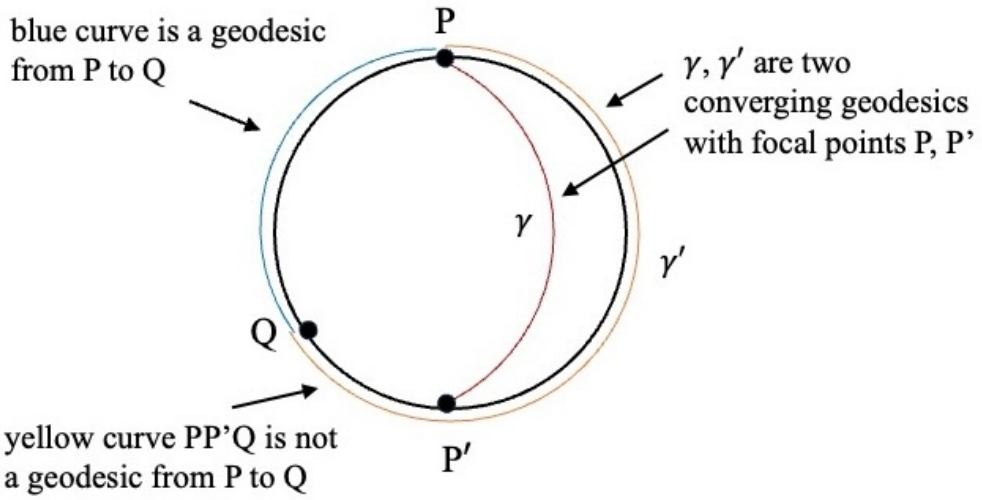


Figure 18: The yellow and red paths PP' are both geodesics. Continuing either one beyond P' leads to a curve $PP'Q$ that does not minimise distance.

starts to maximise (rather than minimise) the distance. It technically still counts as a geodesic, in the sense that it satisfies the geodesic equation and the total length is still stationary. But the its total length has changed from being a minimum to a maximum.

Indeed, there is another shorter geodesic directly from P to Q that does not pass through the south pole P', namely the blue curve in Fig. 18. This is now the geodesic that locally (and in this example, also globally) minimises the distance between P and Q. In the sphere example it is easy to visualise this "shortening phenomenon" explicitly, but we can make a more general statement: whenever a segment of a given smooth geodesic can be deformed to another segment that is also geodesic, the original geodesic will no longer be of minimal length.

If the two focusing geodesics are widely separated, the analysis becomes difficult for generic manifolds. Hence we consider a situation where two geodesic segments that start from a common point are *infinitesimally* separated from each other and then meet at a focal point (Fig. 19). Now, as first explained by Penrose [21] [13], one can always find a shorter path than the original geodesic. We find the path of minimal length in two steps, which I will call "switching" and "smoothing". First we replace the original segment (yellow curve PP' in Fig. 19) by a segment of a different, nearby geodesic (red curve PP'). This is the switching step. At this stage the total length has not changed,

[13] See page 56 of this reference. Penrose modestly calls the argument "crude", but it is now considered standard [16, 17] and I have followed the explanation in [17] which is particularly clear.

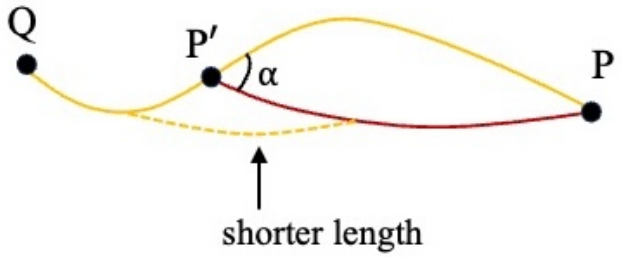


Figure 19: The yellow and red segments from P to P' are both geodesics of equal length. However neither, when continued to Q, minimises distance, since a shorter segment can be found by smoothing the kink.

since we have replaced one geodesic between a pair of points by a neighbouring one between the same pair of points. Next, we note that this procedure has introduced a kink at P', where the curve abruptly changes direction by a small angle α. Now a kink can always be locally smoothed out by taking a more "direct" path, using the triangle inequality. This is illustrated by the dotted yellow segment in the figure. This is the smoothing step, and leads to a shorter path.

Because the length of a geodesic is by definition stationary under an infinitesimal deformation, the reduction in length by smoothing an infinitesimal kink will vanish to first order, and will therefore generically be of order α^2. This has an interesting consequence. In the above discussion the focusing of geodesics gave rise to the possibility of switching, such that after smoothing we reduced the total path length. However there is a slightly more general possibility. Suppose two infinitesimally nearby paths meet at a point, where one is a geodesic but the other is only approximately a geodesic. Again we parametrise the displacement between them by an infinitesimal angle α. Technically this is not a case of "focusing", since the second path is non-geodesic. Indeed, by definition a non-geodesic fails to solve the geodesic equation to all orders in α (other than zero-th order, where it is the original geodesic). Hence it will be longer than the first one at order α^2 (to first order in α the length cannot change, due to stationarity of the geodesic). This lengthening competes against the shortening effect coming from smoothing the kink, so we cannot be sure which one wins.

However there is a condition under which the total length can still be reduced by switching to it and smoothing – we may consider non-geodesics whose failure to solve the geodesic equation starts at order α^2. In this case the change of length is of order

α^3. Then it cannot compete against the shortening phenomenon, which is always of order α^2, and therefore we can still shorten a path by going to a non-geodesic segment under the above conditions.

Analogous considerations hold when we consider geodesic paths from a point P to a sub-manifold S of M [14]. A geodesic from P to S is a stationary path ending at any point Q on S. In particular, such a path must intersect the sub-manifold S orthogonally at Q (otherwise we could reduce the length by displacing Q slightly inside S). So we consider possible geodesics from a point P outside S to the sub-manifold S, namely curves that are locally geodesic at each point and that intersect S orthogonally. We again find a possible shortening phenomenon, and the procedure for shortening again consists of the two steps of switching and smoothing. Let the original geodesic go from P to P' to Q. Let Q' be a nearby point on S such that there is a nearby geodesic from P' to Q', also ending orthogonally on S. Then we switch to the new segment, which introduces a kink. The smoothing process works the same way as before and reduces the length.

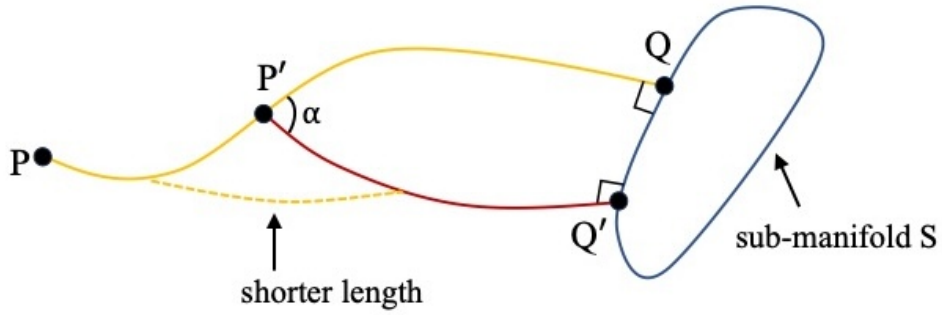


Figure 20: Geodesic focusing where the destination is a sub-manifold S rather than a point.

3.2 Focal points and path lengthening: Lorentzian case

As reviewed in Appendix A, in Lorentzian signature, a geodesic satisfies:

$$\ddot{x}^\mu + \Gamma^\mu_{\nu\lambda}\dot{x}^\nu\dot{x}^\lambda = 0 \tag{3.2}$$

[14] Here it is convenient to assume that the segments that focus arise at the end, rather than the beginning, of the path.

where λ is called an affine parameter. For the time-like or space-like case this parameter can be taken to be the proper time or proper distance respectively, while for the null case there is no geometrical quantity to which it can be related, though it can be related to a physical quantity, for example by declaring that the momentum of a particle is $P^\mu = \frac{dx^\mu}{d\lambda}$ [15]. The "length" of a path between two points P and Q is now replaced by the invariant intervals:

$$\begin{aligned} L &= \int_P^Q \sqrt{g_{\mu\nu} dx^\mu dx^\nu}, \qquad \text{space-like} \\ L &= \int_P^Q \sqrt{-g_{\mu\nu} dx^\mu dx^\nu}, \qquad \text{time-like} \\ L &= 0, \qquad \text{null} \end{aligned} \tag{3.3}$$

where $g_{\mu\nu}$ is a Lorentzian metric. Note that this makes the length ill-defined for generic paths in space-time. However we will be considering the lengths of geodesics, which are defined by parallel-transporting a tangent vector starting from a given point. The norm of the tangent then remains constant along the path and so a smooth geodesic cannot change between space-like, time-like and null along its path but has to be just one of these throughout [16].

Let us now ask about the possible lengths of geodesics, and more general paths, between two space-like-separated points P and Q. Start with Minkowski space-time and write $P = (t_P, \vec{x}_P)$ and $Q = (t_Q, \vec{x}_Q)$ with $t_P \leq t_Q$. Since the Christoffel symbol vanishes in rectilinear coordinates, a straight line joining P and Q will be a geodesic. However, there is a "shorter" path between the points! This is obtained by replacing the straight line by a "sawtooth" pattern of space-like lines, each of which is nearly null [17]. Then, as shown in Fig. 21, we get a space-like curve that is arbitrarily close to zero length. Since the proper distance L for a space-like path can never go negative, the lower bound on L is 0. Next, let us try to maximise the length. Here we see that there is no limit, since we can start at P and zig-zag all over 3d space, say at a fixed time, for as much distance as we want before getting to Q.

Similar results hold for space-like-separated points on a generic manifold. We may construct local light-cones everywhere along a space-like path between P and Q and

[15] A modern way to understand these considerations is to extremise an action defined using an arbitrary parameter, as well as an arbitrary metric, along the trajectory.

[16] A very useful discussion of geodesics can be found in [16], Section 3.3.

[17] Parts of this sawtooth curve correspond to travelling backwards in time – but we are, in any case, not ascribing any physical significance to a space-like path.

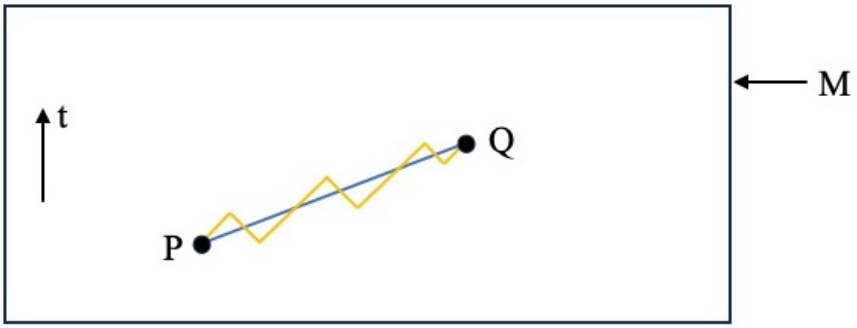


Figure 21: Sawtooth path along connecting space-like-separated points.

use them to replace this path by a nearly null one, showing that the lower bound on L is again 0. And we can use paths that travel all over space to increase L as much as we like. Thus, we see that L has a trivial lower bound of 0, and no upper bound, for any pair of space-like separated points in any space-time.

Things are different for time-like-separated points. In this situation, L is referred to as the "proper time". The minimum proper time is zero by a construction similar to the one above, except that now we use time-like segments that are nearly null. However the maximum proper time is non-trivial. We cannot zig-zag back and forth in time, since the geodesic will necessarily turn space-like in between. We have seen that in Minkowski space-time the straight line joining a pair of points is a geodesic, and for time-like separation this line actually maximises L between the points, because any additional oscillations in time-like directions will reduce the length.

We are now in a position to ask what are the proper-time-maximising geodesics on general manifolds. Here, as in the Euclidean signature case, the possibility of focusing arises, and this time it allows us to *increase* the proper time. To see this, we follow a time-like geodesic past a focal point and then find a segment of it that can be "switched" to another time-like geodesic. This "switching" part works the same way as before and introduces a kink. However in Lorentzian signature, when we smoothen a kink we get a longer proper time (this is already obvious from the fact that the sawtooth-type path has much less proper time than a straight line!).

As before, this procedure works even if the new path is not strictly geodesic, but satisfies the geodesic equation up to first order in the displacement parameter. In this case, too, we can increase the proper time by switching. Finally, if we consider time-like geodesics from a point P to a sub-manifold S, we will be able to increase the proper

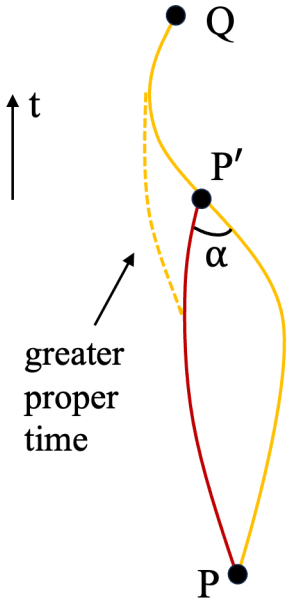


Figure 22: The yellow and red curves between P and P' are both time-like geodesics that maximise proper time between these two points. However the yellow curve $PP'Q$ does not maximise proper time, instead we find the maximising curve by switching and then smoothing as shown by the dotted curve.

time if there is a focal point in the sense that we defined previously for sub-manifolds: a point from where the initial geodesic, meeting S orthogonally, can be switched to another one also meeting S orthogonally.

Thus, to summarise, among time-like geodesics there are some that maximise proper time and others that do not. If a time-like curve does maximise proper time, it must be a geodesic. However if a geodesic encounters a focal point and we continue past it, the proper time will not be maximised.

Exercise 3.1. *Show that in a globally hyperbolic space-time, any time-like geodesic has an initial segment that maximises the proper time between its end-points.*

3.3 The Raychaudhuri equation

The $3\frac{1}{2}$-page paper [3] starts with a physical motivation for the ensuing calculation. To paraphrase the motivation in brief, it is to understand the time evolution of a gravitating system from the point of view of an observer "in its neighbourhood" and in this way

to address the problem of cosmology. However, as the paper notes, previous works by Eddington, Tolman, Bondi et al from the late 1930's to the early 1950's made use of very symmetric situations (such as homogeneous and isotropic systems) to draw broad physical conclusions. The conclusions drawn in this way had a problematic feature, namely the existence of a space-time singularity in the past or future. However, it was unclear whether this feature was due to the symmetry assumptions or to some sort of breakdown of General Relativity. For this reason, it became important to understand time evolution without making any symmetry assumptions.

PHYSICAL REVIEW VOLUME 98, NUMBER 4 MAY 15, 1955

Relativistic Cosmology. I

AMALKUMAR RAYCHAUDHURI
Theoretical Physics Department, Indian Association for the Cultivation of Science, Jadavpur, Calcutta, India
(Received December 28, 1953)

The paper presents some general relations obtaining in relativistic cosmology. It appears from these that a simple change over to anisotropy without the introduction of spin does not solve any of the outstanding difficulties of isotropic cosmological models.

Ref. [3] attempted to study this problem without making any assumptions about the symmetry of the solution [18]. Accordingly, no restrictions are placed on the type of space-time[19]. One simply assumes a space-time with Lorentzian-signature metric $g_{\mu\nu}(\vec{x}, t)$ and an initial-value hypersurface Σ. In General Relativity we can perform a general coordinate transformation $x^\mu \to x'^\mu(x)$ to go to the most convenient coordinate system for the problem at hand. Since there are four such transformations, we can fix four components of the metric. Here, we choose $g_{00} = -1, g_{0i} = 0$ because with this choice, the metric becomes:

$$ds^2 = -dt^2 + g_{ij}(\vec{x}, t)\, dx^i dx^j \tag{3.4}$$

These are called "synchronous coordinates". It is straightforward to make such a coordinate choice for a sufficiently short time proper time away from Σ, but whether this can be done far away from Σ is less clear. One of the many nice features of this coordinate system is that the negative of the determinant of the 4-metric $g_{\mu\nu}$ is the same as the determinant of the spatial 3-metric g_{ij}, and both will be denoted $|g|$.

[18] A year later, a paper with similar motivations and results was published by Komar [22].

[19] In particular it was not assumed to be globally hyperbolic. This concept, and the associated one of a Cauchy hypersurface, had in any case originated only in 1953 with the work of Leray on hyperbolic differential equations [20]. The General Relativity setting for these concepts was developed much later between 1965 and 1970 in the work of Penrose, Hawking and Geroch [4, 5, 15].

An insight of [3], though not stated in precisely these words, was to realise that the above coordinate system is equivalent to a geometric coordinate system defined in terms of geodesics, permitting physical insights. This is done as follows. We first define the families of geodesics that will be of interest:

Definition 3.2. *Given a space-time M and an open subset O of it, a* ***congruence*** *is a family of curves in the space-time such that every point $P \in O$ lies on precisely one member of the family. A congruence made up of geodesics is called a* ***geodesic congruence****.*

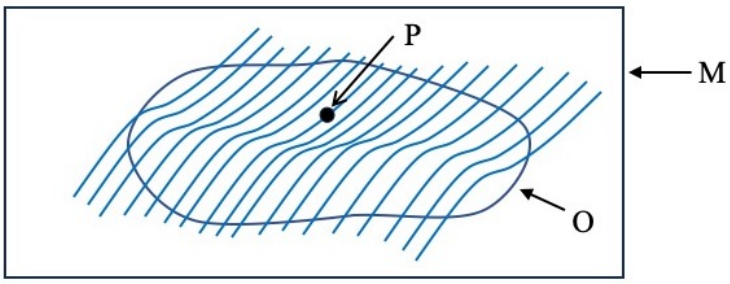


Figure 23: Illustration of a congruence. O is a co-dimension 1 hypersurface in M.

Next, we pick an initial-value hypersurface Σ and choose a spatial coordinate system $\vec{x}$ on it (this only requires ordinary Riemannian geometry, so it can always be done for a differentiable manifold). Consider a congruence of time-like geodesics passing orthogonally through Σ. Now we describe a natural choice of coordinates in the immediate causal future and past of Σ. Pick a point Q in the future of Σ, identify the unique geodesic on which it lies, and follow this geodesic back until it intersects with Σ. Since we have chosen a coordinate system on Σ, this intersection point has some definite coordinate $\vec{x}$. Then we assign to Q the same spatial coordinate $\vec{x}$, and a time coordinate t equal to the proper time along that geodesic from Σ. This is illustrated in Fig. 24. Points on Σ itself are labelled $(\vec{x}, 0)$ and points in the future are labelled $(\vec{x}, t)$ where t is the proper time measured along the corresponding future-directed geodesic.

It is easy to see that the coordinate system we have chosen using geodesics is equivalent to Eq.(3.4). To see this, start with the general metric:

$$d\tau^2 = -(g_{00}\, dt^2 + 2g_{0i}\, dt dx^i + g_{ij}\, dx^i dx^j) \tag{3.5}$$

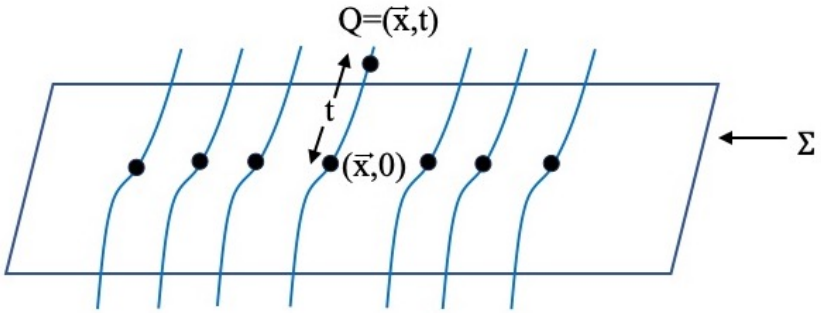


Figure 24: Assigning coordinates to a point Q in the future of Σ.

Along the geodesic, $d\tau = dt$ and $dx^i = 0$, from which we find $g_{00} = -1$. Next, because the geodesic orthogonally intersects Σ at $t = 0$ we have $g_{0i}(\vec{x}, 0) = 0$. The last remaining requirement is to show $g_{0i}(\vec{x}, t) = 0$ for $t \neq 0$. This requires the geodesic equation:

$$\ddot{x}^\mu + \Gamma^\mu_{\nu\lambda}\dot{x}^\nu\dot{x}^\lambda = 0 \tag{3.6}$$

In general the dot means $\frac{d}{d\tau}$, but in our coordinate system this is the same as $\frac{d}{dt}$. Using this in the above equation gives:

$$\Gamma^\mu_{00} = 0 \tag{3.7}$$

for every μ. This in turn sets $g_{0i,0} = 0$. Hence g_{0i}, which was set to vanish at $t = 0$, now vanishes for all time and we recover the metric Eq.(3.4) that corresponds to synchronous coordinates. In this way we have provided a geometric meaning to our chosen coordinate system.

We can now see a potential problem. If the geodesics propagating outward from Σ have a focal point, we can follow the focal point back to Σ along multiple geodesics and land on different points of Σ, giving us different $\vec{x}$ coordinates for the same point. Thus, a focal point marks a place where the coordinate system we have chosen breaks down.

This may or may not mean that something goes wrong with the manifold. There are well-known examples where a coordinate system breaks down but the geometry does not, such as the horizon of a Schwarzschild black hole, and also examples where both the coordinate system and the geometry break down, such as the singularity of a Schwarzschild black hole. So when we observe a breakdown of the coordinate system, additional analysis is needed to understand the implications for the manifold.

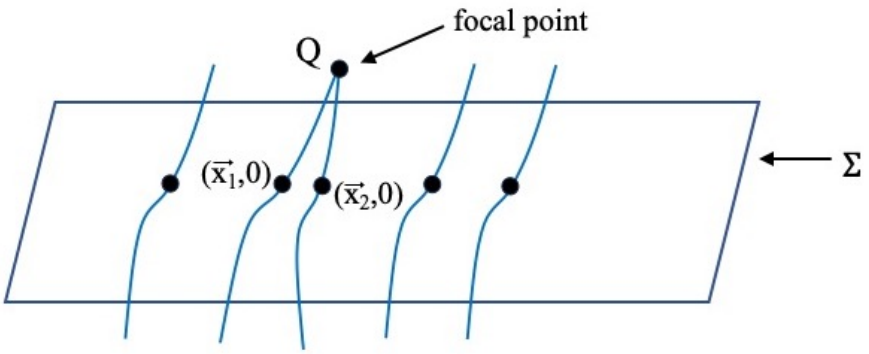


Figure 25: A focal point implies a breakdown of the coordinate system: Q has a spatial coordinate $\vec{x}_1$ as well as $\vec{x}_2$.

In terms of the spatial metric $g_{ij}(\vec{x}, t)$ in the given coordinates, a breakdown would mean that at least one of its eigenvalues vanishes – the one describing the direction along which the actual distance between coordinate-separated points tends to zero. Since it is often hard to diagonalise the metric, we use a criterion that is easier to implement, namely $|g| = 0$ [20]. Accordingly, we examine the conditions under which the determinant of the spatial metric can vanish after time evolution from an initial-value hypersurface.

Let us now calculate R_{00} in our metric. We have:

$$R_{00} = \partial_\alpha \Gamma^\alpha_{00} - \partial_0 \Gamma^\alpha_{0\alpha} + \Gamma^\alpha_{\alpha\beta}\Gamma^\beta_{00} - \Gamma^\alpha_{0\beta}\Gamma^\beta_{\alpha 0} \tag{3.8}$$

The first and third terms vanish due to Eq.(3.7). To evaluate the remaining terms, we need:

$$\begin{aligned} \Gamma^\alpha_{0\beta} &\equiv \tfrac{1}{2} g^{\alpha\gamma}(g_{\gamma 0,\beta} + g_{\gamma\beta,0} - g_{0\beta,\gamma}) \\ &= \tfrac{1}{2} g^{\alpha\gamma}\dot{g}_{\beta\gamma} \end{aligned} \tag{3.9}$$

where again we dropped terms that vanish due to Eq.(3.7). This then leads to:

$$\Gamma^0_{0\beta} = \Gamma^\alpha_{00} = 0, \quad \Gamma^i_{0j} = \tfrac{1}{2} g^{ik}\dot{g}_{jk} \tag{3.10}$$

[20] As noted in [17], $\det g_{ij} \neq 0$ may not imply that all components of g_{ij} are non-vanishing – for example, one eigenvalue of g_{ij} could go to 0 while another goes to ∞ such that the product remains fixed. But, as explained in [17], eigenvalues of g cannot diverge if the geometry of M remains smooth. Hence if $\det g_{ij}$ is non-vanishing we can be sure M is smooth. If it does vanish then, as noted above, more work is needed to figure out what is happening

and hence:

$$R_{00} = -\tfrac{1}{2}\partial_0(g^{ik}\,\dot{g}_{ik}) - \tfrac{1}{4}g^{ik}\,\dot{g}_{jk}\,g^{jl}\,\dot{g}_{il} \tag{3.11}$$

Now we define two physically relevant quantities:

$$\begin{aligned} \text{Expansion:} \qquad & \theta \equiv \tfrac{1}{2}|g|^{-1}|\dot{g}| = \tfrac{1}{2}g^{ik}\dot{g}_{ik} \\ \text{Shear:} \qquad & \sigma^i_j \equiv \tfrac{1}{2}\left(g^{ik}\dot{g}_{jk} - \frac{1}{D-1}\delta^i_j\, g^{kl}\dot{g}_{kl}\right) \end{aligned} \tag{3.12}$$

where in the first line we used Eq.(B.2). The expansion tells us how local volumes change with time (more precisely it measures the rate at which the logarithm of $|g|$ changes).

Turning now to the second definition in Eq.(3.12), we see that the shear tensor σ^i_j is traceless and vanishes for isotropic expansion $\dot{g}_{ij} = \text{const.}\, g_{ij}$, as one can easily verify. Thus we can think of it as a measure of anisotropy. Applying Eq.(B.3) to σ we have:

$$\text{tr}\,\sigma^2 = \tfrac{1}{4}\Big(g^{ik}\,\dot{g}_{jk}\,g^{jl}\dot{g}_{il}\Big) - \frac{1}{D-1}\theta^2 \tag{3.13}$$

Inserting the definitions of θ and σ in Eq.(3.11), we find:

$$\dot{\theta} = -\frac{1}{D-1}\theta^2 - \text{tr}\,\sigma^2 - R_{00} \tag{3.14}$$

This is often called the Raychaudhuri equation (for example, in [16]). It is a general identity relating geodesic expansion and shear to the time-time component of the Ricci tensor, and does not rely on Einstein's equations [21]. However, to apply it one expresses R_{00} in terms of the matter distribution, as we will do below.

We now introduce the Einstein equations (for the first time in these notes!) in the form:

$$R_{\mu\nu} = 8\pi G\,\hat{T}_{\mu\nu} + \frac{2}{D-2}g_{\mu\nu}\Lambda \tag{3.15}$$

where:

$$\hat{T}_{\mu\nu} \equiv T_{\mu\nu} - \frac{1}{D-2}g_{\mu\nu}T^{\alpha}_{\alpha} \tag{3.16}$$

We have kept an arbitrary number of space-time dimensions, since this involves no

[21] Hawking [7] notes that Raychaudhuri [3] makes simplifying assumptions, such as non-rotating dust. However, up to this point the equation is general – simplifying assumptions are only invoked after using Einstein's equation and then specialising to a choice of energy-momentum tensor.

extra work.

Exercise 3.2. *Derive the above form of Einstein's equations Eq.(3.15) by eliminating the Ricci scalar R from the usual form of the equations.*

Upon inserting Eq.(3.15) into Eq.(3.14), we find what is more widely understood to be the Raychaudhuri equation:

$$\dot{\theta} = -\frac{1}{D-1}\theta^2 - \operatorname{tr}\sigma^2 - 8\pi G\hat{T}_{00} + \frac{2}{D-2}\Lambda \tag{3.17}$$

At this point Ref.[3] assumes a specific form for the stress tensor, namely $T_{00} = \rho$ with all other components vanishing, which gives $\hat{T}_{00} = \frac{D-3}{D-2}\rho$, and takes $D = 4$ from the beginning. Imposing these additional restrictions, Eq.(3.17) is Eq.(11) of [3] [22]. As noted in [3], special cases of this equation had previously been derived by Tolman and by Synge in the presence of special symmetries and/or at special points, but the present derivation makes no such assumptions. After deriving this equation, Raychaudhuri went on to consider various implications for cosmological models.

Looking at the RHS of Eq.(3.17), we see that the first term is negative-semi-definite, while the sign of the second term depends on what kind of matter and/or cosmological constant we allow. To get a useful bound we must therefore impose some "energy conditions" on $T_{\mu\nu}$. We have intentionally separated out the cosmological constant contribution and have to consider that as well.

The "weak energy condition" says $T_{00} \geq 0$ while the "strong energy condition" says $\hat{T}_{00} = T_{00} - \frac{1}{D-2}g_{00}T^{\alpha}_{\alpha} > 0$ [23]. Both conditions are obeyed by "normal" matter. Also from Eq.(3.17) we see that a negative cosmological constant contributes with the same sign as a positive $\hat{T}_{00}$. However, a positive cosmological constant contributes to the RHS of Eq.(3.17) with the opposite sign of normal matter.

To extract a useful bound from the Raychaudhuri equation, we assume the strong energy condition and also that Λ is either non-positive or sufficiently small. With these

[22] Note that [3] does not take $g_{0i} = 0$ everywhere, but only along the time axis, i.e. at $\vec{x} = 0$. This means that only the $\vec{x} = 0$ geodesic intersects the initial-value hypersurface orthogonally. In this situation there is an additional quantity $\omega_{ij} \equiv g_{0i,j} - g_{0j,i}$ that contributes a term $-\operatorname{tr}\omega^2$ (which is positive since ω is anti-symmetric) to the RHS of Eq.(3.17). This is variously known as the "rotation" or "twist" or "spin". For our purposes it is sufficient to consider geodesics orthogonal to the hypersurface, for which the rotation vanishes.

[23] Note that, depending on the sign of T^{α}_{α}, the latter can actually be *weaker* than the former! Thus these are logically independent conditions.

assumptions, we find:

$$\dot{\theta} + \frac{1}{D-1}\theta^2 \leq 0 \qquad \text{for all } t \tag{3.18}$$

It is important to keep in mind that the validity of the bound depends on the validity of the above conditions. The above equation can equivalently be written:

$$\frac{d}{dt}\left(\frac{1}{\theta}\right) \geq \frac{1}{D-1} \tag{3.19}$$

Assuming the bound to hold, we can now derive a condition for the existence of focal points, which was our original motivation [24]. Setting $\theta(t=0) = \theta_0$ and integrating Eq.(3.18), we find:

$$\theta^{-1}(t) \geq \theta_0^{-1} + \frac{1}{D-1}t \tag{3.20}$$

This tells us that if the initial expansion is negative, i.e. $\theta_0 < 0$, then $\theta^{-1}(t)$ goes to zero in the future within a time $t_{\text{focal}} = \frac{D-1}{|\theta_0|}$. If the initial expansion is positive, we simply repeat the above argument after sending $t \to -t$ and find that there is a focal point in the past within the same amount of time. Thus to summarise, orthogonal geodesics always focus within a time:

$$|t_{\text{focal}}| = \frac{D-1}{|\theta_0|} \tag{3.21}$$

with the focal point being in the future for negative initial expansion and in the past for positive initial expansion.

Let us now comment on some physical interpretations of the variables defined in Eq.(3.12). First we consider the expansion θ. The initial-value surface Σ we have been discussing has an extrinsic curvature K_{ij} that measures, as usual, the bending of Σ in the space-time. It is shown in Section 9.3 of [16] that on Σ, the trace K of the extrinsic curvature is equal to the expansion:

$$\theta(\vec{x}, 0) = K(\vec{x}) \tag{3.22}$$

Next we define:

$$\mathcal{G} \equiv |g|^{\frac{1}{2(D-1)}} \tag{3.23}$$

[24] The motivation in [3] is similar in part, as the paper considers cosmological models with a focal point in the past – referred to as "the singular state $|g| = 0$".

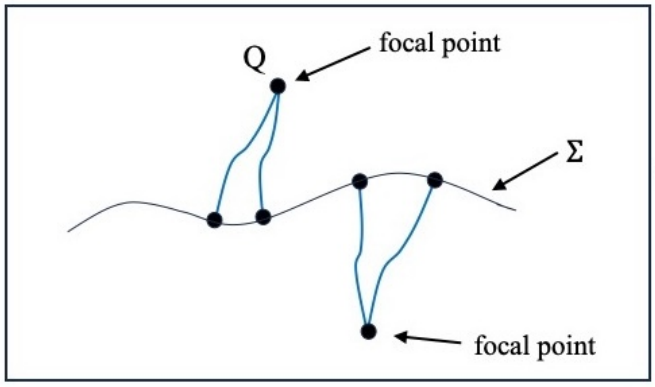


Figure 26: If $\hat{T}_{00} > 0$ there is always a focal point, either in the future or in the past.

In terms of this variable we have:

$$\theta = (D-1)\frac{\dot{\mathcal{G}}}{\mathcal{G}} \tag{3.24}$$

and the Raychaudhuri equation Eq.(3.17) is:

$$\frac{\ddot{\mathcal{G}}}{\mathcal{G}} \leq \frac{1}{D-1}\left(-\operatorname{tr}\sigma^2 - 8\pi G\,\hat{T}_{00} + \frac{2}{D-2}\Lambda\right) \tag{3.25}$$

Now suppose, temporarily, that we are looking at a homogeneous and isotropic universe described by the well-known FLRW [25] metric:

$$ds^2 = -dt^2 + a(t)^2 dx^i dx^i \tag{3.26}$$

Isotropy is encoded in the fact that the metric is rotationally invariant in 3-space. Thus the spatial part is governed by a single function $a(\vec{x}, t)$. Taking this function to be independent of $\vec{x}$ is the further assumption of homogeneity. In this metric, the Hubble parameter H is defined as:

$$H(t) \equiv \frac{\dot{a}(t)}{a(t)} \tag{3.27}$$

Now we would like to consider the situation without the symmetry assumptions

[25] Friedman-Lemaitre-Robertson-Walker

of homogeneity and isotropy. For this, notice that Eq.(3.26) is just a special case of Eq.(3.4) with the identification:

$$g_{ij}(\vec{x},t) = a(t)^2\delta_{ij} \tag{3.28}$$

It follows that in this special case:

$$\det g_{ij} = a^{2(D-1)} \tag{3.29}$$

so that:

$$a = (\det g)^{\frac{1}{2(D-1)}} = \mathcal{G} \tag{3.30}$$

In other words, our quantity $\mathcal{G}(\vec{x},t)$ generalises the scale factor $a(t)$ away from the assumptions of homogeneity and isotropic. For the metric in Eq.(3.26) it is space-independent. This motivates us to define, in the general case, a *local Hubble parameter* as:

$$H(\vec{x},t) = \frac{\dot{\mathcal{G}}(\vec{x},t)}{\mathcal{G}(\vec{x},t)} = \frac{1}{D-1}\theta \tag{3.31}$$

Now we can continue with our main task, to decide if the focusing implied by the Raychaudhuri equations corresponds to merely a breakdown of coordinates or a genuine singularity of space-time. As noted by [16, 17], such focal points can be found even in Minkowski space-time, just by taking a "wiggly" initial-value hypersurface. In this case the focal point definitely does not signal any problem of space-time, but only of our coordinate system. To determine whether space-time itself breaks down, we need additional assumptions. We now turn to this analysis, initiated by Hawking following the seminal work of Penrose [4].

3.4 Time-like geodesics and Hawking theorem

In 1933 Robertson [23] showed, using what is today called the FLRW metric, that a spatially homogeneous and isotropic universe would inevitably have a physical singularity in the past. The question was whether something analogous would be true upon relaxing the assumption of this high degree of symmetry. This was the question Raychaudhuri and Komar sought to address, apparently finding that singularities exist even in the absence of any symmetry. But several years later, in 1963, it was argued by Lifshitz and Khalatnikov [24] that the singularities they had found were purely a consequence of the special choice of the synchronous coordinates being used. It was

Hawking who then provided a serious of increasingly rigorous arguments to show that – under quite reasonable additional conditions, that however were not associated to symmetry – the occurrence of the singularities indicated by the Raychaudhuri equations was a generic reality.

As a first step, in 1965 Hawking [5] considered open FLRW universes with $k = -1$ or 0. He recast Robertson's observation into a statement about the formation of a closed trapped surface in the past, following an analogous result due to Penrose [4] in the context of black holes (which we will discuss below). He then noted, following [4], that if the space-time is time-oriented and globally hyperbolic and the weak energy condition holds then a physical singularity must occur, or else space-time is geodesically incomplete. Finally he pointed out that the above conclusions are robust under any reasonably small perturbation.

Subsequently in a series of papers [6–9] he provided different conditions under which singularities can be shown to exist. The key new idea was the existence of a Cauchy hypersurface, in other words the fact that the space-time under consideration be globally hyperbolic [26].

The most straightforward result is embodied in the following theorem:

Theorem 3.1. *Consider a globally hyperbolic space-time satisfying the strong energy condition. On a given Cauchy hypersurface Σ, let the expansion θ be everywhere positive and bounded below by some constant $-C$ with $C > 0$ (equivalently, the local Hubble constant H is bounded below by $h_{\min} = \frac{C}{D-1}$). Then all past-directed time-like geodesics reach focal points by proper time $t = -\frac{1}{h_{\min}}$ and the space-time is geodesically incomplete (Theorem 9.5.1 of [16])* [27].

To prove this result, we assume the local Hubble constant on Σ is bounded below by a positive constant $h_{\min} > 0$. From Eq.(3.31) this is the same as $\frac{D-1}{\theta_0} < \frac{1}{h_{\min}}$, which identifies Hawking's constant C with θ_0. Now, our discussion of the Raychaudhuri equation tells us that *all* past-directed time-like geodesics from Σ must reach a focal point at or before a time

$$|t_{\text{focal}}| = \frac{D-1}{\theta_0} < \frac{1}{h_{\min}} \tag{3.32}$$

Now we consider a time-like geodesic congruence that orthogonally intersects a

[26] Later the assumption of global hyperbolicity was relaxed in various ways [9, 10]

[27] In [16] one works with past-directed geodesics and consequently the conditions on θ or H have the opposite sign. This point is emphasised in Eq 4.18 of [17].

Cauchy hypersurface Σ, and work in the metric Eq.(3.4). Consider a point R in the causal past of Σ. Global hyperbolicity implies that there is a unique time-like geodesic of maximal proper time from R to Σ, that moreover intersects Σ orthogonally. However we have just shown that there is no past-directed time-like geodesic that can be continued to the past of $t = -\frac{1}{h_{\min}}$. Hence there can be no point R before this time – if there were such points, they could not have a future-directed time-like geodesic reaching Σ, which contradicts global hyperbolicity.

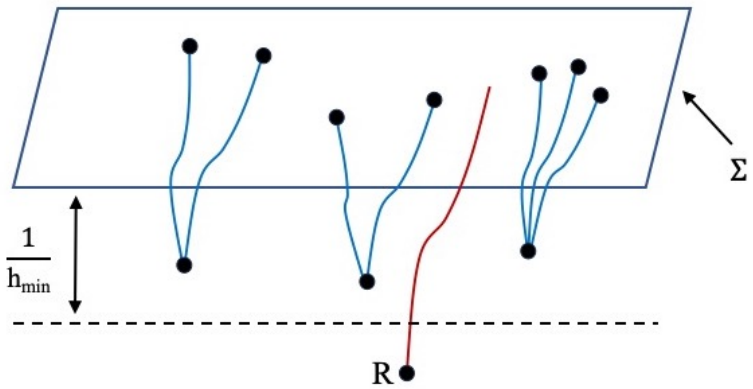


Figure 27: All past-directed geodesics must focus before $t = \frac{1}{h_{\min}}$.

Thus the space-time is *geodesically incomplete*. In the literature the situation is commonly described by saying that space-time develops a *singularity* at a finite time in the past. But as emphasised by [16, 17] among others, this is not the most appropriate phrase – what are popularly known as "singularity theorems" are really statements of geodesic incompleteness [28].

Instead of accepting the conclusions of the above argument, one could instead discard the assumption of global hyperbolicity. Then the contradiction that we found above would go away. However, there is a weaker assumption that still leads to a similar result. One assumes that M is a *strongly causal* space-time as in Definition 2.11, and also that there is a *compact* initial-value surface – an achronal hypersurface Σ satisfying suitable conditions, though not necessarily Cauchy. Under these conditions, it was shown by Hawking [9] (see also [16] Theorem 9.5.2) that there is at least *one* past-directed time-like geodesic that is incomplete, and ends at a time $t \lesssim -\frac{1}{h_{min}}$ [29]. Note

[28] A detailed discussion of this point can be found in Section 9.1 of [16].

[29] In fact, this conclusion can be reached even dropping the strong causality condition ([16], Chapter 9, Exercise 3).

that we did not previously have to assume compactness of Σ. In fact the compactness requirement in this version of the theorem means it does not apply to Minkowski space-time (and many others) where the Cauchy hypersurface is non-compact.

The conclusion from these results is that under a variety of reasonable conditions, there really is something wrong with a space-time when a focal point arises, something that we either call a singularity (following Penrose and Hawking) or geodesic incompleteness, following more modern treatments. These results give considerable physical importance to the results of Raychaudhuri and Komar [3, 22], going far beyond the possibility that they only predict the breakdown of a coordinate system. They also contradict the main result of [24], which claimed the opposite.

The above discussion has been limited to time-like geodesics. The role of null geodesics requires a separate discussion, to which we now turn, which is relevant for the physics of black holes.

3.5 Null geodesics

In the time-like case we were able to compare geodesics by asking which one maximises the proper time. However as we have seen, null geodesics have zero elapsed proper time – thus we apparently have no way to compare null geodesics! There is, however, a useful concept called "promptness" [17] that will allow us to make analogous statements to the null case.

Definition 3.3. *A causal path from P to Q is called **prompt** if there is no other causal path from P that reaches the spatial location of Q at an earlier time.*

This definition is over all causal paths, not just geodesics. However we will soon see that a prompt path is a null geodesic, although every null geodesic need not be prompt. Note also that this definition is tied to a choice of coordinates, though it is intuitively clear that it is invariant under causal changes of coordinates.

Let us assign coordinates $(\vec{x}_P, t_P), (\vec{x}_Q, t_Q)$ to the initial and final space-time points P and Q. Then we are saying that a causal path that starts at the spatial point $\vec{x}_P$ at time t_P and reaches another spatial point $\vec{x}_Q$ at time t_Q is prompt if there is no other causal path that starts at $(\vec{x}_P, t_P)$ and ends at $(\vec{x}_Q, t' < t_Q)$. Importantly, we are *not* comparing two paths from P to Q, but rather two paths from the spatial point $\vec{x}_P$ to $\vec{x}_Q$, with the latter arriving at an earlier time. We will sometimes say the second path

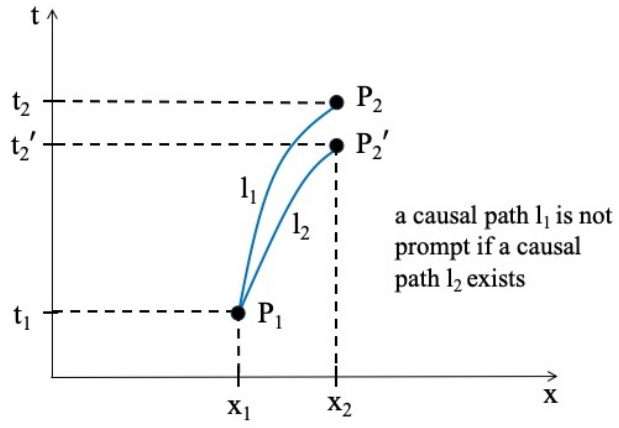


Figure 28: Illustrating promptness.

is "more prompt" than the first, but – again – this does not mean the second path has the same end-points in space-time as the first.

We can understand the situation better by recalling our discussions of the causal future of a point. For there to be any causal path from P to Q we must have $Q \in J^+(P)$. However if the path is prompt then there is no point Q' immediately to the past of Q (i.e. a point Q' with coordinates $(\vec{x}_Q, t' < t_Q)$ that can be reached by a causal path. Thus we must have $Q \in \partial J^+(P)$.

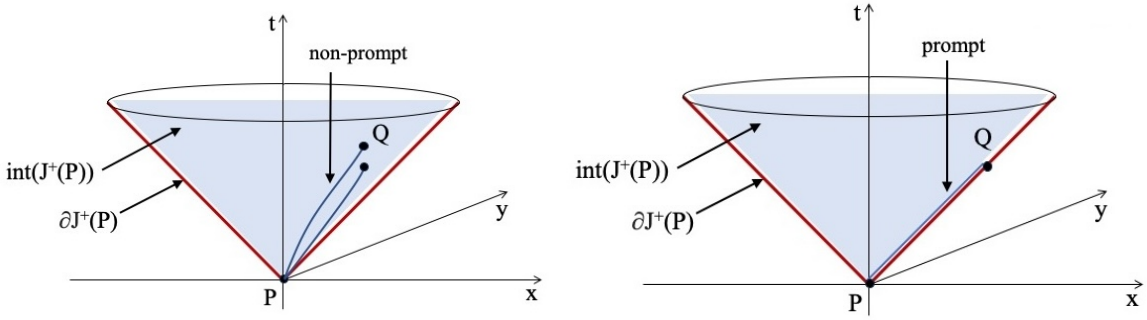


Figure 29: Illustration of promptness in Minkowski space-time.

In Minkowski space-time this means Q lies on the future light-cone of P. In this case there is a unique null geodesic from P to every $Q \in \partial J^+(P)$, and this is a prompt geodesic. This shows us clearly that non-prompt null geodesics can exist only in space-times having non-trivial curvature/topology. To see that they do in fact exist, we can use the example of gravitational lensing. In this phenomenon, light from a distant

star reaches us via multiple paths, *each of which is a null geodesic*. Different parts of the lensed image arrive at different times, so only the earliest one is prompt and the others are all non-prompt. To summarise, every null geodesic is not prompt. This is the analogue of the statement that every time-like geodesic does not globally maximise the proper time.

A simple model of non-prompt null geodesics arises from topology even when the geometry is flat. Consider a cylindrical $(1+1)$-dimensional space-time where the space direction is a circle of some radius R, with metric $ds^2 = -dt^2 + R^2 d\theta^2$ with $0 \leq \theta < 2\pi$. Looking at the causal future $J^+(P)$ of a point P, we see that it is the entire region above the two outgoing null geodesics from P. It is shaded blue in the figure. We also see that these two geodesics have a focal point at Q (on the other side of the cylinder in the drawing). The boundary of $J^+(P)$ then consists only of the two geodesics from P to Q. No part of the boundary extends to the future of Q, and in the future of Q the entire cylinder is in $J^+(P)$.

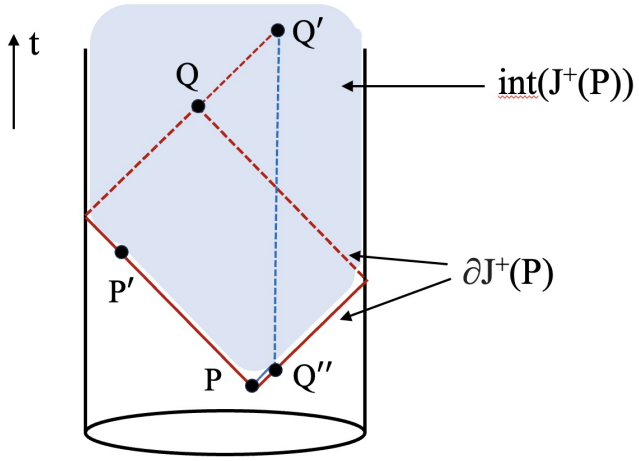


Figure 30: Example of non-prompt null geodesics.

Now the null geodesic from P to P' is prompt and lies entirely in $\partial J^+(P)$. But if we consider the point Q' beyond the focal point Q, things are different. First of all Q' is not on $\partial J^+(P)$, since the boundary ends at the time represented by Q.There is a null geodesic from P to Q', going upward toward the left of P and then continuing at the back of the cylinder to reach Q'. But this geodesic is not prompt, since one can reach the same spatial location as Q' much earlier by the short blue line, which joins P to Q''. This curve is prompt.

This example illustrates a general feature: in any globally hyperbolic space-time, a short initial segment of a null geodesic is always prompt. Now suppose a null geodesic is non-prompt up to some point, can it become prompt by continuing it further? We show that this is not possible, as follows. Take a null geodesic from $P_1 = (\vec{x}_1, t_1)$ to $P_2 = (\vec{x}_2, t_2)$ that is non-prompt, and continue it further to $P_3 = (\vec{x}_3, t_3)$. The fact that the P_1P_2 segment is non-prompt means there is another causal path starting at P_1 that reaches $(\vec{x}_2, t_2 - \epsilon)$ with $\epsilon > 0$. Now we can continue this second path from $(\vec{x}_2, t_2 - \epsilon)$ to $(\vec{x}_3, t_3 - \epsilon')$ in steps. First we extend it by a small enough amount so that the ambient space-time can effectively be treated as Minkowski. Then clearly there is such a new path, and its existence means the original path was non-prompt. We continue in this way until we reach $(\vec{x}_3, t_3 - \epsilon')$ (ϵ' does not have to be equal to ϵ, it just has to be positive). This shows that the original null geodesic remains non-prompt forever. We can call this the "shadowing" argument (the second causal curve is just under the first one and follows just below it on a space-time diagram, like a shadow. Remember that in space-time, "shadowing" actually means "reaching sooner"). We therefore conclude that a null geodesic that is non-prompt for an initial segment can never become prompt.

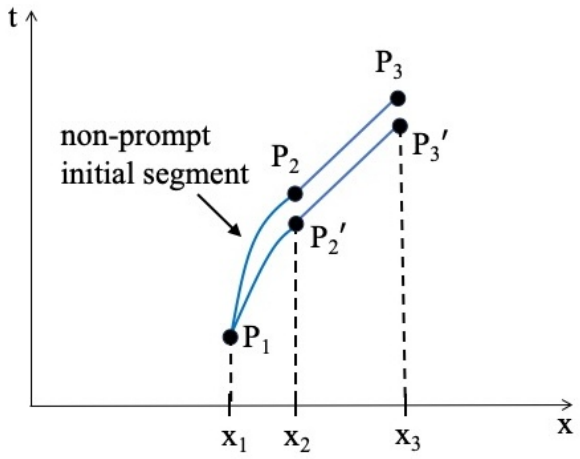


Figure 31: A null geodesic that is initially non-prompt cannot become prompt.

This reasoning also tells us that every prompt causal curve must be a null geodesic. For if there were any non-null (i.e. time-like) segment, we could "smooth it out" by a null curve just below it on a space-time diagram so the original curve would not be prompt.

We can ask a different question: if continued for a long enough distance, can a null

geodesic that was initially prompt later become non-prompt? This is indeed possible. In fact, following [17] we can argue that every non-prompt geodesic has an initial part that is prompt. Consider a point P_1, a small neighbourhood U of P_1 and a point $P_2 \in \partial J^+(P_1) \cap U$. A small neighbourhood can be well-approximated by Minkowski space-time, hence there must be a unique null geodesic from P_1 to P_2 that lies in U. We cannot yet say this geodesic is prompt, since there may be other causal paths not fully contained in U that arrive at $\vec{x}_2$ earlier. However, strong causality says there is a sub-neighbourhood $U' \subset U$ such that, when our P_1, P_2 lie in U', all causal paths from P_1 to P_2 are fully contained in V. Now we can say that the null geodesic between them is prompt. This part can be thought of as an initial segment of an extended null geodesic (which may eventually be non-prompt), showing that there is always an initial prompt part. In Figure 32, the extended part P_2 to P_3 is non-prompt while P_2 to P_3' is prompt. The cylindrical space-time depicted in Figure 30 provides another example of a non-prompt null geodesic (P to Q') whose initial part (P to P') is prompt.

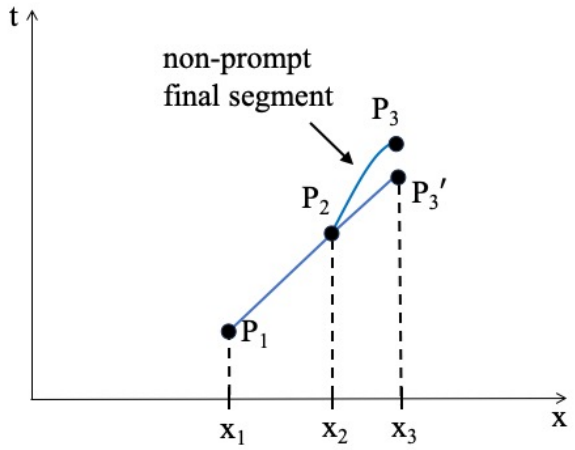


Figure 32: A null geodesic is always prompt initially.

We can now consider focal points for null geodesics. The concept is the same as for time-like geodesics: two null geodesics starting from a common space-time point (or one null geodesic and a curve that is approximately a null geodesic up to first order) may meet again in the future. We now argue that if this happens, the original null geodesic when continued past the focal point is no longer prompt. This argument is analogous to the time-like case, but there is an essential difference as we will see. We start by deforming the segment of the original geodesic before the focal point to the new one.

This creates a kink at the focal point, as in Figure 33. Now a path with a kink cannot be a geodesic, hence it is not prompt. Then we can find another causal path that gets to the final point earlier, and is therefore "more prompt" than the original path. The key point here is that we don't deform a small segment to smooth out the kink, but rather we use the kink as a tool to find a new, more prompt, path that "shadows" the original path, as explained before, reaching the destination sooner. A more detailed and rigorous discussion of promptness can be found in [17], Chapter 8. The cylindrical space-time example above illustrates how promptness fails after a focal point.

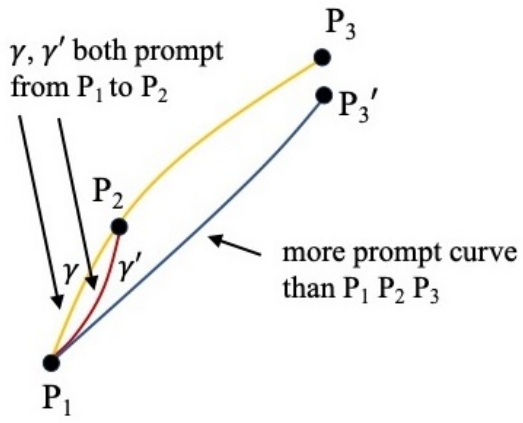


Figure 33: Focal point for null geodesics.

It should be mentioned that, while a focal point renders a null geodesic non-prompt, there could be other ways for it to become non-prompt without necessarily having a focal point.

One can show that promptness and achronality of a null geodesic are equivalent. Recall that to be achronal, a subset of M had to be everywhere space-like or null, and moreover have no time-like path in M connecting two points. Now take a null geodesic and assume it to be prompt but not achronal. Then it is possible to connect two points on it by a time-like path. This path is obviously non-prompt in its segment, and one can therefore find a more prompt "shadow" path below it (Figure 34). But now this shadow can be continued until the spatial end-point of the original geodesic, all the time remaining below it, which contradicts promptness. Thus our assumption was inconsistent and we have proved that a prompt geodesic has to be achronal.

Now suppose that the given null geodesic is non-prompt. In this case there is a shadow causal path that gets to the final destination earlier. Adding in the vertical line

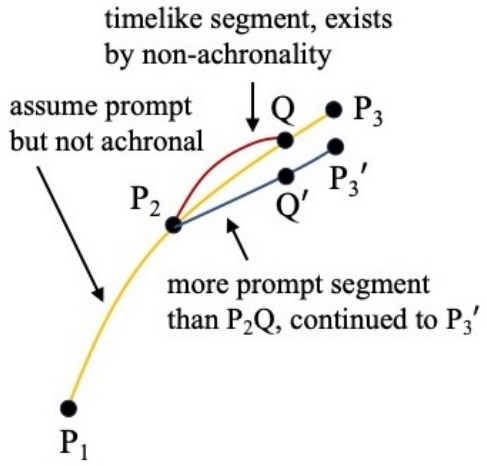


Figure 34: A prompt geodesic must be achronal.

at the end (which is time-like) gives us a (non-smooth) causal path that is partly time-like. Such a path can always be smoothened to be fully time-like. Then the geodesic is not achronal.

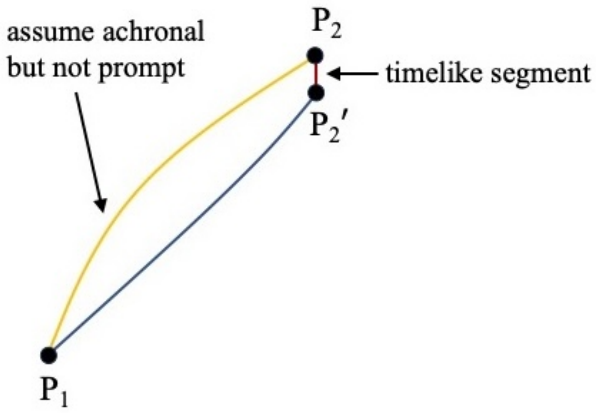


Figure 35: A non-prompt geodesic cannot be achronal.

Thus we have proved:

Theorem 3.2. *A null geodesic is prompt if and only if it is achronal.*

This gives us some new intuition about the meaning of "promptness". Referring back to the cylindrical space-time example of Figure 30, we see that the null geodesic is achronal from P to Q but once continued beyond Q, points like Q' can be connected by

a time-like path to P violating achronality. Thus in this example achronal fails exactly where promptness fails.

Exercise 3.3. *Use the equivalence of prompt geodesics to achronal null geodesics to give an alternate proof that if the initial segment of a geodesic is not prompt, then the whole geodesic is not prompt. Also use it to illustrate how an initially prompt geodesic can later become non-prompt.*

One can also define promptness for paths between a sub-manifold S of M and a point P. Such a path is called prompt it there is no other path from anywhere on S that arrives at the spatial location of P sooner (see Figure 36). A prompt causal path from the set S to the point P is necessarily an achronal null geodesic. Further, if S is everywhere space-like then the path must intersect S orthogonally. It is also true that an orthogonal prompt null geodesic from P to S must have no focal point – where a "focal point" between S and P is defined as a point from where there is more than one null geodesic that connects P orthogonally to S (for proofs, see [17]).

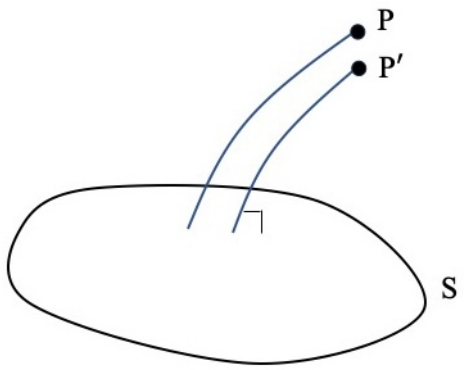


Figure 36: Promptness between a manifold S and a point P.

Now we have an elementary but important result: if a space-like sub-manifold is orthogonal to a null direction, it is orthogonal to *two independent* null directions. To see this, consider a vector W satisfying $W \cdot W = 0$. What is the sub-manifold orthogonal to W? By definition, all its tangent vectors V satisfy $V \cdot W = 0$ and this constraint defines a codimension-1 hypersurface. But W lies within this hypersurface, since it satisfies the same constraint. Hence the hypersurface we have defined cannot be space-like. In fact it is called a *null hypersurface*. To get a space-like surface we need to impose an additional condition $V \cdot X = 0$, where X is some vector satisfying $X \cdot W \neq 0$, in order

to project out the W direction. If X is also null, then we are done. If not, consider the vector:

$$W' = X - \frac{X \cdot X}{2W \cdot X} W \tag{3.33}$$

It is easily verified that: (i) W' is null, (ii) $W' \cdot W \neq 0$. Thus we have two linearly independent null vectors W, W' and a manifold S of codimension 2 whose tangent vectors satisfy: $\{V \in M | V \cdot W = V \cdot W' = 0\}$. The second condition ensures that neither W nor W' lie in the sub-manifold, which is therefore space-like [30].

Exercise 3.4. *Show that if a space-like manifold is of dimension $(D-2)$ then we can always find two independent null vectors orthogonal to it.*

An example of this phenomenon may be helpful. Suppose we are in 1+1 dimensional Minkowski space-time. Then, starting from any given point, a photon has two options to propagate – towards the positive x-axis or towards the negative x-axis. While these directions are back-to-back in space (space has just one dimension in this case!) these are two non-parallel null geodesics in a space-time diagram. Alternatively we may work in $(3+1)$ dimensions and choose a surface S to be the $x^1 - x^2$ plane at some fixed t, x^3. A photon can leave S in two ways, along the $+x^3$ direction or the $-x^3$ direction. Both are null geodesics orthogonal to the surface S. Later we will consider the case when S is a 2-sphere.

We list here three mathematical results that will be important for our future discussions on black hole singularities.

Theorem 3.3. *Consider a compact subset K of a globally hyperbolic manifold M. Then $J^+(K)$ and $\partial J^+(K)$ are both closed.*

This is a consequence of Theorems 2.7, 2.8. In the former, we proved that the causal diamond of $D_P^Q = J^+(P) \cap J^-(Q)$ of two points is closed, while in the latter we proved that $J^+(P)$ is closed for any point P. The proof we need now is similar to that used to obtain these results, and ultimately derives from compactness of the space of causal curves $C(P, Q)$ which we proved a while ago. We skip the details because they are somewhat tedious [31].

[30] A clever way to do this, due to Sachs [25], is to define two sub-manifolds S_1, S_2, each of dimension $(D-1)$ but not space-like, by the conditions that one is orthogonal to a null vector W_1 and the other to a linearly independent null vector W_2. Then the intersection $S = S_1 \cap S_2$ is the desired $(D-2)$-dimensional space-like manifold S.

[31] A sketch of the proof can be found in [17], end of Section 5.2, while in [16] this result is stated as Theorem 8.3.11 but the proof is left as an exercise.

Another important result that is easy to prove is the following:

Theorem 3.4. *The space $\partial J^+(K)$ is achronal.*

The proof is similar to that for prompt null geodesics. Suppose two points $P, Q \in \partial J^+(S)$ are connected by a time-like path. Then this path is not prompt and there is a "shadow" path below it that reaches the spatial location of Q at a time before the first one. This means Q is actually in $J^+(S)$ and not $\partial J^+(S)$. Hence we have a contradiction. This theorem has the nice corollary that any $P \in \partial J^+(S)$ can be reached from S by a prompt null geodesic lying entirely in $\partial J^+(S)$, and similarly for any pair $P, Q \in \partial J^+(S)$ there is a prompt null geodesic connecting them that lies entirely in $\partial J^+(S)$.

The last result we mention here is more subtle and would need a fairly lengthy discussion, so we refer the reader to Section 5.3 of [17] for the details.

Theorem 3.5. *The space $\partial J^+(K)$ is a manifold, though not smoothly embedded in space-time.*

While the above theorems hold for any compact subset K of M, we will be using them in the case where K is a compact $(D-2)$-dimensional space-like surface, which we have denoted S.

3.6 The null Raychaudhuri (Sachs) equation

Now we can turn to the derivation of equations for null geodesic congruences. There are both similarities and differences from the time-like situation. We start with a compact space-like sub-manifold S of dimension $(D-2)$ in M, on which we assume a coordinate system $x^a, a = 1, 2, \cdots, D-2$. Now we want to consider null geodesics emanating from it. Such geodesics satisfy the usual geodesic equation:

$$\frac{d^2x^\alpha}{d\tau^2} + \Gamma^\alpha_{\lambda\rho}\frac{dx^\lambda}{d\tau}\frac{dx^\rho}{d\tau} = 0 \tag{3.34}$$

where τ is an affine parameter, defined up to shifts and scaling $\tau \to a\tau + b$. Recall that for time-like geodesics the affine parameter was chosen to be the proper time, which specifies it uniquely apart from a shift – which was fixed by choosing the origin to lie on the initial surface. For null geodesics, the choice of the parameter is arbitrary and admits the same shift and scaling. Again we can fix the shift by choosing the origin on

the initial surface, but there is no invariant way to fix the scale a. Therefore we simply make an arbitrary choice of τ such that the affine version of the geodesic equation holds and such that $\tau = 0$ on S.

As we saw in the previous subsection, there are two null geodesics from each point of S, both of which are orthogonal to S [32]. We pick one of these, call it the u direction, and continue it all over S as in Figure 37. This generates a $(D-1)$-dimensional space Y, on which we choose coordinates as follows: (i) each point has a spatial coordinate $x^a, a = 1, \cdots, D-2$ obtained by following the null geodesic on which it lies back to S and choosing the coordinates of that intersection point, as well as a null coordinate u, chosen to be the value of the affine parameter τ along the null geodesic, starting with $u = 0$ on S. For future reference we note that in these coordinates the metric component g_{uu} vanishes, since u is a null direction. Also recall that there is an ambiguity in scaling of u and this can be space-dependent: we are allowed to replace u by $e^{f(x^a)}u$. This scales u by a positive function dependent on x^a.

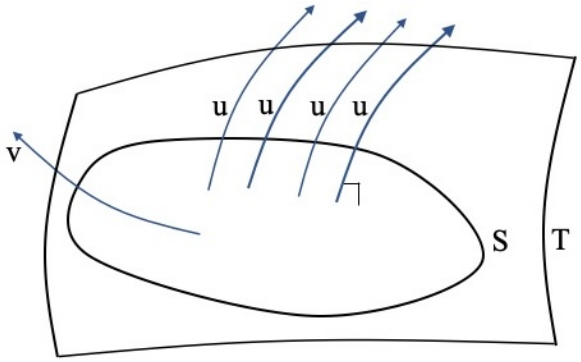


Figure 37: Forming the null hypersurface Y.

The above procedure gives us coordinates (u, x^a) on the $(D-1)$-dimensional subspace Y. Since $g_{uu} = 0$ while the x^a are spatial coordinates, the signature of Y is $(0, +, \cdots +)$. As stated earlier, such a space is called a null hypersurface. Now we want to extend this to a neighbourhood of all of M. For this we embed S into a space-like hypersurface T of dimension $(D-1)$ and choose a function v on T that vanishes along S, but whose normal derivative to S is non-zero everywhere. Then we treat the value of v as a coordinate transverse to S. Thus we can parametrise T in terms of "slices"

[32] If for example S is a 2-sphere in 4d, then one family of future-directed null geodesics goes "outward" and the other family goes "inward". In contrast, for a space-like 3-manifold in 4d there is only one set of future-directed time-like geodesics.

$S(v)$ that are all copies of S, where $S(0) = S$, as shown in Figure 38. So far, x^a were defined only on S, and now we extend them arbitrarily – but smoothly – along the v direction so that they provide spatial coordinates for every $S(v)$. Finally, consider a future-directed null geodesic from every $S(v)$, intersecting $S(v)$ orthogonally, and define the coordinate u of a point on it to be the affine parameter starting from an origin at $S(v)$. Thus we have swept out a D-dimensional neighbourhood of M with coordinates (u, v, x^a) defined as follows: follow each point P back towards the past along a null u-geodesic. This geodesic will land on $S(v)$ which determines the v-coordinate of P. The amount of affine parameter between $S(v)$ and P gives us the coordinate u. Finally x^a for P is taken to be the spatial coordinate of the intersection point on $S(v)$. This is illustrated in Figure 38.

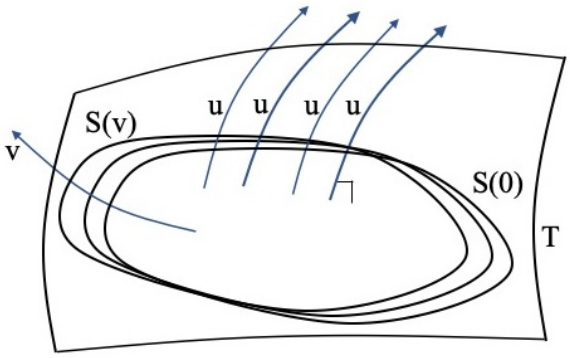


Figure 38: Constructing a neighbourhood of M in terms of slices $S(v)$.

What is the form of a general metric in these coordinates? We already saw that $g_{uu} = 0$ because the u direction is null. The metric g_{ab} of the $(D-2)$-dimensional space-like manifold S is arbitrary. Next, $g_{ua} = 0$ at $u = 0$ (i.e. on each $S(v)$) since u is defined along a null geodesic orthogonal to $S(v)$. We have not yet found that it vanishes on all of M. There are no constraints on g_{va}, g_{vv} so we write $g_{va} = c_a(u, v, x^a)$ and similarly $g_{vv} = g_{vv}(u, v, x^a)$. Finally g_{uv} must be generically nonzero, otherwise we would have $\det g_{\mu\nu} = 0$ and the space-time would be singular from the start.

This is as far as we can get without using the geodesic equation Eq.(3.34). Imposing this equation (with $\tau = u$) we get:

$$\Gamma^{\alpha}_{uu} = 0 \tag{3.35}$$

which, using $g_{uu} = 0$, translates to:

$$g_{u\beta,u} = 0 \tag{3.36}$$

for all β. Choosing $\beta = a$ we find g_{ua} vanishes everywhere since it vanishes at $u = 0$. Taking $\beta = v$ we learn that g_{uv} is independent of u, while $\beta = u$ gives us no new information since we already have $g_{uu} = 0$ everywhere. We parametrise $g_{uv} = -e^{q(v,x^a)}$. Also we found no constraints on g_{ab} so it is a general function of (u, v, x^a). Thus our metric is:

$$ds^2 = -2e^q du\, dv + 2c_a\, dv\, dx^a + g_{ab}\, dx^a\, dx^b + g_{vv}\, dv^2 \tag{3.37}$$

This has apparently left g_{vv} as an arbitrary function but in fact we have not used up all our freedom. We can make a reparametrisation $u \to u + h(u, v, x^a)$ where h vanishes at $v = 0$, which leads to:

$$du\, dv \to du\, dv + \partial_v h\, dv^2 + \partial_u h\, du\, dv + \partial_a h\, dv\, dx^a \tag{3.38}$$

This modifies the $du\, dv$ and $dv\, dx^a$ terms, but they were arbitrary anyway. Thus, at least locally, we can set g_{vv} to anything we like. Setting it to zero would make v a null direction, which is consistent with the geometric constructions above. However there is another choice: $g_{vv} = g_{ab} c^a c^b$ where $c^a = g^{ab} c_b$ [33], which turns out extremely convenient as we will see below. This choice means, of course, that the v direction is no longer null.

The metric now becomes [34]:

$$ds^2 = -2e^q du dv + g_{ab}(dx^a + c^a\, dv)(dx^b + c^b\, dv) \tag{3.39}$$

The arbitrary functions are $q(v, x^a), g_{ab}(u, v, x^a), c_a(u, v, x^a)$ with $a = 1, \cdots, D-2$, making up a total of $\frac{D(D-1)}{2}$ functions. This is the same number of functions that we had in the time-like case.

If now we compute the inverse of the above metric, we find that $g^{uu} = g^{vv} = g^{av} = 0$ everywhere. In particular, $g^{vv} = 0$ is an immediate consequence of the special form we chose for g_{vv}. Another nice feature is that taking the inverse 4d metric $g^{\mu\nu}$ and restricting the components to $\mu = a, \nu = b$ leads to a matrix g^{ab} that is actually the inverse of g_{ab} in the 2d sense. Hence we can think of $c^a = g^{ab} c_b$ with no ambiguity in the meaning of g^{ab}.

Exercise 3.5. *For the metric in Eq.(3.39), calculate all components $g^{\mu\nu}$ of the inverse*

[33] Here g^{ab} are the a, b components of $g^{\mu\nu}$, the inverse of the 4d metric.

[34] The derivation here has followed [17]. A more elegant derivation of this metric can be found in [25].

metric and verify the above statements.

We are finally in a position to write the null Raychaudhuri equation, which is essentially the component:

$$R_{uu} = 8\pi G T_{uu} + \frac{2}{D-2} g_{uu}\Lambda \tag{3.40}$$

of Einstein's equations. Recall that $\hat{T}$ was defined in Eq.(3.16):

$$\hat{T}_{\mu\nu} \equiv T_{\mu\nu} - \frac{1}{D-2} g_{\mu\nu} T^{\alpha}_{\alpha} \tag{3.41}$$

Now since $g_{uu} = 0$ in the coordinate system of Eq.(3.39), the cosmological term on the RHS of Eq.(3.40) vanishes and so does the trace term in $\hat{T}_{uu}$ [35]. Thus the energy condition we will need in order to get a bound on the expansion is the *null energy condition*:

$$T_{\mu\nu} k^{\mu} k^{\nu} > 0 \tag{3.42}$$

where k is any null vector. This is actually true for all known types of matter. Choosing $k^{\mu} = (1, 0, 0, \cdots, 0)$ in our coordinate system (u, v, x^a) leads to $T_{uu} > 0$.

Next we compute R_{uu} in terms of the functions q, c_a, g_{ab}. Similar to Eq.(3.8), we have:

$$R_{uu} = \partial_{\alpha}\Gamma^{\alpha}_{uu} - \partial_{u}\Gamma^{\alpha}_{u\alpha} + \Gamma^{\alpha}_{\alpha\beta}\Gamma^{\beta}_{uu} - \Gamma^{\alpha}_{u\beta}\Gamma^{\beta}_{\alpha u} \tag{3.43}$$

Using Eq.(3.35), this simplifies to:

$$R_{uu} = -\partial_{u}\Gamma^{\alpha}_{u\alpha} - \Gamma^{\alpha}_{u\beta}\Gamma^{\beta}_{\alpha u} \tag{3.44}$$

Thus we only need to evaluate Γ^{α}_{uv} and Γ^{α}_{ub}:

$$\begin{aligned} \Gamma^{\alpha}_{uv} &= \tfrac{1}{2} g^{\alpha\gamma}\big(g_{\gamma u,v} + g_{\gamma v,u} - g_{uv,\gamma}\big) \\ &= \tfrac{1}{2} g^{\alpha v} g_{vv,u} + \tfrac{1}{2} g^{\alpha a}(g_{av,u} - g_{uv,a}) \\ \Gamma^{\alpha}_{ub} &= \tfrac{1}{2} g^{\alpha\gamma}\big(g_{\gamma u,b} + g_{\gamma b,u} - g_{ub,\gamma}\big) \\ &= \tfrac{1}{2} g^{\alpha v}(g_{uv,b} + g_{vb,u}) + \tfrac{1}{2} g^{\alpha a} g_{ab,u} \end{aligned} \tag{3.45}$$

[35] Of course, these terms still contribute to the R_{uv}, R_{va} and R_{ab} components of Einstein's equations.

where we dropped a number of vanishing terms. From this we get:

$$\begin{aligned}\Gamma^{\alpha}_{u\alpha} &= \tfrac{1}{2}\left(g^{vv}g_{vv,u} + g^{av}g_{av,u} - g^{av}g_{uv,a} + g^{av}g_{uv,a} + g^{av}g_{av,u} + g^{ab}g_{ab,u}\right) \\ &= \tfrac{1}{2}g^{ab}g_{ab,u}\end{aligned} \tag{3.46}$$

where all other terms vanished due to $g^{uu} = g^{vv} = g^{va} = 0$. With this, the dependence of $\Gamma^{\alpha}_{u\alpha}$ on q, c^a has dropped out. The same happens for the other term in R_{uu}, which reduces to:

$$\Gamma^{\alpha}_{u\beta}\Gamma^{\beta}_{u\alpha} = \tfrac{1}{4}g^{ac}g_{bc,u}g^{bd}g_{ad,u} \tag{3.47}$$

It follows that:

$$R_{uu} = -\tfrac{1}{2}\partial_u(g^{ab}g_{ab,u}) - \tfrac{1}{4}g^{ac}g_{bc,u}g^{bd}g_{ad,u} \tag{3.48}$$

Notice the similarity with Eq.(3.11), but with a $(D-2)$-dimensional metric.

Exercise 3.6. *Explicitly verify the second term in R_{uu} in Eq.(3.48).*

Next, denote:

$$A = \det{}^{\frac{1}{2}} g_{ab} \tag{3.49}$$

We want to study the time evolution of this determinant, which will play a role analogous to that of V in the time-like case. We now define:

$$\begin{aligned}\text{Null expansion:} \qquad \theta &\equiv \frac{\partial_u A}{A} = \tfrac{1}{2}g^{ab}\partial_u g_{ab} \\ \text{Null shear:} \qquad \sigma^a_b &\equiv \tfrac{1}{2}\left(g^{ac}\partial_u g_{cb} - \frac{1}{D-2}\delta^a_b g^{cd}\partial_u g_{cd}\right)\end{aligned} \tag{3.50}$$

Again, these are very similar to their time-like versions Eq.(3.12) with the replacement of $(D-1)$ by $(D-2)$ dimensions. Inserting these definitions in the uu component of the Einstein equation, we arrive at the null version of the Raychaudhuri equation:

$$\partial_u\theta + \frac{1}{D-2}\theta^2 = -\operatorname{tr}\sigma^2 - 8\pi G T_{uu} \tag{3.51}$$

The RHS is < 0 due to the null energy condition so, as before, we get a bound. This is most easily expressed and solved in terms of a quantity:

$$\hat{\mathcal{G}} \equiv \det{}^{\frac{1}{D-2}} g_{ab} \tag{3.52}$$

which plays the role of $\mathcal{G}$ in Raychaudhuri for the time-like case. The bound then

becomes:

$$\partial_u^2 \hat{\mathcal{G}} < 0 \tag{3.53}$$

If at $u = 0$ we have $\frac{\hat{\mathcal{G}}}{\partial_u \hat{\mathcal{G}}} = -\frac{1}{\alpha}$, then one finds that $\hat{\mathcal{G}} \to 0$ by a null time $u = \alpha$ in the future if α is positive (negative null expansion), and in the past if α is negative (positive null expansion).

3.7 Trapped surfaces and Penrose's singularity theorem

Recall that there are two families of null geodesics from a given space-like surface S of dimension $(D-2)$, and we arbitrarily picked one of these families to define a coordinate system. Then we studied the null expansion θ for this family. We could repeat the whole exercise with the other family of null geodesics. This would give us a second null expansion that obeys analogous equations. Now depending on the geometry and topology of the initial space-like surface, one can imagine different situations (going towards the future): both null expansions are positive, or one is positive and the other negative, or both are negative.

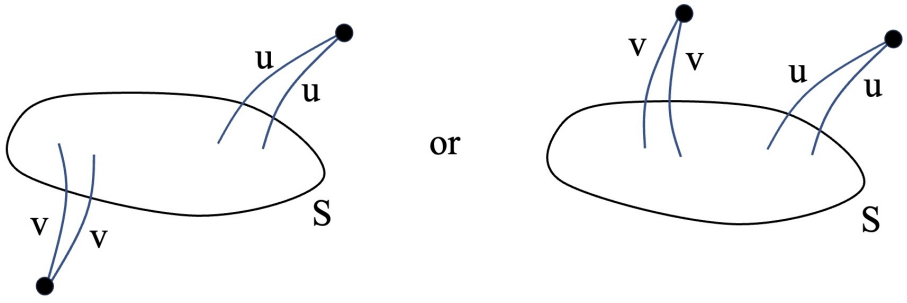


Figure 39: Left: u geodesics focus in the future and v geodesics focus in the past, Right: Both sets of geodesics focus in the future.

For example if S is the spatial 2-sphere in Minkowski space-time, the null geodesics coming out of the sphere have positive expansion towards the future while those going into the sphere have negative expansion towards the future. This can be seen intuitively by the fact that if we translate the original sphere along outward-pointing null geodesics then we get a series of concentric expanding spheres, while if we do the same along inward-pointing null geodesics then we get concentric shrinking spheres. However the other cases are also possible in some situations, and this is important for us.

Definition 3.4. *A **trapped surface** is a compact $(D-2)$-dimensional space-like submanifold S such that* both *its future-directed null expansions are everywhere negative.*

Clearly the example above of a round 2-sphere in Minkowski space-time is not a trapped surface. We will now see that such a surface cannot in fact arise in Minskowski space-time, and that it arises instead in the more complicated black hole geometry.

To understand the situation in more detail, it is simpler to go to $(2+1)$-dimensional Minkowski space-time and replace the sphere by a circle in the xy plane. Now there is a simple way to see why the null expansion starts out positive for one family of geodesics and negative for the other. The initial circle, though independent of time, is dependent on u since its metric is $ds^2 = r^2 d\theta^2 = (u-v)^2 d\theta^2$ where $u = t + r, v = t - r$. Thus $A = u - v$ and we have $\partial_u A = 1 > 0$ while $\partial_v A = -1 < 0$, proving the intuitive fact that the circle embedded in this way has one positive and one negative null expansion.

We can make more complicated initial surfaces S in Minkowski space-time by deforming the sphere. In this case it can happen that both null expansions are positive in some regions and both are negative in some other regions. However one cannot find a surface such that both its null expansions are negative (or both are positive) *everywhere*. Thus there is no trapped surface in Minkowski space-time. But this does not rule out that we could find such a surface in some other space-time. As we will see, a black hole is a place where we do find trapped surfaces.

Before asking when trapped surfaces arise, let us ask what are the consequences if they do arise. The consequences of a trapped surface are embodied in Penrose's famous theorem, which we now state as follows:

Theorem 3.6. *Let M be a globally hyperbolic space-time having a non-compact Cauchy surface Σ and satisfying the null energy condition. If M contains a compact trapped surface, then it is not geodesically complete.*

First let us provide some intuition about this theorem. The basic strategy is to carefully examine the $(D-1)$-dimensional null hypersurface $\partial J^+(S)$. The argument that prompt geodesics are achronal can be used to show that, in fact, the whole of $\partial J^+(S)$ is achronal. Next we show that $\partial J^+(S)$ is compact, using the fact that it is generated by null geodesics out of S. Next we assume that M has a non-compact Cauchy hypersurface Σ. Finally we recall that any achronal hypersurface must be homeomorphic (topologically equivalent) to part of a given Cauchy hypersurface– but,

as Σ is non-compact while $\partial J^+(S)$ is compact, this is impossible. Hence there is a contradiction and M is incomplete.

A very simplified proof is given below. We start with the two null vectors orthogonal to S, let us call them W and W' as before. Consider a null geodesic with affine parameter u going to the future along W from the trapped surface S, and another one with affine parameter v similarly going to the future but along W' from S. We now assume S is a trapped surface, so $\frac{\hat{\mathcal{G}}}{\partial_u \hat{\mathcal{G}}} = -\frac{1}{\alpha} < 0$ at $u = 0$ (initial negative null expansion along W) and along the second geodesic with $\frac{\hat{\mathcal{G}}}{\partial_v \hat{\mathcal{G}}} = -\frac{1}{\alpha'} < 0$ at $v = 0$ (initial negative null expansion along W').

In each of these cases, if the starting assumption holds (and space-time is globally hyperbolic and satisfies the null energy condition) then the null Raychaudhuri equations imply that $\hat{\mathcal{G}} \to 0$ along the first geodesic by a null time $u = \alpha$ and along the second by a null time $u' = \alpha'$ respectively. Thus both the geodesics reach a focal point within a finite affine time. However, as is always the case with the Raychaudhuri equations, we cannot necessarily conclude that something is wrong with the space-time from these facts. It may be possible to extend the geodesics in some way beyond these points. But Penrose's theorem says that, if we impose the additional condition of a non-compact Cauchy hypersurface, then *at least one* of these geodesics cannot be extended beyond the focal point. Thus the space-time is truly geodesically incomplete. As discussed earlier in these notes, this may or may not mean there is a "singularity" and it is best to avoid that word since we don't know precisely what physical phenomenon is taking place there. All we know is that in a precise sense space-time is breaking down.

To prove the theorem, our main goal is to prove that $\partial J^+(S)$ is compact, and thereby find a contradiction. First we prove a lemma associated to any prompt null geodesic γ from S to a point $Q \in J^+(S)$. It starts at some point $P \in S$, travels along u to a focal point Q and is extended beyond that. The initial segment γ_{PQ} of this geodesic is prompt, which means $Q \in \partial J^+(S)$. Also this segment is compact as a set (it contains its end-points and is bounded). Now consider the intersection $\gamma \cap \partial J^+(S)$ where γ is the whole (extended) curve. We are going to show that $\gamma \cap \partial J^+(S)$ is compact.

Here we can invoke the theorems we stated (and in some cases proved) earlier. First, $\gamma \cap \partial J^+(S) \subset \gamma_{PQ}$, because after Q the geodesic becomes non-prompt and no longer travels in $\partial J^+(S)$. Moreover it is a closed subset of γ_{PQ} since $\partial J^+(S)$ is closed in the space-time M (Theorem 3.3 of these notes). Since γ_{PQ} is compact, the intersection $\gamma \cap \partial J^+(S)$ is also compact, which proves the lemma.

Now this can be repeated for every point $Q \in \partial J^+(S)$ and the prompt null geodesic to it from S. Thus every such intersection $\gamma \cap \partial J^+(S)$ is compact. We then repeat this for curves γ' along u' to find that their intersections $\gamma' \cap \partial J^+(S)$ are also compact. Together, this shows that $\partial J^+(S)$ is itself compact. Now in Theorem 3.4 we showed that $\partial J^+(S)$ is achronal. However in sub-section 2.3 we showed that an achronal $(D-1)$-dimensional hypersurface cannot be compact if the Cauchy surface Σ is non-compact. This is a contradiction. Therefore, at least one of the two families of geodesics cannot be extended beyond the focal point, and Penrose's theorem is proved.

=========

4 Black holes

4.1 The Schwarzschild solution

From now on we will work in $D = 4$, though most of the results have generalisations in higher dimensions. It has been known since over a century that the vacuum Einstein equations $R_{\mu\nu} = 0$ have a simple spherically symmetric solution:

$$ds^2 = -\left(1 - \frac{r_H}{r}\right) dt^2 + \left(1 - \frac{r_H}{r}\right)^{-1} dr^2 + r^2 (d\Omega_2)^2 \tag{4.1}$$

where $r_h > 0$ is a free parameter. Suppose the mass distribution of the solution is non-zero up to a radius $\tilde{r}$. Then this describes the gravitational field in the vacuum prevailing in the region $r > \tilde{r}$. Now in the weak-field (Newtonian) approximation it is easy to show that the 00 component of the metric is:

$$g_{00} = -\left(1 - \frac{2GM}{r}\right) \tag{4.2}$$

from which we get the identification $r_H = 2GM$.

The coordinates θ, ϕ parametrise a 2-sphere. Radial light rays are null and orthogonal to this sphere, so we find them by solving $ds^2 = 0$ with $d\theta = d\phi = 0$. This leads to:

$$dt = \pm \frac{dr}{1 - \frac{2GM}{r}} \tag{4.3}$$

We only consider future-directed rays, so dt is always positive. Now the $+$ sign above corresponds to outgoing rays (r increases with t) while the $-$ sign corresponds to ingoing

rays. As a check we can take $r \gg 2GM$ and then the rays are $dt = \pm dr$ just as in Minkowski space-time.

Now, the solution Eq.(4.1) appears to break down at $r = 2GM$ because the dt^2 term vanishes and the dr^2 term blows up. This is not a problem as long as $\tilde{r} > 2GM$, because in that case we are not supposed to use a vacuum solution all the way down to $r = 2GM$ in the first place. Indeed $\tilde{r} \gg 2GM$ for the earth, moon, sun and any other "normal" massive spherical body (for the earth, $\tilde{r} \simeq 6400$ km while $2GM \simeq 1$ cm!). However if $\tilde{r} < 2GM$ then we are allowed to go down to $r = 2GM$ and there we encounter the apparently singular behaviour of the solution. As long as we keep working in the (t, r, θ, ϕ) coordinates our space-time appears to end at $r = 2GM$ and we cannot go to $r < 2GM$. If we try to do so, we will come to wrong conclusions or – at least – will not be able to claim this region is connected in any way to the $r > 2GM$ region. A related problem is that, from Eq.(4.3), we seem to have just one null geodesic $dr = 0$ at $r = 2GM$. Something is clearly wrong at $r = 2GM$.

Yet, one finds that the scalar quantity $R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma}$ is finite and well-behaved at $r = 2GM$, which suggests that this could be just a "coordinate singularity". We can show that this is so by finding coordinates in which the metric is finite at $r = 2GM$. We define:

$$u = t - r^* \text{ (outgoing)}, \quad v = t + r*, \text{ (ingoing)} \tag{4.4}$$

where:

$$r^* = r + 2GM \log \left| \frac{r}{2GM} - 1 \right| \tag{4.5}$$

Note that:

$$\frac{dr^*}{dr} = \frac{1}{1 - \frac{2GM}{r}} \tag{4.6}$$

and observe that that there is no modulus sign in this equation.

Exercise 4.1. *Verify the above statement.*

Next we choose our independent coordinates to be (u, r, θ, ϕ) or (v, r, θ, ϕ) [36]. These are called outgoing/ingoing Eddington-Finkelstein coordinates. Choosing the ingoing one, we find the metric to be:

$$ds^2 = -\left(1 - \frac{2GM}{r}\right) dv^2 + 2dv\,dr + r^2(d\Omega_2)^2 \tag{4.7}$$

[36] Note that we are not using r^* as a coordinate, but only as a way to define u, v. Also we are not using both u, v together. There are other treatments that do both of these things, but here we are restricting to the coordinate systems that are most useful in our discussion.

In this coordinate system, r has become a "radial null coordinate" ($g_{rr} = 0$). We see that the above metric is smooth for $r = 2GM$. At this value, both v and r become null coordinates but there is no singularity in the metric or change in the signature of space-time. Thus, in Eddington-Finkelstein, coordinates we are entitled to go continuously from $r > 2GM$ to $r < 2GM$. Note that in the original coordinates (t, r, θ, ϕ) we could take either $r > 2GM$ or $r < 2GM$, but we cannot smoothly interpolate between them since they become ill-defined at $r = 2GM$, hence we cannot use those coordinates to relate any object outside (such as a null vector) to the corresponding object inside.

Exercise 4.2. *Derive Eq.(4.7) by starting with Eq.(4.1) and making the given changes of variable.*

The two radial null rays in these coordinates are found by setting $ds^2 = 0$ in Eq.(4.7) together with $d\theta = d\phi = 0$ as before. We get:

$$dv = 0, \quad dv = \frac{2dr}{1 - \frac{2GM}{r}} \tag{4.8}$$

By going to large r, it is easily seen that the first one is $dt = -dr$ so it is ingoing. The second null ray can be written:

$$dt + dr^* = \frac{2dr}{1 - \frac{2GM}{r}} \tag{4.9}$$

and from this we see that it is outgoing. In fact for large r it just reproduces $dt = dr$.

So far we have not done much beyond changing coordinates from Eq.(4.3). But we now have a single space-time for all $r > 0$ so we can smoothly continue these rays to the region $r \leq 2GM$. Let us follow a null ray starting at infinite radial distance, continue it to small radial distance and see how it behaves along the way.

The first (originally ingoing) null ray $dv = 0$ does not depend on r so it does not change as we vary r, and continues to correspond to an ingoing null ray. However for $r < 2GM$ the second null ray becomes:

$$dt + dr^* = -\frac{2dr}{\frac{2GM}{r} - 1} \tag{4.10}$$

Now as t increases, r *decreases*, so this ray (which was outgoing far away) has also become ingoing! Thus a sphere of radius $r < 2GM$ has two independent ingoing null rays. It is a trapped surface.

Let us now consider the other singular point of the original solution Eq.(4.1), namely $r = 0$. Unlike the apparent singularity at $r = 2GM$ that we were able to remove by changing coordinates, this one is a genuine non-removable singularity. To show this, one can calculate that $R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma} \sim \frac{1}{r^6}$ as $r \to 0$. Since this quantity is invariant under general coordinate transformations, the singularity as $r \to 0$ is a genuine feature of the geometry that cannot be removed by any change of coordinates. The overall picture of black hole formation is illustrated in Figure 40.

From the above considerations, we find two somewhat distinct key features of this black hole. One is that no signal can go out of it: once we are inside, all directions towards the future point inwards. Thus $r = 2GM$ is like a one-way membrane, and is called the "event horizon". The interior is a "black hole" – a region of space-time from where nothing can escape, not even light signals. The other is that any freely falling object inside the black hole will reach the singularity. In fact, it can be shown that along null geodesics, it takes a finite amount of affine parameter to reach the origin along each of the null directions.

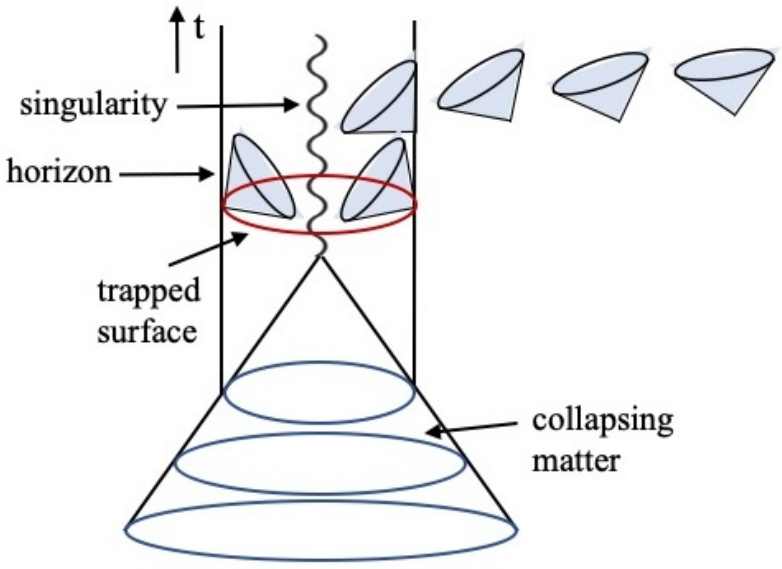


Figure 40: Black hole formation from collapsing matter.

It is important to realise that the word "horizon" can mean two slightly different things. It is defined as the hypersurface $r = 2GM$, so it has a total of three dimensions (more generally $(D-1)$ dimensions). Two of these are (θ, ϕ), the angular dimensions. The third one can be understood by examining the two null rays in Eq.(4.8). The null

ray $dv = 0$ is, as we saw earlier, is unaffected by the radial coordinate. But for the other null ray, $dv = \frac{2dr}{1-\frac{2GM}{r}}$, the denominator of the RHS vanishes as $r \to 2GM$ which means it has become $dr = 0$. This is as expected, since the metric Eq.(4.7) loses its first term and then both v and r are null coordinates. Hence on the horizon, this null ray is neither outgoing nor ingoing but *stays parallel to the horizon all the way into the future*, with v being the affine parameter along it. This constitutes the third direction of the black hole horizon. Since this direction is null, the horizon is a null hypersurface of signature $(0, +, +)$.

But sometimes we talk about the spatial horizon, which is $\partial(B \cap \Sigma) \equiv H \cap \Sigma$, namely the boundary of the intersection of the black hole with a Cauchy hypersurface Σ. This is what we would see in an experiment (which is conducted in a fixed frame of reference in the neighbourhood of a fixed time). When we talk of "horizon area", as we will do below, we mean the area of the section of the horizon at a fixed value of v.

Note that Penrose's theorem predicts somewhat less than what we know about the Schwarzschild solution. It says that at least one of the future-directed null geodesics from the trapped surface becomes inextendible after some affine time, but here we know what happens to both the null geodesics – they terminate on the singularity at $r = 0$. In fact this result pre-dates Penrose's work, and we did not need his theorem to discover it. But the key point, as in all our previous discussions, is that we are not interested only in the highly symmetric Schwarzschild solution. By itself it is not sufficient to acquire a general understanding of black holes, for which we have to understand generic situations and not highly symmetric ones. We now move on to a discussion of generic black holes.

4.2 Cosmic censorship

In going beyond spherically symmetric black holes we must be careful to make suitable assumptions that allow us to use the known laws of physics. To see why this is essential, let us briefly think once more about the Schwarzschild black hole. It has two special features that arise together: the fact that there is an event horizon that does not allow signals to escape outwards, and the fact that there is a singularity. This is true of several other black hole solutions. Because these two properties occur together, the existence of singularities will never matter for us living in the world outside all black holes. In order to observe singularities (or geodesic incompleteness) we would have go inside a black hole, and would never be able to come back to report their effects.

Hence, whatever these effects may be, they cannot influence the results of experiments performed outside, and the usual laws of physics continue to apply.

If instead we allow spacetimes having a singularity but no horizon, things would be very different. Such an object is called a "naked singularity" and it could affect experiments anywhere in its causal future and thereby destroy the predictivity of GR. We could not assume globally hyperbolic space-times or postulate things like "there exists an asymptotic observer at infinity" because such an observer may not exist.

This motivates us to put forward the conjecture of "cosmic censorship". This states that any singularity in space-time arising from gravitational collapse is hidden from the outside world by an event horizon. In the form usually called "weak cosmic censorship" it is assumed that space-time is asymptotically flat (Minkowski). With this hypothesis, the region of a space-time outside all black holes is normal, in the sense that asymptotic regions in spatial and time directions extend indefinitely and there is no geodesic incompleteness.

We will not spell out all forms of the cosmic censorship conjecture here. Indeed there is not a complete consensus whether any form of the conjecture is true or whether naked singularities are allowed to arise in some situations [37]. Instead of trying to be most general, we will see where minimal versions of the cosmic censorship conjecture lead us.

4.3 Generic black holes

Once we go away from specific classical solutions, we need to define what a black hole and an event horizon mean in general. We consider a globally hyperbolic and asymptotically flat space-time M, and an asymptotic observer, roughly at rest arbitrarily far in the future so that their world-line I is (approximately) vertical. The fact that such an observer exists is a consequence of cosmic censorship, because otherwise there might not be points of M very far to the future.

Now recall that the causal past $J^-(I)$ of this world-line is the set of all space-time

[37] However there are both theoretical and experimental inputs. First, numerical simulations of gravity fail to provide any counterexamples to cosmic censorship – since these are not constrained by symmetries or simplifications in the equations, failures of cosmic censorship could well show up, but they don't. Second, in 1972, Penrose speculated that the collision of two black holes might lead to a naked singularity but today, gravitational wave experiments have provided large numbers of events where there is no sign of such an outcome. On the contrary, the data clearly support the notion that when two black holes merge, the result is a bigger black hole.

points from where causal (time-like or null) paths can reach this observer. If there is a black hole in M having an event horizon, then clearly its interior will not lie in $J^-(I)$. This can be used to define the interior without knowing any details of the black hole. We simply define $B = M \setminus J^-(I)$, the complement of $J^-(I)$ in M, and declare that B is the "black hole region".

What topological information do we have about $J^-(I)$ and B? There is a result ([16], page 308) that $J^-(I)$ is an open set. The formal proof relies on a corollary to our Theorem 2.1 on the existence of convex normal neighbourhoods. A more intuitive proof ([17], Section 6.2) says that if P is a point on I and Q is a point in $J^-(I)$, then a causal path from Q to P can always be deformed to a causal path from Q' near Q, to the spatial point $\vec{x}_P$. Thus every point in $J^-(I)$ has an open neighbourhood completely contained in $J^-(I)$, which is the definition of an open set. Now in a topological space the complement of an open set is closed (by definition), so we see that the black hole region B is closed. This means it contains its boundary, which we define to be the event horizon: $H = \partial B$.

Now suppose that somewhere in M there is a trapped surface S. Let us prove that it is entirely inside B. If any part of it is outside B, then it is in $J^-(I)$. But this means that null geodesics from S do reach points P on I. Now among the null geodesics, one of them is prompt. We can show this as follows. Take families of geodesics that reach the world-line I earlier and earlier, and use the fact that families of curves are compact. That means there is a limiting curve that reaches earliest, and by definition that is an orthogonal prompt null geodesic. However all such null geodesics were proved to focus in finite affine parameter, while the observer on I is arbitrarily far away. This is a contradiction. Hence S has to be completely inside B. A formal proof is in [16] Theorem 12.2.2.

We have now essentially proved Penrose's fundamental physical result: the formation of black holes occurs generically even without the assumption of spherical symmetry. The statement is that if a trapped surface exists then in the future, a "singularity" (geodesic incompleteness) is inevitable. Now a trapped surface will form in gravitational collapse even if the initial conditions deviate from spherical symmetry, at least within some range of parameters. The reason is that once we have both focal points in the future (the defining property of a trapped surface) then one or both focal point cannot suddenly change to the past. Moreover we proved that a trapped surface is completely inside the black hole region. Then at least one of the families of orthogonal geodesics focuses and terminates, so the space-time is geodesically incomplete. Note

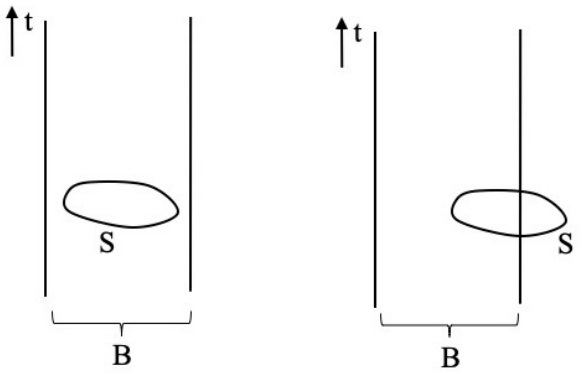


Figure 41: Left: S inside B, Right: S partly outside B.

that all this holds only assuming cosmic censorship.

A corollary of the above is that if a set $S \subset B$ then so is its entire causal future: $J^+(S) \subset B$. The proof is as follows. Take a point $Q \in J^+(S)$. It must lie in $J^+(P)$ for some $P \in S$. Now if Q were outside B, it could influence an observer's world line in the far future which means it lies in $J^-(I)$. But since there is a causal curve from P to Q, this means $P \in J^-(I)$. Since B and $J^-(I)$ are disjoint, this contradicts the fact that $P \in S \subset B$.

From this we can deduce the following important result:

Theorem 4.1. *A black hole cannot split into two black holes. However two black holes can join into one.*

To prove this, suppose (as a very unrealistic idealisation!) that at some time T there is exactly one black hole B in the universe. Then any Cauchy surface Σ after that time (i.e. a surface containing only points $(\vec{x}, t)$ with $t > T$) will intersect it. Likewise suppose at a later time there are two black holes B_1, B_2 in the universe. Then a Cauchy surface Σ' after t_2 intersects both black holes. Now consider a future-directed null geodesic coming out of $B \cap \Sigma$. We already showed that this geodesic must remain inside the black hole region of space-time forever. Therefore if it reaches Σ' then it must enter either $B_1' \cap \Sigma'$ or $B_2' \cap \Sigma'$.

This much was for a single geodesic. Now consider the entire family of null geodesics coming out of $B \cap \Sigma$. Either this whole family goes into $B_1 \cap \Sigma'$, or the whole family goes into $B_2 \cap \Sigma'$, or the family splits and part of it goes into $B_1 \cap \Sigma'$ and the rest goes into $B_2 \cap \Sigma'$. If the whole family goes into one black hole, say B_1, then it means B_1 is

the same black hole as B and it has simply evolved in time. The presence of B_2 then has to be explained by assuming it formed through independent collapse of some other spherical shell. In this scenario, B did not split. In the contrary scenario, part of the family of geodesics enters $B_1 \cap \Sigma'$ and the rest enters $B_2 \cap \Sigma'$. This is exactly what we mean by saying B has split into B_1 and B_2. However if this was the case, it would mean we can divide the space of null geodesics from $B \cap \Sigma$ into two disjoint parts, contradicting the fact that the future light cone is connected (for a detailed proof, see [16], theorem 12.2.1). Thus we have proved that a black hole cannot split. On the way we have also proved that a black hole cannot disappear (classically) since then the outgoing geodesics from $B \cap \Sigma$ could never intersect any future Cauchy surface, contradicting global hyperbolicity.

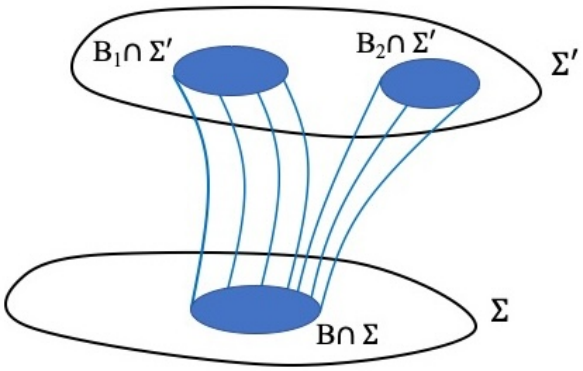


Figure 42: A black hole cannot split in two.

Now consider the converse: can two black holes merge? This seems to be the time-reverse of something we just ruled out. However this is possible for a simple reason – geodesics are certainly allowed to enter a black hole even if they were outside in the past. So if we start with two black holes B_1, B_2 and later on have just one black hole B', it need not be true that the geodesics from $B_1 \cap \Sigma$, $B_2 \cap \Sigma$ are the *only* ones to enter $B' \cap \Sigma'$. In fact this cannot be true by the time-reverse of the above theorem. However there can certainly be many other geodesics passing through the initial Cauchy hypersurface that are not in any of the black hole regions $B_1 \cap \Sigma$, $B_2 \cap \Sigma$, that in the future end up in $B' \cap \Sigma'$. Thus there one can have a connected geodesic bundle that comes partly from the merging black holes and partly from outside, that together disappears into the future black hole. This would not violate our connectedness theorem above. Clearly the one-way nature of the black hole horizon has led to the fact that splitting and

joining of black holes are not just time-reverses of each other, and we have shown that the first is impossible while the second is possible.

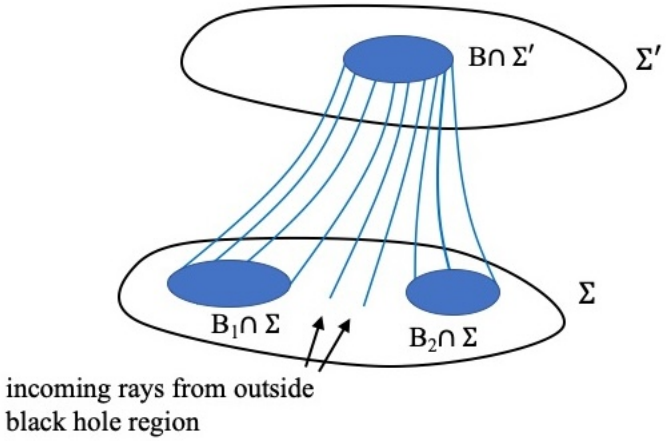


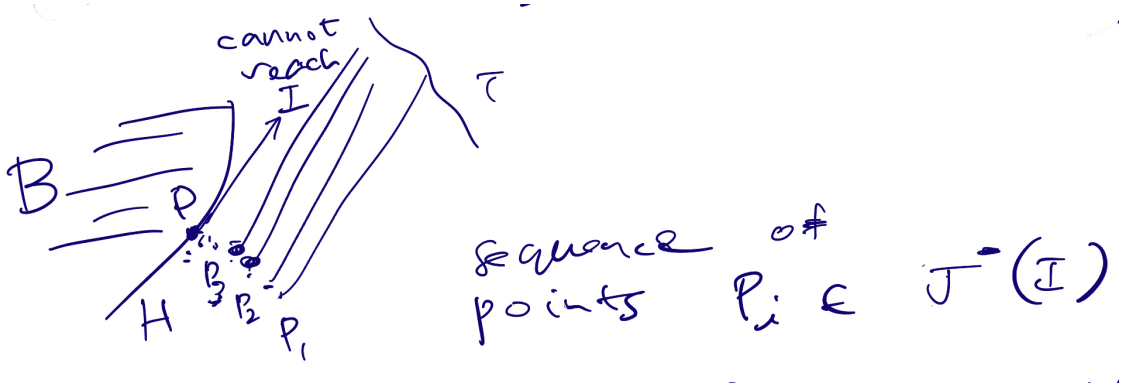
Figure 43: Two black holes can merge into one.

4.4 Hawking's area theorem

The last theorem we want to prove is the Hawking area-increase theorem: in any process, the area of a black hole can only increase (if two black holes merge during the process then the statement is that the final area is greater than the sum of areas of the initial black holes). Note that the increase in area of black holes in all physical processes had already been noted, in the context of specific types of black hole solutions, well before Hawking presented his theorem, notably by Floyd and Penrose and by Christodoulou in the context of Kerr black holes. In fact Floyd-Penrose suggested this is a general feature of black holes, and this is what Hawking set out to prove. The understanding of geodesics initiated by Raychaudhuri, and the subsequent study of singularities by Penrose and Hawking, had set the stage for this to be done.

To prove Hawking's theorem, we need to discuss the black hole horizon in a little more detail. In our discussion of the Schwarzschild solution we saw that the horizon is a sphere translated along the null direction v. In the general case too, it can be argued that future-directed null geodesics starting on the horizon remain on it forever towards the future. To see this, consider a sequence of points $P_1, P_2, \cdots$ just outside the horizon H that has a limit correspond to a point $P \in H$. From each of these points there is a prompt null geodesic to the world-line I of an observer in the far future. However the

limit of these world-lines, which starts from P, cannot reach I, since H is part of B and does not have any points in $J^-(I)$. This means that the limiting world-line either stays on H or enters B. But it cannot enter B since it is the limit of world-lines all of which are outside B. Thus we have proved that one of the two families of null geodesics from the horizon always stay on the horizon. These are called "horizon generators".



Thus horizon generators do not have a focal point. This can be understood by continuity – outside B, the focal point of this family of geodesics is to the past (since they are outgoing), while inside B it is to the future (since they are now ingoing). So right on the boundary H, they focus neither to the past nor the future [38]. Thus each horizon generator just continues along the horizon forever towards the future.

Now let us take two Cauchy hypersurfaces Σ_1, Σ_2 with the latter in the future of the former. Consider $S_i = \Sigma_i \cap H$ and let us examine the null expansion $\theta = \frac{\partial_v A}{A}$ for horizon generators going out normally from S_1 towards the future along the v direction. From the (null) Raychaudhuri equation we see that if θ were negative anywhere on S_1, there would inevitably be a focal point in the future. Since we know this is not so, we must have $\theta \geq 0$ everywhere on S_1. Now take all future-directed horizon generators from S_1 and follow them until they intersect S_2 (this must happen – each of them intersects Σ_2 since that is a Cauchy hypersurface, and also lies in H, so it intersects $S_2 = \Sigma_2 \cap H$). But as we go from Σ_1 to Σ_2, the fact that $\theta \geq 0$ means that the area of the image of the horizon generators in S_2 is greater than or equal to the area of S_1 (we allow the image to not be all of S_2 to allow for the possibility of new black holes forming in the intervening time). This then proves that:

$$\text{Area}(S_2) \geq \text{Area}(S_1) \tag{4.11}$$

[38] Of course in the past, this applies only if the horizon was always present – if the black hole formed from a collapsing shell then before that time there was no black hole and the discussion does not apply.

which is Hawking's area increase theorem. This proof relies on the fact that horizon generates are a "complete set", an alternate without this assumption is sketched in [17]. A more formal proof is in [16] Theorem 12.2.6.

4.5 Emergence of black hole thermodynamics

In the preceding notes we have summarised a number of developments in a very intense period in the study of General Relativity, from Raychaudhuri's seminal paper to the Hawking area increase theorem. The focusing of geodesics still informs much of our thinking about time evolution in gravity, while the singularity theorems to which it led have posed challenges for a better, quantum, understanding of gravity that are still being addressed today. However it is the area theorem that soon led to an overhaul of physics itself. In this section I will try to briefly summarise the ideas.

Once Hawking showed that the area of a black hole always increases, the next significant step was taken by Bekenstein in a 1973 paper called "Black Holes and Entropy", just a year after he completed his Ph.D. He noted that this area increase parallels the behaviour of entropy in thermodynamics, and took a bold step by declaring: "In this paper, we attempt a unification of black hole physics and thermodynamics". He then went on to propose that the area of a black hole is proportional to its entropy: $S \propto A$, and studied this relation from a physical point of view as well as in examples involving specific black holes. Much of his reasoning was based on the Kerr-Newman black hole.

Concretely, Bekenstein attempted to, in his own words, "construct the black-hole analogue of the thermodynamic expression":

$$dE = TdS - PdV \tag{4.12}$$

He did so using the Kerr-Newman black hole of mass M, angular momentum $\vec{J}$ and charge Q. This is the most general stationary [39] black hole solution. Let us first list, without derivation, some properties of this solution (they can be easily found in standard textbooks). It has two horizons, an "inner" and "outer" horizon, whose radial locations are given by:

$$r_{\pm}(M, Q, J) = GM \pm \sqrt{G^2M^2 - GQ^2 - \frac{\vec{J}^2}{M^2}} \tag{4.13}$$

[39] A space-time is said to be stationary if it has a time-like Killing vector field.

The area of this black hole is:

$$A(M,Q,J) = 4\pi\left(r_+^2 + \frac{\vec{J}^2}{M^2}\right) = 4\pi(2GMr_+ - GQ^2) \tag{4.14}$$

In the limit $J, Q \to 0$ these quantities reduce to $r_+ = 2GM, r_- = 0$ and $A = 16\pi G^2 M^2$ which are the familiar results for the Schwarzschild black hole.

Since this black hole is charged, it has an electrostatic potential at the horizon that is found to be:

$$\Phi = \frac{Qr_+}{r_+^2 + \frac{\vec{J}^2}{M^2}} \tag{4.15}$$

In the limit $\vec{J} \to 0$ this reduces to the familiar electrostatic potential $\frac{Q}{r_+}$ on the surface of a non-rotating body of radius r_+. Also since the black hole is rotating, it will have an angular velocity at the horizon which turns out to be:

$$\vec{\Omega} \equiv= \frac{\frac{\vec{J}}{M}}{r_+^2 + \frac{\vec{J}^2}{M^2}} \tag{4.16}$$

Finally, there is a useful quantity called the "surface gravity" that is defined as the acceleration in the frame of an observer at infinity, that is needed to keep a body suspended just above the horizon. It is constant all over the horizon, and is given by:

$$\kappa = \frac{2\pi(r_+ - r_-)}{GA} \tag{4.17}$$

Now we can derive a relation between the variation of the area under a change of mass, charge and angular momentum. Take an infinitesimal variation of Eq.(4.14) to get:

$$dA = 8\pi G\left(r_+ dM + M dr_+ - Q dQ\right) \tag{4.18}$$

Using:

$$\begin{aligned} dr_+ &= G\,dM + \frac{G^2 M\,dM - GQ\,dQ - \frac{\vec{J}\cdot d\vec{J}}{M^2} + \frac{\vec{J}^2 dM}{M^3}}{\sqrt{G^2M^2 - GQ^2 - \frac{\vec{J}^2}{M^2}}} \\ &= \frac{2\left(Gr_+ + \frac{\vec{J}^2}{M^3}\right)dM - 2GQ\,dQ - 2\frac{\vec{J}\cdot d\vec{J}}{M^2}}{r_+ - r_-} \end{aligned} \tag{4.19}$$

Collecting the coefficient of dM on the RHS of Eq.(4.18), we have:

$$\begin{aligned}\text{coeff of } dM &= 8\pi G\left(r_+ + \frac{2GMr_+ + 2\frac{\vec{J}^2}{M^2}}{r_+ - r_-}\right) \\ &= 8\pi G\left(\frac{r_+^2 + (2GM - r_-)r_+ + 2\frac{\vec{J}^2}{M^2}}{r_+ - r_-}\right) \\ &= \frac{4GA}{r_+ - r_-}\end{aligned} \tag{4.20}$$

Next, the coefficient of dQ on the RHS of Eq.(4.18) is:

$$\begin{aligned}\text{coeff of } dQ &= -8\pi GQ\left(1 + \frac{2GM}{r_+ - r_-}\right) \\ &= -16\pi GQ\,\frac{r_+}{r_+ - r_-}\end{aligned} \tag{4.21}$$

Finally, the coefficient of $d\vec{J}$ on the RHS of Eq.(4.18) is:

$$\text{coeff of } d\vec{J} = -\frac{16\pi G}{r_+ - r_-}\frac{\vec{J}\cdot\, d\vec{J}}{M} \tag{4.22}$$

Then, defining:

$$\Theta \equiv \frac{r_+ - r_-}{4GA} \tag{4.23}$$

and recalling Eqs.(4.15),(4.16), we get:

$$dM = \Theta\, dA + \Phi\, dQ + \vec{\Omega}\cdot d\vec{J} \tag{4.24}$$

After deriving this formula [40], Bekenstein interpreted the last two terms as the work done on the black hole to increase its charge and angular momentum. Thus, he argued, they should be thought of as the analogues of $-PdV$ in the thermodynamic relation Eq.(4.12). Then the first term should play the role of TdS, with S being proportional to the area of the black hole. He noted that the parameter Θ playing the role of T is non-negative, just like temperature.

Just over two months later, Bardeen, Carter and Hawking (BCH) submitted a paper on what they called the "four laws of black hole mechanics". They considered

[40] Note that Bekenstein works in terms of the "rationalised area" $\alpha = A/4\pi$ so the expressions here differ slightly from those in his paper.

a more general class of stationary axisymmetric black holes than the Kerr-Newman vacuum solution, namely those with "rings of matter" outside the horizon. Then they essentially derived the same thermodynamic relation as Bekenstein (they do not refer to his paper, since their work was probably concurrent, but they do refer to his Ph.D. thesis). There was one significant new feature: they realised that the coefficient Θ of dA in the thermodynamic relation is related to the surface gravity of the black hole:

$$\Theta = \frac{\kappa}{8\pi} \tag{4.25}$$

which can be verified by comparing Eq.(4.23) and Eq.(4.17).

In their papers, both Bekenstein and BCH then went beyond this "first law"-like relation to discuss other analogies with regular thermodynamics. BCH systematically discussed the analogy for the zeroth, second and third laws of thermodynamics. Some of this was quite straightforward – the zeroth law says temperature is constant in equilibrium, which is also true of surface gravity at the horizon (and of the explicit expression that Bekenstein found). The second law says entropy increases, which is Hawking's theorem. The third law is more non-trivial – it says the temperature of a black hole cannot be reduced to absolute zero by a finite sequence of operations, and BCH argued this is also true of surface gravity for Kerr-Newman black holes.

Returning to Bekenstein, he noted that there are two obvious problems with the formula: (i) entropy and area have different dimensions, (ii) any constant in front of A can be absorbed into the definition of Θ so we cannnot unambiguously fix the constant factor relating area to entropy. He tried to fix these problems in the same paper, obtaining (in units where $c = 1$):

$$\frac{S}{k_B} = \frac{\ln 2}{8\pi}\frac{c^3 A}{G\hbar} \tag{4.26}$$

As it turns out, he fixed the dimensions correctly. This was possible because only one combination of $c, \hbar, G$ has the right dimensions to cancel out the dimensions of A so that it can match S/k_B, as Bekenstein noted – possibly for the first time. However it raised another question: why should $\hbar$ be present given that all our considerations so far are classical? In his paper, Bekenstein says "in desperation we appeal to quantum physics" and then justifies it by saying that $\hbar$ also appears in the expression for the entropy of other thermodynamic systems that are treated as classical.

Apart from the dimensional factors, Bekenstein tentatively proposed a result for the

constant $\frac{\ln 2}{8\pi}$ by a thought experiment involving the loss of one "bit" of information into the black hole [41]. This constant has since then turned out to be incorrect. Meanwhile BCH simply chose the constant to be 1 and quoted their "temperature" as $\frac{\kappa}{8\pi}$. Both sets of authors agreed that this should not be thought of as a true temperature, since the temperature of a black hole was "obviously" zero. Yet, towards the end of his paper, Bekenstein reveals his deep belief that his relation to thermodynamics is not a mere analogy: "The common entropy in the black-hole exterior plus the black-hole entropy never decreases. This statement means that we must regard black-hole entropy as a genuine contribution to the entropy content of the universe."

After these papers were written, there remained two glaring open questions: (i) what exactly was the coefficient of A (apart from the dimensional factors)? This in turn would determine the precise value of the "temperature" variable in the thermodynamic equation. (ii) what was $\hbar$ doing? Could there be a quantum-mechanical source of actual temperature?

It turned out that neither of these questions could possibly be answered in the domain of classical physics, and both were resolved by Hawking in his seminal 1975 paper "Particle creation by black holes". Hawking found the mechanism by which a black hole actually acquires a temperature, describing it in physical terms as the process of quantum pair-creation just outside the horizon. A computation of this effect enabled him to argue that black holes emit thermal radiation at a temperature $\frac{\kappa}{2\pi}$ where κ is their surface gravity. This is 4 times what it was thought to be in BCH, and hence the constant in the relation of entropy to area is $\frac{1}{4}$. The correct formula is then:

$$S = \frac{c^3 A}{4G\hbar} \tag{4.27}$$

This is certainly one of the most beautiful formulae in all of physics, unifying the three fundamental constants $c, \hbar$ and G in a fundamental relation. It also marks the beginning of the quantum era in the study of black holes, which is an appropriate place to conclude these notes.

[41] For much of the paper, Bekenstein tries to interpret his thermodynamic analogy in the language of classical information theory. Again, this was a futuristic idea that had far-reaching consequences when it was implemented with quantum, rather than classical, information.

Acknowledgements

I am grateful to Rajesh Gopakumar for his encouragement and support to deliver these lectures as an activity of ICTS, Bengaluru. The students and staff of ICTS, including ..., were most helpful in organising the video sessions and recordings. I thank Johnny Gleeson, ... for their comments that helped improve the manuscript.

A Notation and conventions

The metric of space-time is taken to be $(-,+,+,+)$.

Metric: $g_{\mu\nu}(t,\vec{x}), \quad \mu,\nu \in \{0,1,\cdots,D-1\}$.

Determinant: $g(t,\vec{x}) \equiv -\det g_{\mu\nu}(t,\vec{x}) = \big|\det g_{\mu\nu}(t,\vec{x})\big|$.

Often the arguments $(t,\vec{x})$ will be collectively denoted by x, or else suppressed altogether.

Christoffel symbol:

$$\Gamma^{\alpha}_{\mu\nu} \equiv \frac{1}{2} g^{\alpha\beta}(g_{\beta\mu,\nu} + g_{\beta\nu,\mu} - g_{\mu\nu,\beta}) \tag{A.1}$$

Riemann curvature tensor:

$$R^{\alpha}{}_{\lambda\mu\nu} \equiv \partial_\mu \Gamma^{\alpha}_{\nu\lambda} - \partial_\nu \Gamma^{\alpha}_{\mu\lambda} + \Gamma^{\alpha}_{\mu\beta}\Gamma^{\beta}_{\nu\lambda} - \Gamma^{\alpha}_{\nu\beta}\Gamma^{\beta}_{\mu\lambda} \tag{A.2}$$

Ricci tensor:

$$R_{\mu\nu} \equiv R^{\alpha}{}_{\mu\alpha\nu} \tag{A.3}$$

Ricci scalar:

$$R \equiv g^{\mu\nu} R_{\mu\nu} \tag{A.4}$$

Energy-momentum tensor:

$$T_{\mu\nu} \equiv -2\frac{1}{\sqrt{g}} \frac{\delta S_{\text{matter}}}{\delta g_{\mu\nu}} \tag{A.5}$$

Einstein-Hilbert action:

$$S = \frac{1}{16\pi G}\int d^D x\,\sqrt{g}(R - 2\Lambda) \tag{A.6}$$

Einstein equations:

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R + \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu} \tag{A.7}$$

Geodesics: A geodesic is a path whose tangent vector propagates parallel to itself. In Riemannian geometry, a geodesic is the path that minimises [42] the distance S between two points of space:

$$S = \int_{\lambda_1}^{\lambda_2} d\lambda \sqrt{g_{ij}(x)dx^i dx^j} \tag{A.8}$$

where the path is described by $x^i = x^i(\lambda)$ and $x_1^i = x^i(\lambda_1), x_2^i = x^i(\lambda_2)$ are the chosen end-points.

The tangent vector of a curve in Riemannian geometry is written $t^i = \frac{dx^i}{d\lambda}$. The statement that this tangent vector propagates parallel to itself is tells us that its directional derivative $D_\lambda \equiv t^m D_m$ along itself is again parallel to it:

$$t^m D_m t^i = \alpha t^i \tag{A.9}$$

We have:

$$t^m D_m t^i = D_\lambda \frac{dx^i}{d\lambda} = \frac{d^2x^i}{d\lambda^2} + \Gamma^i_{jk}\frac{dx^j}{d\lambda}\frac{dx^k}{d\lambda} \tag{A.10}$$

so Eq.(A.9) can be re-written:

$$\frac{d^2x^i}{d\lambda^2} + \Gamma^i_{jk}\frac{dx^j}{d\lambda}\frac{dx^k}{d\lambda} = \alpha\frac{dx^i}{d\lambda} \tag{A.11}$$

Given a path $x^i(\lambda)$ satisfying the above equation, we can always find a reparametrisation $\lambda \to \lambda'(\lambda)$ of the parameter that sets the coefficient α on the RHS of Eqs. (A.9, A.11) to 0. In other words, in the new parametrisation (and after dropping the prime on λ')

[42] This means it corresponds to a local minimum of the distance, not that it is necessarily the shortest distance of all.

we have:

$$D_\lambda t^i = t^m D_m t^i = \frac{d^2x^i}{d\lambda^2} + \Gamma^i_{jk}\frac{dx^j}{d\lambda}\frac{dx^k}{d\lambda} = 0 \tag{A.12}$$

The parameter λ in which this equation is satisfied is called an *affine parameter* and the geodesic said to be affinely parametrised. In this case the norm $t_i t^i = g_{ij}t^i t^j$ of the tangent vector is preserved along the path, since:

$$\frac{d}{d\lambda}(g_{ij}t^i t^j) = 2g_{ij}t^i D_\lambda t^j = 0 \tag{A.13}$$

One can show that given an affine parameter λ, all other possible affine parameters have the form:

$$\lambda' = a\lambda + b \tag{A.14}$$

for constants a, b. One useful fact is that the invariant distance $\lambda = \int \sqrt{g_{ij}(x)dx^i dx^j}$ corresponds to a choice of affine parameter.

In the Lorentzian case, similar considerations apply but with some new features. The equations analogous to Eqs. (A.9) and (A.11), which describe non-affinely parametrised geodesics, are:

$$t^\kappa D_\kappa t^\mu = \frac{d^2x^\mu}{d\lambda^2} + \Gamma^\mu_{\nu\lambda}\frac{dx^\nu}{d\lambda}\frac{dx^\lambda}{d\lambda} = \alpha t^\mu \tag{A.15}$$

For the affine case, the same equations hold but with $\alpha = 0$. Now there are three types of geodesics in Lorentzian signature:

$$\begin{aligned} g_{\mu\nu}\frac{dx^\mu}{d\lambda}\frac{dx^\nu}{d\lambda} &> 0: \quad \text{space-like} \\ g_{\mu\nu}\frac{dx^\mu}{d\lambda}\frac{dx^\nu}{d\lambda} &< 0: \quad \text{time-like} \\ g_{\mu\nu}\frac{dx^\mu}{d\lambda}\frac{dx^\nu}{d\lambda} &= 0: \quad \text{null} \end{aligned} \tag{A.16}$$

Correspondingly we have a *proper distance* between space-like-separated points, a *proper time* between time-like-separated points, and zero invariant distance between null-separated points. As we have seen, for an affinely parametrised geodesic the norm of the tangent vector is constant and therefore any geodesic must be space-like, time-like or null throughout its entire trajectory.

In practice we will mainly be interested in the time-like and null cases, which are the ones along with signals can propagate. For time-like-separated space-time points,

a geodesic extremises the proper time between them:

$$S = \int_{\lambda_1}^{\lambda_2} d\lambda \sqrt{-g_{\mu\nu}(x) dx^\mu dx^\nu} \tag{A.17}$$

where the space-time path is described by $x^\mu = x^\mu(\lambda)$ and $x_1^\mu = x^\mu(\lambda_1), x_2^\mu = x^\mu(\lambda_2)$ are the chosen end-points. But for the null case we simply have $S = 0$.

The relevant new features are: (i) A time-like geodesic actually *maximises* the proper time between points. This is shown in Section 3.2. (ii) The affine parameter for a null geodesic cannot be identified with the proper interval, since the latter is always zero. Thus, in a sense, there is no "standard" affine parameter λ (or, more generally, linear family of affine parameters $a\lambda + b$) in the null case. (iii) We can normalise the tangent vector of a time-like geodesic to unity: $g_{\mu\nu} t^\mu t^\nu = 1$ (and we typically do so). However the tangent vector of a null geodesic satisfies $g_{\mu\nu} t^\mu t^\nu = 0$, which is preserved under any constant rescaling of t^μ [43]. Hence there is no "standard normalisation" in the null case.

B Useful identities

The following is a useful identity for the infinitesimal variation of the determinant of a $d \times d$ invertible matrix M_{ij}, where we denote $\det M_{ij}$ by $|M|$:

$$\begin{aligned} \delta|M| &= \delta \exp\left(\log|M|\right) \\ &= \delta \exp\left(\,\mathrm{tr}\,\log M\right) \\ &= \left(\,\mathrm{tr}\, M^{-1}\delta M\right) \exp\left(\,\mathrm{tr}\,\log M\right) \\ &= \left(\,\mathrm{tr}\, M^{-1}\delta M\right)|M| \end{aligned} \tag{B.1}$$

Hence,

$$|M|^{-1}\delta|M| = M^{-1}_{ij}\delta M_{ji} \tag{B.2}$$

This can be applied to calculate derivatives, replacing δ by $\frac{d}{dx}$ where x is some variable on which the matrix depends.

[43] The rescaling factor only needs to be constant along each geodesic, but can vary from one geodesic to another.

Another helpful identity, that is trivial to derive, is:

$$\mathrm{tr}\left(M - \frac{1}{d}(\mathrm{tr}\, M)\right)^2 = \mathrm{tr}\,(M^2) - \frac{1}{d}(\mathrm{tr}\, M)^2 \tag{B.3}$$

C Compactness

Here we summarise the relevant definitions of various types of compactness.

Definition C.1. *A subset S of a topological space is called* ***compact*** *if every open cover admits a finite sub-cover.*

Here an open cover means a union of open sets U_n that completely contains S, $(\cup_n U_n) \supset S$. Such a union can be over a finite cover or an infinite cover, depending on whether n takes finitely many or infinitely many values. The theorem then says that S is compact if and only if *for every* infinite cover $\{U_n\}$, there is a finite subset of values of n for which the corresponding U_n are also a cover. Note that the definition is *not* equivalent to saying that we can find *some* finite cover for S!

As an example, take $\mathbf{R}^2$ with the usual topology, and let S be an open disc centred at the origin: the set of points $|\vec{x}| < 1$. We can easily find a finite cover for S – in fact a single open disc of radius > 1 centred at the origin covers it completely. But this tells us nothing about compactness of S. We now consider the infinite cover $\cup_{n=2}^{\infty} D_n$ where D_n is the open disc centred at the origin with radius $R_n = 1 - \frac{1}{n}$. Clearly this infinite union contains every point of S, and hence can be called a cover. For, if we take an arbitrary point lying at a distance $r < 1$ from the origin, it will lie inside all discs D_n with $n > \frac{1}{1-r}$ and hence in their union. And yet, there is no finite subcover of this particular cover that contains the whole of S. This proves that S is not compact.

If we replace S by its closure $\bar{S}$, namely the points $|\vec{x}| \leq 1$, then things are different. The infinite cover we defined above does not contain any point with $|\vec{x}| = 1$, so it is not a valid cover of $\bar{S}$. Indeed, $\bar{S}$ is compact – though to prove it, we really need to show that *every* open cover has a finite sub-cover. Proofs exist in the literature, for example [13]. However it is not easy to apply this definition to non-trivial spaces, including the space of causal paths that is of interest here. Fortunately there are other notions of compactness that are easier to verify, and to which we now turn.

For any subset of $\mathbf{R}^n$, and more generally for any metric space, compactness as defined above is equivalent to being *closed and bounded* (Heine-Borel theorem). This is

how we can immediately see that, for example, the open disc in $\mathbf{R}^2$ is not compact (it's bounded but not closed), and the entire real line in $\mathbf{R}^2$ is not compact (it's closed but not bounded), but the closed disc in $\mathbf{R}^2$ is compact (it's both closed and bounded).

Unfortunately some of the spaces of interest in these notes are not metric spaces, so we have to rely on another approach to compactness.

Definition C.2. *A set S is **sequentially compact** if every infinite sequence in S has a sub-sequence that converges to a point in S.*

In general, this definition is not equivalent to compactness [44]. However, if a topological space is *second-countable* and sequentially compact, then it is compact. A second-countable space is one which has a countable basis of open sets – one that can be labelled by the set of natural numbers. This is the converse to the Bolzano-Weierstrass theorem (as stated in [16], Theorem A.9 in Appendix A [45]). Thus to prove compactness of a space, it is sufficient to show that it is second-countable as well as sequentially compact.

References

[1] A. Einstein, *Zur Allgemeinen Relativitätstheorie*, *Sitzungsber. Preuss. Akad. Wiss. Berlin (Math. Phys.)* **1915** (1915) 778.

[2] M. Saha and S. Bose, *The principle of relativity: Original papers by a. einstein and h. minkowski, Translated into English), University of Calcutta Publication* (1920) .

[3] A. Raychaudhuri, *Relativistic cosmology. 1.*, *Phys. Rev.* **98** (1955) 1123.

[4] R. Penrose, *Gravitational collapse and space-time singularities*, *Phys. Rev. Lett.* **14** (1965) 57.

[5] S. Hawking, *Occurrence of singularities in open universes*, *Phys. Rev. Lett.* **15** (1965) 689.

[44] The strategy of [17] is to first prove sequential compactness of C_P^Q for Minkowski space-time, starting from the self-evident fact that D_P^Q is compact in Minkowski space-time (it is just a diamond-shaped region including its boundary). Then it is argued that the same proof can be extended to general M as long as D_P^Q is compact, and this leads to the condition of global hyperbolicity – which implies compactness of D_P^Q and allows us to recycle the proof of compactness of C_P^Q that works for Minkowski space-time. However here, following [16], we first take M to be globally hyperbolic and then prove sequential compactness of C_P^Q and finally of D_P^Q.

[45] The original Bolzano-Weierstrass was in the context of real numbers, but the version in [16] refers to general topological spaces.

[6] S. W. Hawking, *Singularities in the universe*, *Phys. Rev. Lett.* **17** (1966) 444.

[7] S. Hawking, *The occurrence of singularities in cosmology*, *Proc. Roy. Soc. Lond. A* **294** (1966) 511.

[8] S. Hawking, *The Occurrence of singularities in cosmology. II*, *Proc. Roy. Soc. Lond. A* **295** (1966) 490.

[9] S. Hawking, *The occurrence of singularities in cosmology. III. Causality and singularities*, *Proc. Roy. Soc. Lond. A* **300** (1967) 187.

[10] S. W. Hawking and R. Penrose, *The Singularities of gravitational collapse and cosmology*, *Proc. Roy. Soc. Lond. A* **314** (1970) 529.

[11] J. B. Hartle, *Gravity: an introduction to Einstein's general relativity*. American Association of Physics Teachers, 2003.

[12] S. M. Carroll, *Spacetime and geometry*. Cambridge University Press, 2019.

[13] I. M. Singer and J. A. Thorpe, *Lecture notes on elementary topology and geometry*. Springer, 2015.

[14] N. Mukunda and S. Mukhi, *Lectures on advanced mathematical methods for physicists*. World Scientific, 2010.

[15] R. Geroch, *Domain of dependence*, *Journal of Mathematical Physics* **11** (1970) 437.

[16] R. M. Wald, *General Relativity*. Chicago Univ. Pr., Chicago, USA, 1984, 10.7208/chicago/9780226870373.001.0001.

[17] E. Witten, *Light rays, singularities, and all that*, *Rev. Mod. Phys.* **92** (2020, arXiv: 1901.03928) 045004 [1901.03928].

[18] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-Time*, Cambridge Monographs on Mathematical Physics. Cambridge University Press, 2, 2023, 10.1017/9781009253161.

[19] N. J. Hicks, *Notes on differential geometry*, vol. 1. van Nostrand Princeton, 1965.

[20] J. Leray, "Hyperbolic Differential Equations." 1952.

[21] R. Penrose, *Techniques in differential topology in relativity*. SIAM, 1972.

[22] A. Komar, *Necessity of singularities in the solution of the field equations of general relativity*, vol. 104. APS, 1956.

[23] H. P. Robertson, *Relativistic cosmology*, *Reviews of modern Physics* **5** (1933) 62.

[24] E. M. Lifshitz and I. M. Khalatnikov, *Investigations in relativistic cosmology*, vol. 12. 1963, 10.1080/00018736300101283.

[25] R. K. Sachs, *On the Characteristic Initial Value Problem in Gravitational Theory*, *J. Math. Phys.* **3** (1962) 908.