

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

Sunil Mukhi

April 28, 2023

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Causality</b>	<b>5</b>
2.1	First look at causality . . . . .	5
2.2	Globally hyperbolic space-times . . . . .	10
2.3	Properties of globally hyperbolic space-times . . . . .	16
2.4	Compactness of the space of paths . . . . .	18
<b>3</b>	<b>Geodesics and focal points</b>	<b>24</b>
3.1	Focal points and path shortening: Euclidean case . . . . .	25
3.2	Focal points and path lengthening: Lorentzian case . . . . .	27
3.3	The Raychaudhuri equation . . . . .	30
3.4	Timelike geodesics and Hawking theorem . . . . .	37
3.5	Null geodesics . . . . .	40
3.6	The null Raychaudhuri (Sachs) equation . . . . .	48
3.7	Trapped surfaces and Penrose's singularity theorem . . . . .	53

<b>4</b>	<b>Black holes</b>	<b>56</b>
4.1	The Schwarzschild solution . . . . .	56
4.2	Cosmic censorship . . . . .	60
4.3	Generic black holes . . . . .	61
4.4	Hawking’s area theorem . . . . .	64
4.5	Emergence of black hole thermodynamics . . . . .	66
	<b>Appendices</b>	<b>70</b>
<b>A</b>	<b>References, notation, conventions, identities</b>	<b>70</b>
A.1	References . . . . .	70
A.2	Notation and conventions . . . . .	71
A.3	Useful identities . . . . .	72

# 1 Introduction

The General Theory of Relativity is one of the most stunning 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 the 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:

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

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

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

with appropriate boundary conditions and/or boundary terms (which we will discuss later as needed). All the relevant notation and conventions can be found in Appendix A.

Despite the conceptual simplicity, unpacking these equations for their mathematical and physical content is extremely complicated. Broadly there are two reasons for this, one math-

emational and the other physical. On the mathematical side, Eq.(3.7) is a set of second-order partial differential equations for 10 independent variables  $g_{\mu\nu}$ . While this makes things tedious, it should hardly be out of reach of modern computing methods. Yet, it is not possible to find the most general solutions to the Einstein equations.

---

*SM: Can one find the most general solution to the Einstein equations?*

*ChatGPT: No, it is not possible to find the most general solution to the Einstein equations.*

*The Einstein equations describe the relationship between the geometry of spacetime and the matter and energy within it. They are highly nonlinear and complex partial differential equations, and it is difficult to find exact solutions to them except in highly symmetric cases.*

*In fact, it is known that there is no general algorithm for finding solutions to the Einstein equations. Even in the absence of matter, the equations are highly nonlinear and difficult to solve.*

*As a result, much of the work in general relativity involves finding approximate solutions or making simplifying assumptions to make the equations more tractable.*

---

To solve differential equations involving space, we need boundary conditions. To solve differential equations in time, we need initial conditions. But 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 out of 10 variables  $g_{\mu\nu}$ , 4 are redundant and can be removed by a *gauge choice*. Another 4 correspond to *constraints* on the initial values and therefore also get removed. That leaves two independent degrees of freedom of the metric, which precisely correspond to the two independent polarisations of a gravitational wave. But the steps leading to the independent degrees of freedom are highly non-unique.

Those were the mathematical problems. On the physical side, one can get space-times to do a lot of strange things, for example time can be a circle so we keep coming back to the past. These are called *unphysical*. But who, or what, decides what is physical and what isn't? In 1926, Arthur Eddington claimed that black holes are unphysical solutions of the Einstein equations. Later, S. Chandrasekhar also doubted they could form in nature. Today we know they are the one of the most common objects in the universe! Today there is a broad consensus that black holes are physical objects, but whether other types of singularities are physical or unphysical is still debated.

---

*SM: How do we decide which solutions of Einstein's equations are physical or unphysical?*

*ChatGPT: In general relativity, we can distinguish physical solutions of the Einstein equations from unphysical solutions by testing whether they satisfy certain physical criteria or assumptions. Here are a few examples of such criteria:*

*Energy conditions: These are inequalities that relate the stress-energy tensor of matter and energy to the curvature of spacetime. For example, the weak energy condition states that the energy density measured by any observer cannot be negative. Violations of these conditions often indicate the presence of unphysical or exotic matter, such as negative mass.*

*Causality conditions: These are conditions that ensure that the propagation of signals in spacetime respects the causal structure of the universe, namely that events can only be influenced by events in their past light-cone. For example, the chronology protection conjecture states that the existence of closed timelike curves, which would allow for time travel, is not allowed in classical general relativity.*

*Symmetry requirements: Some solutions of the Einstein equations may not respect the symmetries of the physical system being modeled. For example, a solution that assumes a static universe may not be physical if the universe is observed to be expanding.*

*Singularity avoidance: Singularities are points in spacetime where the curvature becomes infinite and the laws of physics may break down. Physical solutions of the Einstein equations should avoid singularities, or at least describe them in a way that is consistent with our current understanding of physics.*

*These criteria are not exhaustive, and there may be additional criteria specific to different physical contexts. Ultimately, the physical validity of a solution depends on how well it describes the observed behavior of matter and energy in the universe.*

---

And with that answer, ChatGPT (without much prompting!) has brought us right to the topic of this lecture series. Energy conditions, causality conditions, singularity avoidance and symmetry are themes we will repeatedly encounter. In what follows, we will discuss what the Einstein equations tell us about the geometry and topology of classically allowed spacetimes. This is still way too much material, so a lot of topics will unfortunately have to be excluded.

Before proceeding, let me also 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. Indeed, the problem of quantising gravity is under active investigation. These lectures are intended to help people wishing to participate in this investigation by providing a slightly more detailed understanding of classical gravity than the standard courses. But here, gravity will be treated as classical – and things will be complicated enough despite this limitation.

## 2 Causality

### 2.1 First look at causality

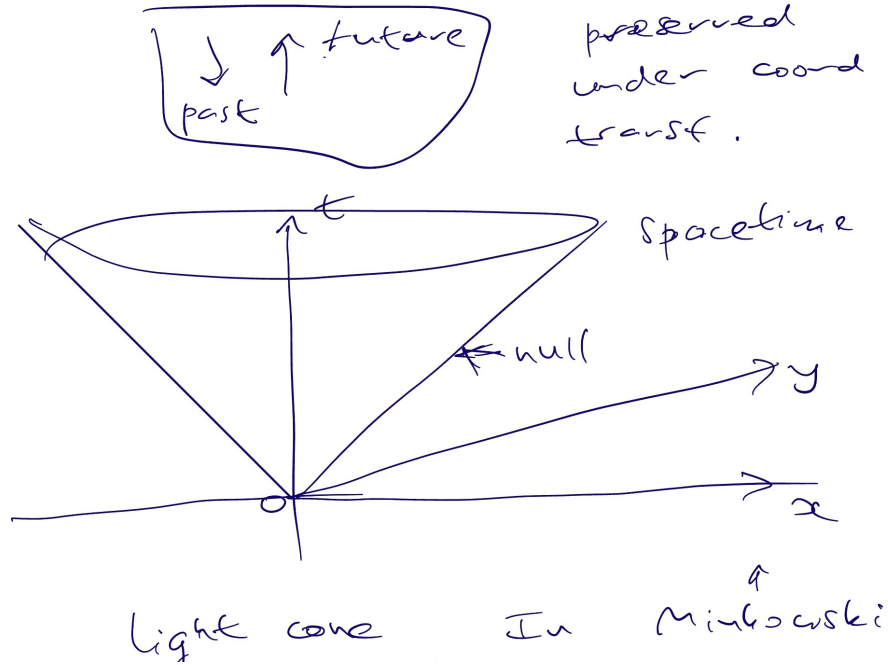
In this section we will study causal properties of generic space-times, as well as space-times satisfying certain properties that we will specify. But we will not assume they are solutions of the Einstein equations.

**Definition 2.1.** A **space-time**  $M$  is a  $D$ -dimensional differentiable manifold (also called  $M$ ) with a smooth metric that, in the coordinate system  $x^\mu$ , is denoted  $g_{\mu\nu}(x)$  of Lorentzian signature.  $M$  is 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<sup>1</sup>.

The key feature of space-times, as compared to spaces of Euclidean signature, is the presence of one coordinate distinct from all the others. We call it *time*. Both in special and general relativity, the special nature of the time coordinate is preserved under coordinate transformations (Lorentz transformations in the former case, and general coordinate transformations in the latter). 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 at  $D = 2$ , and not too hard for  $D = 3$  where we imagine the horizontal direction to be a plane.  $D = 3$  is usually the right situation to try and visualise, as it has the same connectivity properties as higher dimensions. Another approach is to plot the radial direction against time and “imagine” a  $D - 2$ -dimensional sphere at each radial point.

---

<sup>1</sup>Technically  $M$  is also required to be **Hausdorff** and **paracompact**. Details can be found in [Wald], Appendix A.



In a space-time diagram, the meaning of “future” and “past” directions in a small neighbourhood of any given point is given 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 from that point. In Minkowski space-time, in  $D = 2$  this is just a pair of straight lines at  $45^\circ$ , but in  $D = 3$  it is really a cone. In a general space-time  $M$  we should think of the light cone as living not in  $M$  but in the tangent space  $T_p(M)$ .

Now we can begin to discuss causality. Let us consider space-time paths  $x^\mu(s)$ . Here  $s$  is an arbitrary parameter that labels points on the path. Two paths are equivalent if they differ by a non-singular transformation  $\tilde{s}(s)$ . We also identify them with each other if they differ by a reparametrisation of the space-time coordinate,  $x^\mu \rightarrow x'^\mu$ . At this stage we do not assume geodesic paths, so these are not necessarily paths that a freely moving particle can traverse.

**Definition 2.2.** A path <sup>2</sup>  $x^\mu(s)$  is **timelike** if the tangent vector  $\frac{dx^\mu}{ds}$  is everywhere timelike, and **causal** if the tangent vector is everywhere timelike or null:

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

<sup>2</sup>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  $Q$ ”.

In this definition we have used a coordinate system, but obviously the curve is a geometric entity that can be described in any coordinate system. For this reason some authors, e.g. [Wald], denote curves by  $\lambda(s)$  where  $\lambda$  is an abstract point in  $M$  for each  $s$ , which then can be assigned coordinates in any chosen 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  $M$ .

**Definition 2.3.** *The **chronological future**  $I^+(P)$  or **chronological past**  $I^-(P)$  <sup>3</sup> is the set of the points reachable in the future/past by a timelike path starting at  $P$ . 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  $I^-(S)$ .*

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 timelike curve remains timelike under an infinitesimal deformation). A related point is that  $P$  is in its own chronological future/past only if the space-time has closed timelike curves, and not otherwise.

---

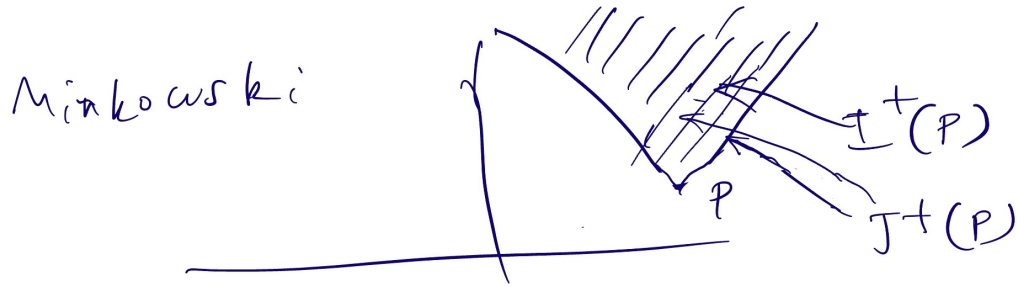
<sup>3</sup>Recall that “chronos” was the Greek god of time.

**Exercise 2.1.** Prove both the statements above.

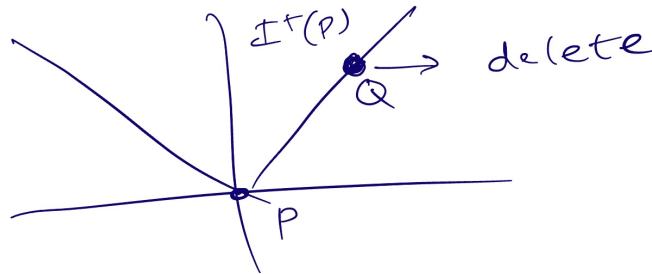
**Definition 2.4.** The **causal future**  $J^+(P)$  or **causal past** of a point  $P$  is the set of points reachable in the future/past by a causal path starting at  $P$ . Again we define the causal future of a set  $S \subset M$  as the union of causal futures of each point in  $S$ :  $J^+(S) \equiv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

Note that this space includes its own boundary – the set of points that can be reached by 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  $\partial I^+(P)$ . One may think that  $I^+(P) \cup \partial I^+(P) = J^+(P)$ , but this need not be true. An example is Minkowski space-time where a point on the future light-cone is deleted. Then  $J^+(P)$  lacks the deleted point, but  $I^+(P) \cup \partial I^+(P)$  includes 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, which we have explained later on in these notes).

We can also define the chronological future of a set  $S$  by:

$$I^+(S) = \cup_{P \in S} I^+(P) \quad (2.2)$$

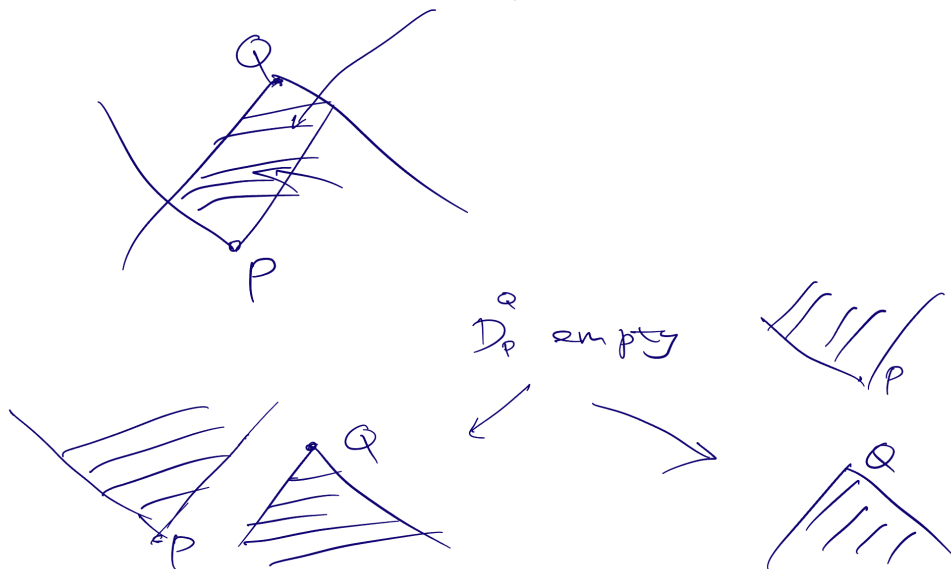
**Exercise 2.4.** Show that (i)  $I^+(S)$  is an open set, (ii)  $I^+(I^+(S)) = I^+(S)$ , (iii)  $I^+(\bar{S}) = I^+(S)$ .

Next we move on to consider two distinct points, say  $P$  and  $Q$ , such that  $Q$  is in the causal future of  $P$ . We will be interested in considering 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  $Q$ .

A causal path from  $P$  to  $Q$  must lie in the causal diamond of these points, (but obviously, not every path that lies in the causal diamond is a causal path).

Causal diamond  $D_P^Q = J^+(P) \cap J^-(Q)$

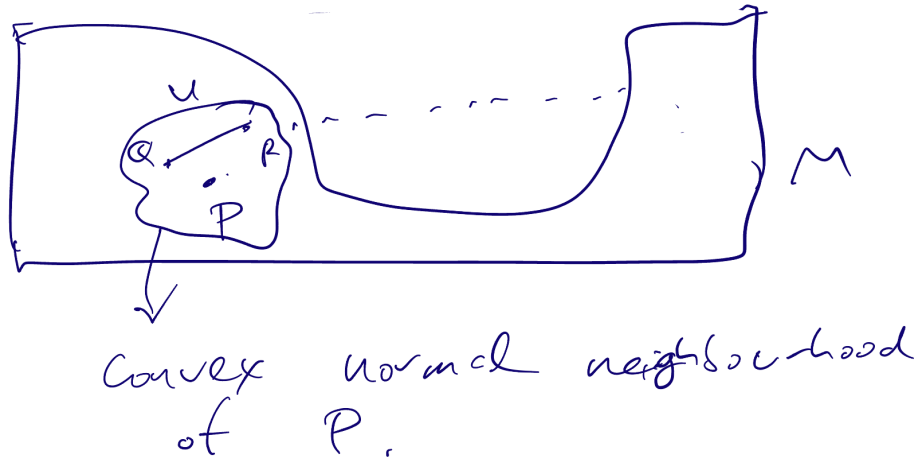


**Exercise 2.5.** Draw paths from  $P$  to  $Q$  that are (i) in the causal diamond, and causal, (ii) in the causal diamond, but not causal, (iii) not in the causal diamond.

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 timelike 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** of  $P$ . 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 timelike or null geodesics from  $P$  and similarly for the causal past (for the proof see [Wald], Theorem 8.1.2).



## 2.2 Globally hyperbolic space-times

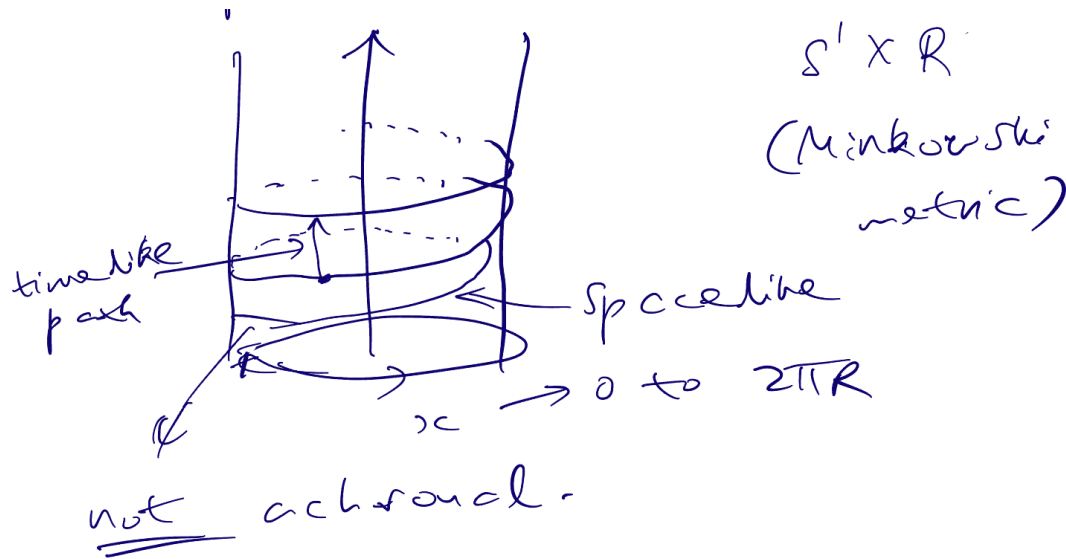
We would like to specify the conditions that a space-time must satisfy for it to be a physically acceptable one. To this end, we start with a few definitions that will help us define physical initial-value hypersurfaces.

First we define subsets of  $M$  in which no two points are separated by a timelike 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  $P, Q \in S$ .

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

Note however that an everywhere locally spacelike or null surface may not be achronal. As an example consider a hypersurface corresponding to a spatial circle rapidly swept out as time goes on. In a space-time diagram this looks like a helix and can easily be taken spacelike. Then there is a timelike path that takes us from a point on the helix at one time to the same point at a future time. Hence the helix, though spacelike, is not achronal.



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

The next concept we need to understand is whether a curve in space-time extends forever or comes 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 endpoint** of this curve if every open neighbourhood  $O \subset M$  around  $P$  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 endpoint is called **future extendible**, while a curve without any endpoint is called **future inextendible**.

Note that the endpoint  $P$  does not need to actually lie on the curve! In other words, there does not have to be a value  $s = s_1$  such that  $x^\mu(s_1)$  is the point  $P$ . The above definition requires the curve to come arbitrarily close to  $P$ , in the topological sense, 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, so 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 endpoint 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 timelike 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 timelike 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  $x^\mu(s)$ , in [Wald] 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 timelike. Because it comes closer and closer to  $x^\mu(s)$  as we go to the past, an endpoint of  $x'^\mu(s)$  would also be an endpoint of  $x^\mu(s)$  but the latter does not exist since we assumed  $x^\mu(s)$  to be past inextendible.

These definitions have 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.** *A **Cauchy hypersurface** in  $M$  is an achronal spacelike<sup>4</sup> surface  $\Sigma$  such that every inextendible causal path through a point  $P \in M, P \notin \Sigma$ , passes through  $\Sigma$ . A space-time  $M$  with a Cauchy hypersurface is said to be **globally hyperbolic**.*

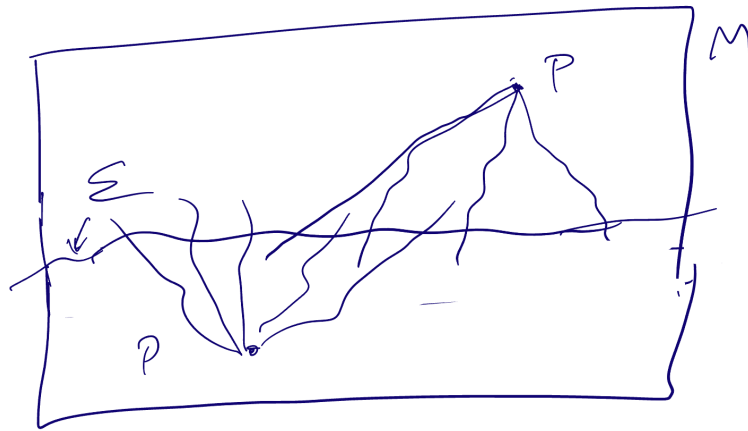
This definition basically says that every point of  $M$  is connected to a Cauchy hypersurface by a causal path. This sets the stage for us to specify initial conditions on the Cauchy hypersurface and be confident that they will determine the system for all future times. Some authors require  $\Sigma$  to be closed, but it can be proved from the above definition that it is closed ([Witten], p. 18) – so it need not be assumed to start with. A Cauchy hypersurface  $\Sigma$  divides the whole space-time into a future and a past. To quote [Wald], “in a globally

---

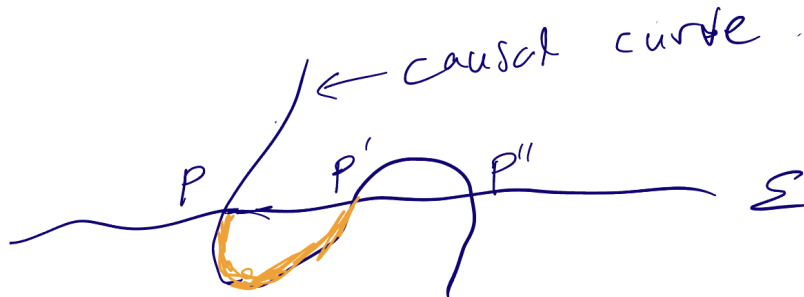
<sup>4</sup>[Witten] requires it to be spacelike, while [Wald] and others do not. For now, we will stick with the spacelike definition and generalise to the null case as needed later on.



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  $\Sigma$ ".



Note that an inextendible timelike 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 timelike path from the Cauchy surface to itself (yellow path in the figure), contradicting achronality.



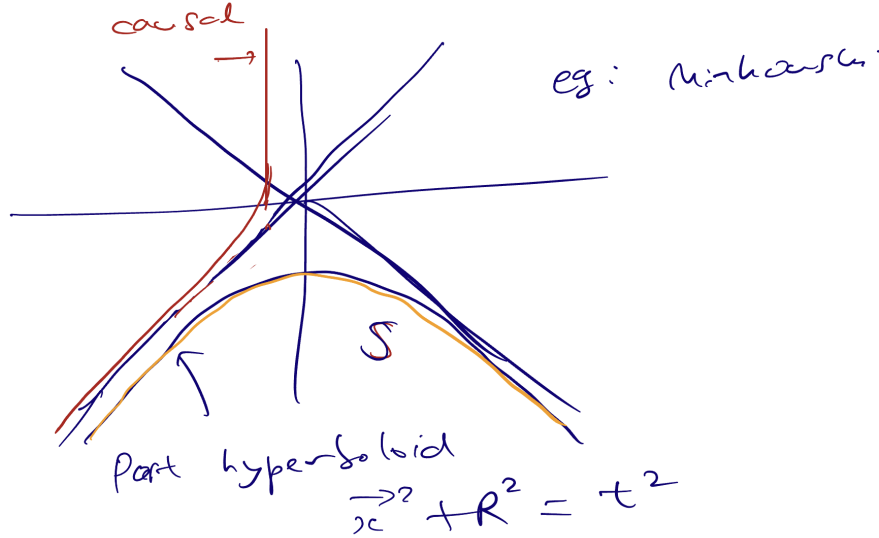
The standard example of a non-globally hyperbolic space-time is Minkowski space-time with one point  $P$  deleted. Take a spacelike surface  $S$  for which  $P$  lies in its causal future. Deletion of  $P$  does not change the fact that other points in its future are in the causal future of  $S$ , since 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 timelike path from  $Q$  that lands on  $P$ . This path cannot be extended to reach  $S$  since it gets “stuck” at  $P$  (as [Witten] puts it). 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.

We will encounter situations where there is an achronal spacelike set that fails to be a Cauchy hypersurface. Let us consider an example of such a hypersurface ([Witten]). Let  $S$  be the past hyperboloid in Minkowski space-time, defined by  $\vec{x}^2 + R^2 = t^2$  (yellow line in

figure). It is everywhere spacelike since:

$$\left| \frac{d\vec{x}}{dt} \right| = \frac{t}{\sqrt{t^2 - R^2}} > 1 \quad (2.3)$$

By drawing it we can see that it is also achronal – there are no timelike curves connecting two points on it. To see that it is not Cauchy, we draw a curve that asymptotes to the time axis in the future and to the light cone in the past (red line in figure).



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  $M$ . This subset is called the domain of dependence.

**Definition 2.9.** The **future domain of dependence**  $D^+(S)$  of an achronal spacelike 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  $D^+(S) \cup D^-(S)$ .

In this language, a Cauchy hypersurface  $\Sigma$  is the special case where  $D(\Sigma) = M$ , i.e. its domain of dependence is the whole space-time.

If  $S$  is achronal and spacelike but not a Cauchy hypersurface, then its domain of dependence  $D(S)$  is clearly not the whole space-time  $M$ . But  $D(S)$  by itself is a globally hyperbolic space-time, essentially by construction. And it is still true that  $S$  divides  $D(S)$  into a future and past.

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

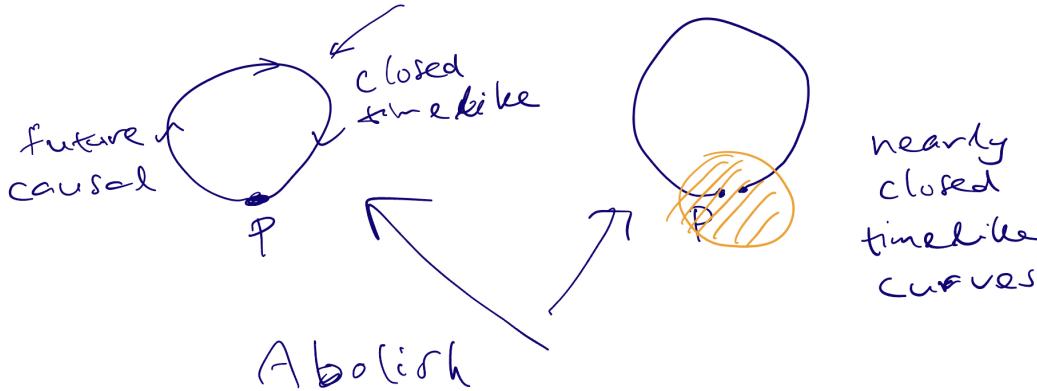
**Theorem 2.4.** *A point  $P$  lies in  $\overline{D^+(S)}$  if and only if every past inextendible timelike curve through  $P$  intersects  $S$ . (The proof is straightforward – see [Wald], Proposition 8.3.2.)*

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

**Definition 2.10.** *The boundary of the closure of the domain of dependence of an achronal spacelike 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 an example 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 timelike curves, so we will take that as a minimal requirement. However we also want to avoid space-times where there are *nearly closed* timelike 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.

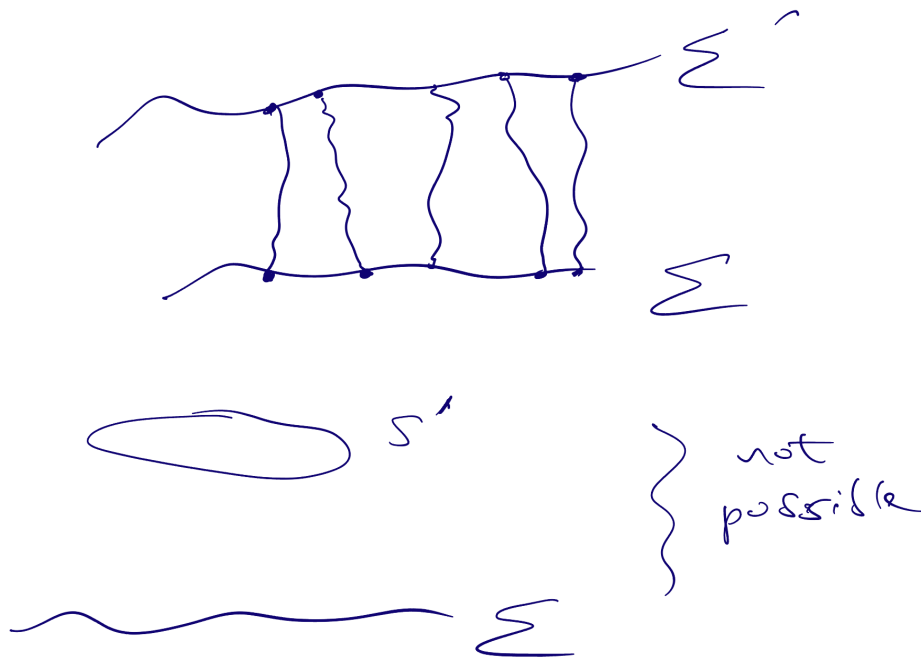


**Definition 2.11.** *A space-time is said to be **strongly causal** if, for every point  $P$  and an arbitrary open neighbourhood  $O$  of  $P$ , there exists a sufficiently small sub-neighbourhood  $V \subset O$  through which no causal curve passes more than once.*

## 2.3 Properties of globally hyperbolic space-times

Globally hyperbolic space-times have several good properties. First, they cannot contain closed timelike paths. This follows from the fact that an inextendible timelike 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.

Another important property is that two different Cauchy hypersurfaces  $\Sigma, \Sigma'$  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, which can be used to set up a smooth 1-1 map between them via timelike curves that are integral curves of a vector field (a more precise proof is in [Witten] Section 3.2). There is also a statement in case  $\Sigma$  is Cauchy but  $S'$  is only achronal but not Cauchy. In this case it can be shown that  $S'$  is topologically equivalent to a portion of  $\Sigma$ . As a corollary, if  $\Sigma$  is noncompact – but connected – then any achronal hypersurface  $S'$  *cannot* be compact.



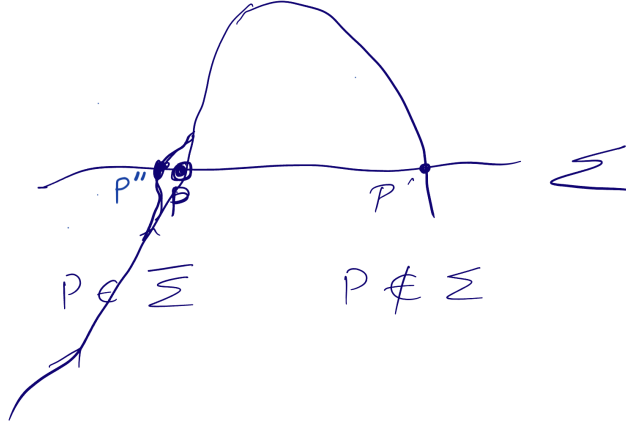
Three more corollaries are:

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

(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 move on Cauchy

hypersurface to another. This type of “translation” sweeps out a line transverse to the Cauchy hypersurface  $\Sigma$ , 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  $\Sigma$  is the same as  $\Sigma \times \{0\}$  where 0 is the location of  $\Sigma$  on the transverse curve  $R$ . Since  $\Sigma$  is closed in itself, and a single point  $\{0\}$  is closed in  $R$ ,  $\Sigma$  is closed in  $\Sigma \times R$ . A more detailed proof ([Witten], page 18) is as follows. Assume the contrary, namely  $\Sigma$  is not closed. Then there exist points  $P$  such that  $P \in \bar{\Sigma}$  but  $P \notin \Sigma$ , i.e.  $P$  is a limit point of  $\Sigma$  that does not lie in  $\Sigma$ . 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  $\Sigma$  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  $\Sigma$  (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.



There is one more point 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-time  $M''$ ,  $M'''$  etc, but at some point the extended space-time will no longer be globally hyperbolic. Thus we are led to the concept of *maximal extension* of a space-time, namely a *maximally extended space-time* 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:

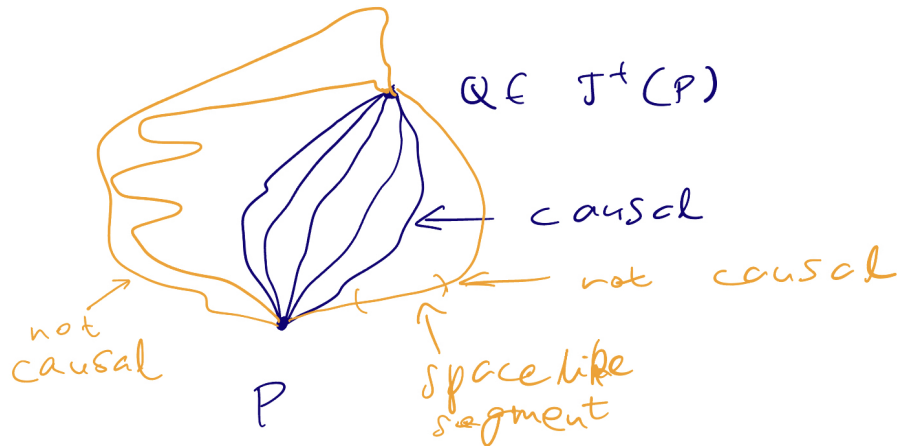
- (D1) The space  $C_P^Q$  of causal paths from  $P$  to  $Q$  is compact for every  $P, Q$  (plus strong causality). [Leray definition]
- (D2) There exists a Cauchy hypersurface through which all inextendible curves pass. [Geroch definition]
- (D3) The causal diamond  $D_P^Q$  is compact for every  $P, Q$  (plus strong causality). [Hawking-Ellis definition]

It turns out these are all equivalent, i.e. any one of them implies the other two. Here, following [Wald] and [Witten], I have used Geroch's definition (D2), so we need to prove (D1) and (D3) as consequences. This is what we will do now. Thus as before, when we say “globally hyperbolic space-time” we will mean a space-time satisfying (D2).

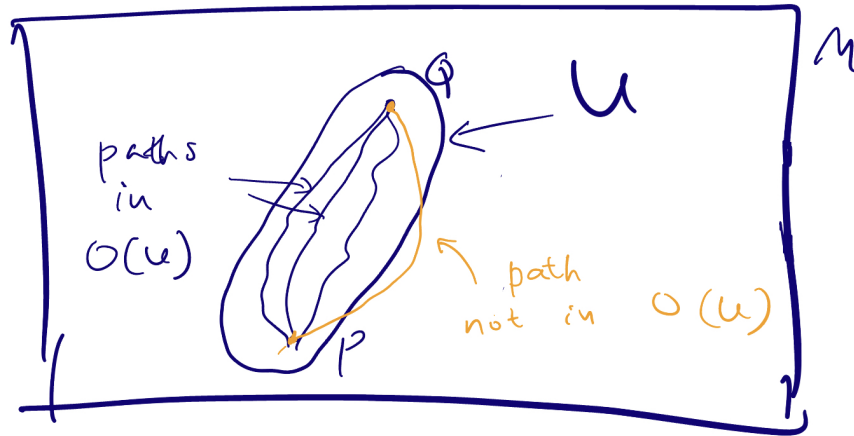
Now consider a globally hyperbolic space-time  $M$  and the set of all continuous <sup>5</sup> causal paths  $x^\mu(s)$  starting at  $P$  and ending at  $Q \in J^+(P)$ . We may choose the parameter  $s$  such that  $x^\mu(0)$  is the point  $P$  and  $x^\mu(1)$  is the point  $Q$ . As always, we identify 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  $C_P^Q$ .

---

<sup>5</sup>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.



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  $U$ . We take the  $O(U)$  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  $Q$  is **compact**.*

Before giving a precise definition of compactness, we mention that it is closely related to convergence of infinite sequences in a space. It is easy to show that sequences of general (non-causal) paths need not converge to a definite limiting path. As an example, consider a sequence whose  $n$ th member is the path  $(x = \sin n\pi t, y = 0, z = 0)$  in Minkowski space-time. These paths connect  $(0, 0, 0)$  to  $(1, 0, 0)$  as  $t$  goes from 0 to 1. However there is no limiting path as  $n \rightarrow \infty$ , instead the paths just oscillate more and more wildly as  $n$  increases. Notice that none of these paths is causal – this is easiest to see if you sketch it. So we have shown

that sequences of non-causal paths can have bad convergence properties. The theorem above says that when we consider the space of *causal* paths, things are quite different.

Now let's examine the definition(s) of compactness in a little detail.

**Definition 2.12.** *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 that completely contains  $S$ , namely  $(\cup_n U_n) \supset S$ . 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  $\cup_{n=1}^{\infty} D_n$  where  $D_n$  is the open disc centred at the origin with radius  $R_n = 1 - \frac{1}{n}$ . This cover contains every point of  $S$ . 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, 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  $\bar{S}$  at all. 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. However it is not clear how 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.

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 not closed), the closed disc in  $\mathbf{R}^2$  is compact (it's both closed and bounded) and the entire real line in  $\mathbf{R}^2$  is not compact (it's closed but not bounded).

Unfortunately  $C_P^Q$  is not a metric space nor a subset of any  $\mathbf{R}^n$  (the space of curves is, at least in a formal sense, infinite-dimensional). So we have to rely on another approach to compactness.

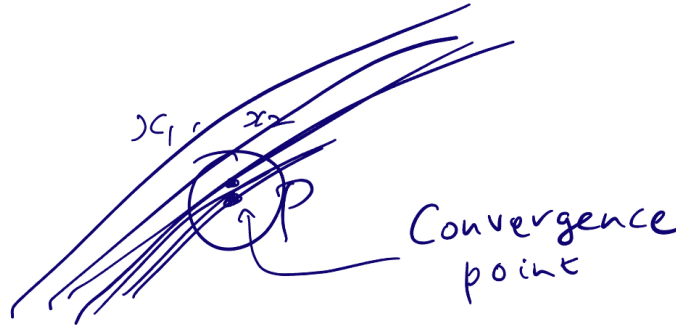
**Definition 2.13.** *A set  $S$  is **sequentially compact** if every infinite sequence in  $S$  has a sub-sequence that converges to a point in  $S$ .*



In general, this definition is not equivalent to compactness<sup>6</sup>. However, if a topological space is *second countable* and sequentially compact, then it is compact. This is the converse to the Bolzano-Weierstrass theorem (as stated in [Wald], Theorem A.9 in Appendix A<sup>7</sup>). Thus to prove compactness we need to show that the space of causal curves is second countable as well as sequentially 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. 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. So it is reasonable to assume that  $C_P^Q$  is also second countable. Then it only remains to show convergence of curves to a limit curve. For this we need a few definitions:

**Definition 2.14.** 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  $n > N$ .



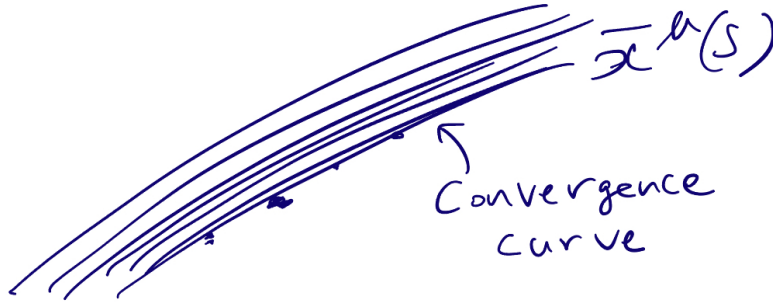
Next we use convergence points to define convergence curves.

**Definition 2.15.** A curve  $x^\mu(s)$  is a **convergence curve** of a sequence  $\{x_n^\mu(s)\}$  of curves if there exists an infinite subsequence  $x_{n_k}^\mu(s)$  for which each point of  $x^\mu(s)$  is a convergence point.

---

<sup>6</sup>The strategy of [Witten] 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 (it is just a diamond-shaped region including its boundary). Then he argues 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 then the proof of compactness of  $C_P^Q$  that works for Minkowski can be recycled. However, following [Wald], here we first take  $M$  to be globally hyperbolic and then prove sequential compactness of  $C_P^Q$  and finally of  $D_P^Q$ .

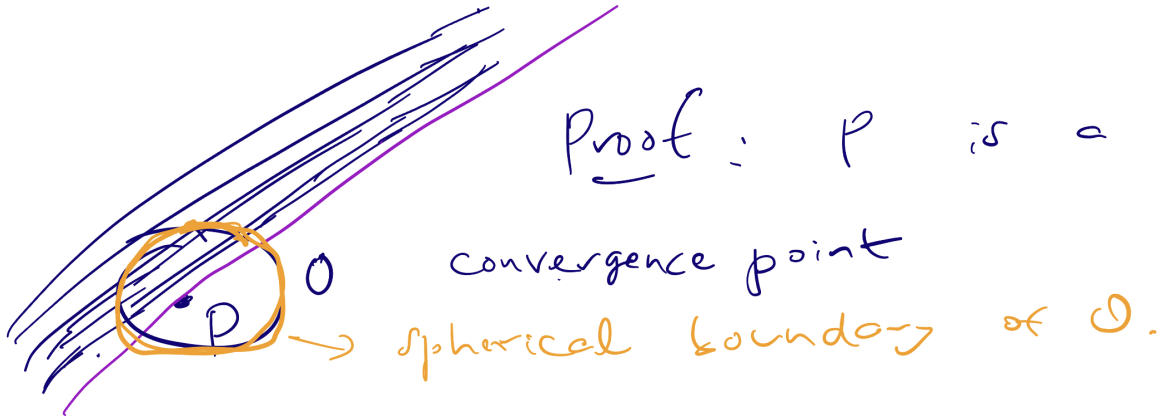
<sup>7</sup>The original Bolzano-Weierstrass was in the context of real numbers, but [Wald]'s version refers to general topological spaces.



We can now state the following theorem:

**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 ([Wald], Lemma 8.1.5).*

Let us sketch how this is proved. We are given that an infinite sub-sequence of curves in the sequence  $x_n^\mu(s)$  converges to  $P$ . This implies that every open neighbourhood  $O$  of  $P$  intersects an infinite sub-sequence of the curves in the sequence  $x_n^\mu(s)$ . Now let us set up Riemann normal coordinates around  $P$  – these are coordinates such that the metric at the point  $P$  is just  $\eta_{\mu\nu}$  and first derivatives of the metric vanish. Next, take a ball around  $P$  of radius  $r$  in these coordinates to be the open neighbourhood  $O$ . So this ball intersects infinitely many of the curves, which must all pass through the spherical boundary of the ball that lies at a Euclidean distance  $r$  (in the chosen coordinates) from  $P$ . Since the sphere is compact, there must be a sub-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.



Finally we return to our main theorem, that  $C_P^Q$  is compact. We prove this by showing that every infinite sequence in  $C_P^Q$  has a limit that lies in  $C_P^Q$ , i.e. every infinite family of causal paths from  $P$  to  $Q$  has a limiting path from  $P$  to  $Q$ . Here is a summary of the proof. First, take both  $P$  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  $M$ , the curves starting at  $P$  become inextendible. Then the previous theorem tells us they have an inextendible convergence curve that also starts at  $P$ . If we now put back the point  $Q$ , this convergence curve either remains inextendible or ends at  $Q$ . The former is impossible since the original curves do not pass through the Cauchy hypersurface, so the convergence curve also cannot pass through the Cauchy hypersurface. Therefore it lies in  $C_P^Q$ . Thus we have shown 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.

There are several important consequence of compactness of the space of causal curves from  $P$  to  $Q$ :

**Theorem 2.7.** *The causal diamond  $D_P^Q = J^+(P) \cap J^-(Q)$  is compact (recall that this was the [Hawking-Ellis] definition of globally hyperbolic space-times). 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 endpoints  $P$  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 ([Wald], Theorem 8.3.10). Using [Wald] Theorem A.2, one immediately sees that  $D_P^Q$  is also closed.

**Theorem 2.8.** *The causal future  $J^+(P)$  of a point  $P$  is closed.*

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  $J^+(P) \cap J^-(Q)$  is closed therefore it should contain all its limit points. Thus we have proved that  $J^+(P)$  is closed.

**Theorem 2.9.** *There is a unique 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). 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 curves  $x_n^\mu(s) \in C_P^Q$  of increasing proper time  $\tau_n$  that grows without limit for large  $n$ . This cannot have a convergent sub-sequence, which contradicts compactness of  $C_P^Q$ . So there must be an upper bound on the proper time of a curve from  $P$  to  $Q$ . Let the smallest value of this upper bound be  $\bar{\tau}$ . Now take a sequence of curves whose proper times  $\tau_n$  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 particle, so this study will lead to physical insights about how actual space-times behave. We will finally be using Einstein's equations!

### 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:

$$\ddot{x}^i + \Gamma_{jk}^i \dot{x}^j \dot{x}^k = 0 \quad (3.1)$$

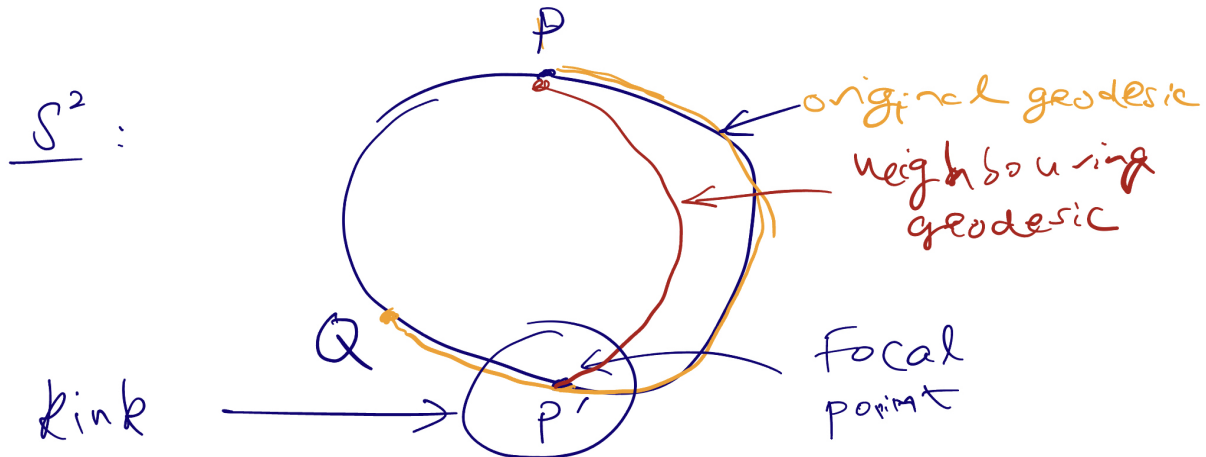
where  $t$  is a parameter along the curve.

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*.

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  $P'$  until a point  $Q$  on the other side of the sphere. The geodesic  $PP'Q$  does not minimise the distance from  $P$  to  $Q$ . Instead, there is another shorter geodesic directly from  $P$  to  $Q$  that does not pass through the south pole  $P'$ .

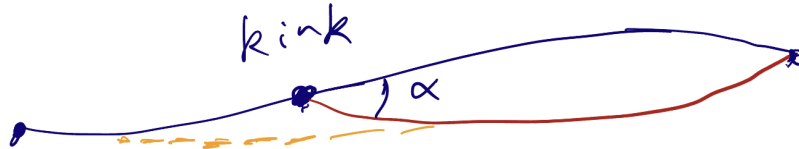


In the sphere example we can easily visualise this “shortening phenomenon”. However we would like to understand it for more general manifolds. As nicely explained in [Witten],

there is a criterion for it to happen on a general manifold: whenever a segment of the original smooth geodesic can be deformed to another nearby geodesic, the former will no longer be of minimal length. We find the minimal length path in two steps, which we can call “switching” and “smoothing”. First we replace the original segment by a segment of a different, nearby geodesic. This is the switching step, and at this stage the total length has not changed. Now we note that this procedure has introduced a kink in the path – since we are switching one segment of the curve, but not the rest of the curve. But a kink can always be locally smoothed out by switching to a more “direct” path, using the triangle inequality. This is the smoothing step, and leads to a shorter path.

In the  $S^2$  example, consider the original geodesic  $PP'Q$  and note that the  $PP'$  segment can easily be switched to other geodesics, since every great circle through  $P$  and  $P'$  is a geodesic. After doing this we have a new path  $PP'Q$  with a kink at  $P'$ . Smoothing out that kink leads to a shorter path  $PQ$  that no longer passes through  $P'$ . In this example we see clearly that the sharpest kink (obtained by switching to the great circle diametrically opposite the original one) leads to the shortest path.

For a general manifold this is not so easy to do explicitly. So we consider the case when the new geodesic segment is infinitesimally separated from the old one. Switching to this geodesic, we get a kink of some infinitesimal angle  $\alpha$  (this angle is  $2\pi$  minus the new angle between incoming and outgoing paths at the kink). Because a geodesic makes the length of a path stationary under any small deformation, the reduction in length by smoothing an infinitesimal kink will vanish to first order, and will therefore generically be of order  $\alpha^2$ .



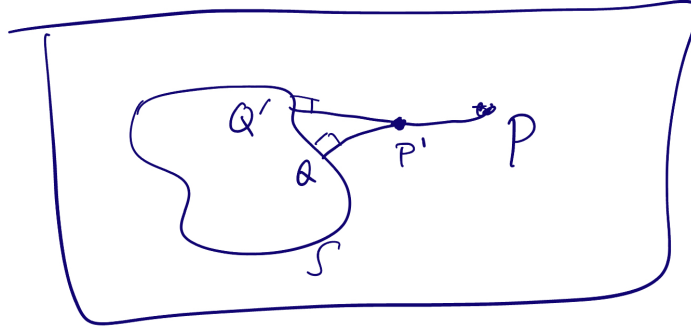
In the above discussion it was clear that the focusing of geodesics gives rise to the possibility of switching, such that after smoothing we reduce the total path length. However there is a slightly more general possibility. Suppose two infinitesimally nearby paths meet at a point, where one is a geodesic but the other is only approximately a geodesic. Again we parametrise the displacement between them by an infinitesimal angle  $\alpha$ . 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  $\alpha$  (other than zero-th order, where it is the original geodesic). Hence it will be longer than the first one at order  $\alpha^2$  (to first order in  $\alpha$  the length cannot change, due to stationarity of the geodesic). This length-

ening 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 consider non-geodesics whose deviation from being geodesic is of order  $\alpha^2$ . In this case the change of length is of order  $\alpha^3$ . Then it cannot compete against the shortening phenomenon, which is always of order  $\alpha^2$ , 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  $M$ . 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 is again by following 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 then works the same way as before and reduces the length.



### 3.2 Focal points and path lengthening: Lorentzian case

In Lorentzian signature, a geodesic satisfies:

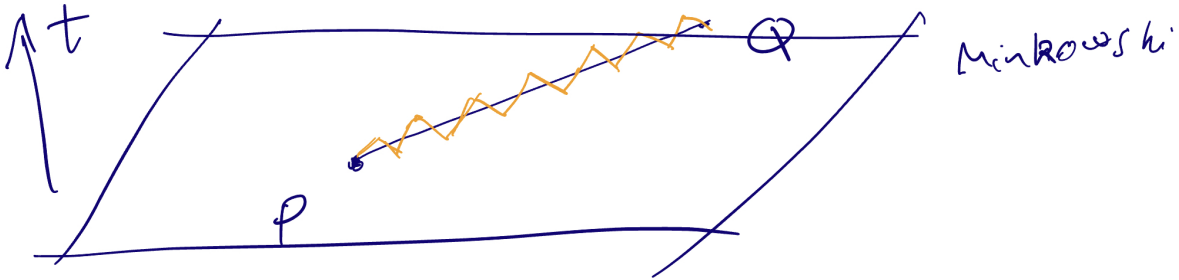
$$\ddot{x}^\mu + \Gamma_{\nu\lambda}^\mu \dot{x}^\nu \dot{x}^\lambda = 0 \quad (3.2)$$

where  $\lambda$  is called an affine parameter. For the timelike or spacelike 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  $P^\mu = \frac{dx^\mu}{d\lambda}$ <sup>8</sup>. The “length” of a path between two points  $P$  and  $Q$  is now replaced by the invariant intervals:

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

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. Thus the norm of the tangent remains constant along the path and so a geodesic cannot change between spacelike, timelike and null along its path but has to be just one of these throughout<sup>9</sup>.

Let us now ask about the possible lengths of geodesics and more general paths between spacelike-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 spacelike lines, each of which is nearly null. Then we get a spacelike curve that is arbitrarily close to zero length. Since  $L$  for a spacelike 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

<sup>8</sup>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. When I have time, I will put in an Appendix on this point of view since it does not feature in most GR textbooks.

<sup>9</sup>A very useful discussion of geodesics can be found in [Wald], Section 3.3

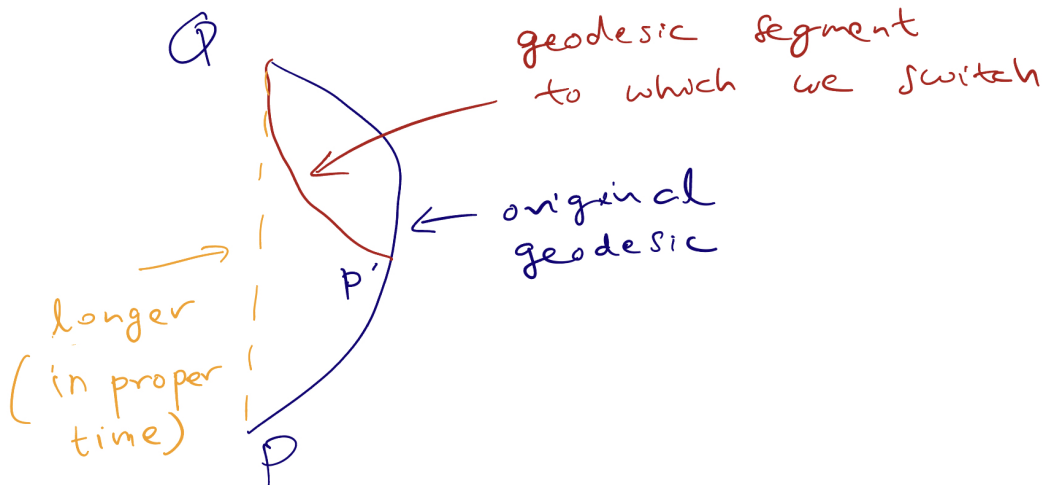


start at  $P$  and zig-zag all over 3d space, say at a fixed time, for as much distance as we want before getting to  $Q$ .

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

Things are different for timelike-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 timelike 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 spacelike in between. We have seen that in Minkowski space-time the straight line joining a pair of points is a geodesic, and for timelike separation this line actually maximises  $L$  between the points, because any additional oscillations in timelike 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 timelike geodesic past a focal point and then find a segment of it that can be “switched” to another timelike 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 timelike geodesics from a point  $P$  to a sub-manifold  $S$ , we will be able to increase the proper 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 timelike geodesics there are some that maximise proper time and others that do not.

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

### 3.3 The Raychaudhuri equation

The  $3\frac{1}{2}$ -page paper [Raychaudhuri] starts with a physical motivation for the problem to be addressed. 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 various difficulties and people could not agree whether these were due to the symmetry assumptions or to a failure of General Relativity. For this reason, it becomes 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.

[Raychaudhuri] does not explicitly place any restrictions on the type of space-time or whether it admits a Cauchy hypersurface. For a modern treatment, following [Witten], we work in a globally hyperbolic space-time with Lorentzian-signature metric  $g_{\mu\nu}(\vec{x}, t)$ . In General Relativity we can perform a general coordinate transformation  $x^\mu \rightarrow x'^\mu(x)$  to go to

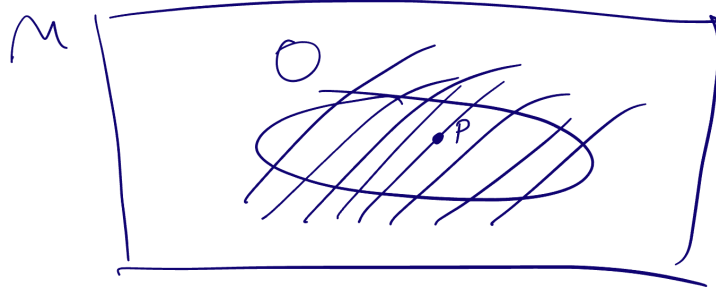
the most convenient coordinate system for the problem at hand. Since there are four such transformations, we can fix four components of the metric. Here, we would like to choose  $g_{00} = -1, g_{0i} = 0$  because with this choice, the metric becomes:

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

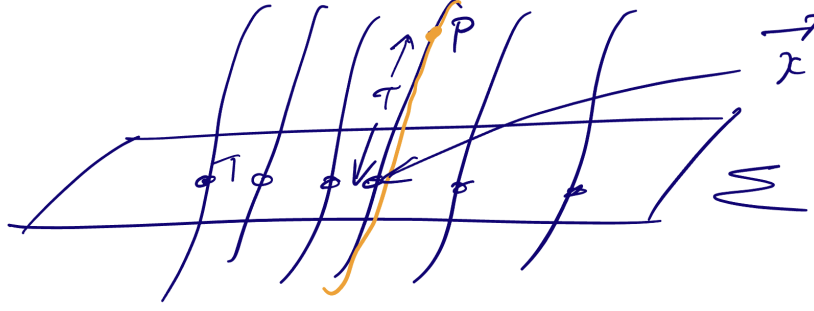
It is straightforward to make such a coordinate choice for a sufficiently short time period away from the Cauchy hypersurface, but whether this can be done far away from the Cauchy hypersurface is less clear. An insight of [Raychaudhuri], though not stated in these words, was to realise that this problem becomes tractable if we give a geometric interpretation for our coordinate choice in terms of geodesics. This is done as follows.

We pick a Cauchy hypersurface  $\Sigma$  and choose a spatial coordinate system  $\vec{x}$  on the Cauchy hypersurface (this only requires ordinary Riemannian geometry, so it can always be done for a differentiable manifold).

**Definition 3.2.** *Given a space-time  $M$  and an open subset  $O$ , a **congruence** is a family of curves 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**.*



Next we consider all timelike geodesics passing orthogonally through  $\Sigma$ . Now there is a natural choice of coordinates in the immediate causal future and past of  $\Sigma$ , as follows. Pick a point  $P$  in the future of  $\Sigma$ , identify the unique geodesic on which it lies, and follow this geodesic back until it intersects with  $\Sigma$ . Since we have chosen a coordinate system on  $\Sigma$ , this intersection point has some definite coordinate  $\vec{x}$ . Then we assign to  $P$  the coordinates  $(\vec{x}, t)$  where  $t$  is the proper time along that geodesic from  $\Sigma$ . Points on  $\Sigma$  itself are labelled  $(\vec{x}, 0)$  and points in the past are labelled  $(\vec{x}, -t)$  where  $t$  is proper time measures along the past-moving geodesic.



There is no guarantee that this defines a good coordinate system at all points in the past and future. This is precisely what we now want to analyse. It is easy to see that the coordinate system we have chosen using geodesics is equivalent to Eq.(3.4). We have  $d\tau^2 \equiv -(g_{00} dt^2 + 2g_{0i} dt dx^i + g_{ij} dx^i dx^j)$ . Along the geodesic,  $d\tau = dt$  and  $dx^i = 0$ , from which we find  $g_{00} = -1$ . Next, because the geodesic orthogonally intersects the Cauchy hypersurface 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:

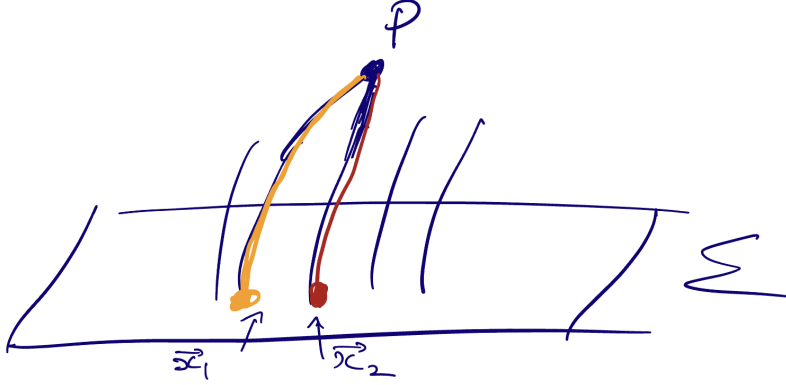
$$\ddot{x}^\mu + \Gamma_{\nu\lambda}^\mu \dot{x}^\nu \dot{x}^\lambda = 0 \quad (3.5)$$

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:

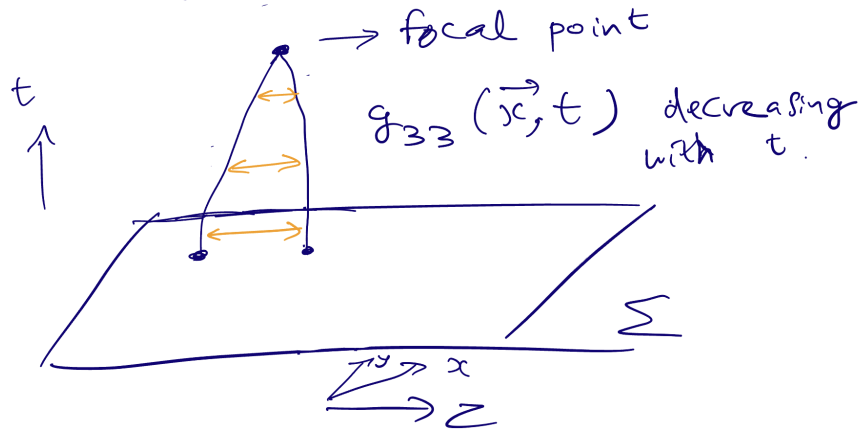
$$\Gamma_{00}^\mu = 0 \quad (3.6)$$

for every  $\mu$ . 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 Eq.(3.4). In this way we have provided a geometric meaning to our preferred coordinate system.

Right away, we can see that this coordinate system will break down if the geodesics propagating outward from  $\Sigma$  have a focal point. In this case, we can follow the focal point back to  $\Sigma$  along multiple geodesics and land on different points of  $\Sigma$ , 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 (horizon of a Schwarzschild black hole) and examples where both the coordinate system and the geometry break down (singularity of a Schwarzschild black hole). So after finding such a breakdown of the coordinate system, additional analysis is needed to understand the implications for the manifold.



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 – since this would describe a direction along which coordinate-separated points have zero distance between them. Since diagonalising the metric is not necessarily easy, we use a simpler criterion – that  $\det g_{ij} = 0$ <sup>10</sup>. Accordingly, we examine the conditions under which the determinant of the spatial metric can vanish after time evolution from a Cauchy hypersurface.

<sup>10</sup>As [Witten] argues, the converse could fail if say one eigenvalue of  $g_{ij}$  goes to 0 while another goes to  $\infty$  such that the product remains fixed. But as he explains, the latter cannot happen if the geometry of  $M$  remains smooth.

We now write Einstein's equations (for the first time in these lectures!) in the form <sup>11</sup>:

$$R_{\mu\nu} = 8\pi G \hat{T}_{\mu\nu} \quad (3.7)$$

where:

$$\hat{T}_{\mu\nu} \equiv T_{\mu\nu} - \frac{1}{D-2} g_{\mu\nu} T^\alpha_\alpha + \frac{1}{4\pi G(D-2)} g_{\mu\nu} \Lambda \quad (3.8)$$

We have kept an arbitrary number of space-time dimensions, since this involves no extra work. Also we have absorbed the cosmological constant into the definition of  $\hat{T}$ .

**Exercise 3.2.** *Derive the above form of Einstein's equations by eliminating  $R$  from the usual form.*

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

$$R_{00} = \partial_\alpha \Gamma^\alpha_{00} - \partial_0 \Gamma^\alpha_{0\alpha} + \Gamma^\alpha_{\alpha\beta} \Gamma^\beta_{00} - \Gamma^\alpha_{0\beta} \Gamma^\beta_{\alpha 0} \quad (3.9)$$

The first and third terms vanish due to Eq.(3.6). To evaluate the remaining terms, we need:

$$\begin{aligned} \Gamma^\alpha_{0\beta} &\equiv \frac{1}{2} g^{\alpha\gamma} (g_{\gamma 0, \beta} + g_{\gamma \beta, 0} - g_{0\beta, \gamma}) \\ &= \frac{1}{2} g^{\alpha\gamma} \dot{g}_{\beta\gamma} \end{aligned} \quad (3.10)$$

where again we dropped terms that vanish due to Eq.(3.6). This then leads to:

$$\Gamma^0_{0\beta} = \Gamma^\alpha_{00} = 0, \quad \Gamma^i_{0j} = \frac{1}{2} g^{ik} \dot{g}_{jk} \quad (3.11)$$

and hence:

$$R_{00} = -\frac{1}{2} \partial_0 (g^{ik} \dot{g}_{ik}) - \frac{1}{4} g^{ik} \dot{g}_{jk} g^{jl} \dot{g}_{il} \quad (3.12)$$

Next we define three quantities which for our convenience. The first is:

$$V(\vec{x}, t) \equiv \det^{\frac{1}{2}} g_{ij}(\vec{x}, t) \quad (3.13)$$

This is the quantity whose possible vanishing we want to study. The remaining two quantities

---

<sup>11</sup>[Raychaudhuri] assumes a specific form for the stress tensor, namely  $T^0_0 = \rho$  with all other components vanishing, and of course he takes  $D = 4$  from the beginning.

are combinations that simplify the equations and also have a physical interpretation:

$$\begin{aligned} \text{Expansion:} \quad \theta &\equiv \frac{1}{2} g^{ik} \dot{g}_{ik} = \frac{\dot{V}}{V} \\ \text{Shear:} \quad \sigma_j^i &\equiv \frac{1}{2} \left( g^{ik} \dot{g}_{jk} - \frac{1}{D-1} \delta_j^i g^{kl} \dot{g}_{kl} \right) \end{aligned} \quad (3.14)$$

where in the first line we used Eq.(A.9). The expansion measures the rate at which the logarithm of  $\det g$  is changing, which tells us how local volumes change with time. The shear tensor  $\sigma_j^i$ , which is traceless, vanishes for isotropic expansion:  $\dot{g}_{ij} = \text{const. } g_{ij}$  as one can easily check. Thus we can think of it as a measure of anisotropy.

Now, using Eq.(A.10) for the shear  $\sigma$  we find:

$$\text{tr } \sigma^2 = \frac{1}{4} \left( g^{ik} \dot{g}_{jk} g^{jl} \dot{g}_{il} \right) - \frac{1}{D-1} \theta^2 \quad (3.15)$$

which immediately leads to:

$$R_{00} = -\dot{\theta} - \frac{1}{D-1} \theta^2 - \text{tr } \sigma^2 \quad (3.16)$$

Thus the 00 component of Einstein's equation Eq.(3.7) becomes:

$$\dot{\theta} + \frac{1}{D-1} \theta^2 = -\text{tr } \sigma^2 - 8\pi G \hat{T}_{00} \quad (3.17)$$

This is the Raychaudhuri equation <sup>12</sup>. It is significant for its generality. As noted in [Raychaudhuri], 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 this 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), the first term is negative-semidefinite, while the sign of the second term depends on what kind of matter and/or cosmological constant we allow.

---

<sup>12</sup>For comparison, [Raychaudhuri] works not with  $V = \det^{\frac{1}{2}} g_{ij}$  but with  $G \equiv \det^{\frac{1}{6}} g_{ij}$ , which is chosen so that (for  $D = 4$ ) the LHS of his equation becomes  $3\frac{\dot{G}}{G}$ . Also he 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 Cauchy hypersurface orthogonally. In this situation there is an additional quantity  $\omega_{ij} \equiv g_{0i,j} - g_{0j,i}$  that contributes a term  $-\text{tr } \omega^2$  (which is positive since  $\omega$  is anti-symmetric) on the RHS of Eq.(3.17). This treatment has also been followed by [Wald], who refers to  $\omega_{ij}$  as the “twist”. However the geodesics we consider are all orthogonal to the Cauchy hypersurface, so this quantity plays no role in our discussion.

Now suppose we arrange that  $\hat{T}_{00} > 0$ , then we get a bound:

$$\dot{\theta} + \frac{1}{D-1}\theta^2 < 0 \quad \text{for all } t \quad (3.18)$$

This can equivalently be expressed as:

$$\ddot{G} < 0 \quad \text{for all } t \quad (3.19)$$

where  $G = V^{\frac{1}{D-1}} = \det^{\frac{1}{2(D-1)}} g_{ij}$  generalises the variable used in [Raychaudhuri] to  $D$  dimensions. Of course the validity of the bound depends on the validity of our assumptions about  $\hat{T}_{00}$ , which involves both the matter distribution and the cosmological constant. From the definition of  $\hat{T}_{\mu\nu}$  in Eq.(3.8), one can verify that the contribution to  $\hat{T}_{00}$  is positive for normal matter and also for a negative cosmological constant. However it is negative for a positive cosmological constant, which is presently believed to be the case in our universe.

Assuming the bound Eq.(3.18) to hold, we can now derive a condition for the existence of focal points, which was our original motivation (the motivation in [Raychaudhuri] is similar in part, as the paper considers cosmological models with a focal point in the past – referred to as “the singular state  $G = 0$ ”). In what follows, it is convenient to use Raychaudhuri’s variable  $G$  rather than  $V$ . Let us pick a point, say  $\vec{x} = \vec{x}_0$ , on the initial hypersurface  $t = 0$ . We choose the initial value  $\frac{\dot{G}(\vec{x}_0, 0)}{G(\vec{x}_0, 0)} = -\frac{1}{\alpha}$  at this point, for some positive number  $\alpha$  (since  $G > 0$ , this means  $\dot{G} < 0$ ). Now:

$$\frac{d}{dt} \left( \frac{G}{\dot{G}} \right) = 1 - \frac{G\ddot{G}}{\dot{G}^2} \geq 1 \quad \text{for all } t \quad (3.20)$$

where we have used Eq.(3.19). It follows that:

$$\frac{G}{\dot{G}}(t) \geq \frac{G}{\dot{G}}(0) + t = t - \alpha \quad (3.21)$$

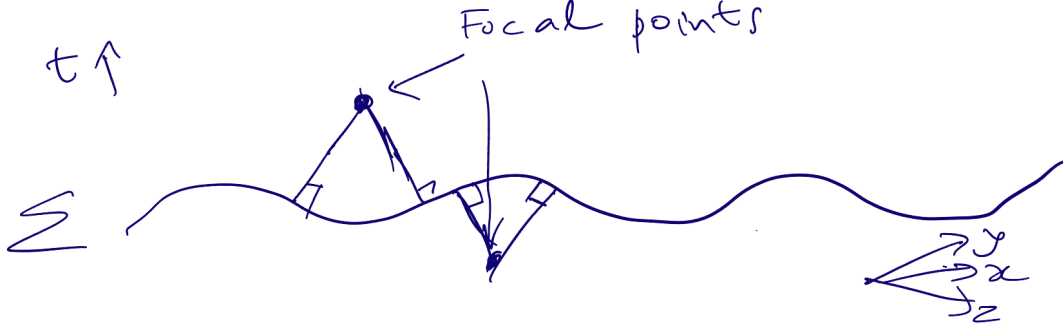
Inverting and integrating, this gives:

$$G(t) \leq G(0) \left( 1 - \frac{t}{\alpha} \right) \quad (3.22)$$

We see that  $G$ , and hence  $\det g$ , goes to 0 at or before the time  $t = \alpha$ . Note that if we had taken  $\dot{G} > 0$  at  $t = 0$ , we would instead have found a focal point in the past of the Cauchy hypersurface. Thus, always assuming  $\hat{T}_{00} > 0$ , there is always a focal point – either in the future or the past.



We still need to decide if this corresponds to just a coordinate singularity or a singularity of space-time. As [Witten] nicely explains, we can even find such focal points in Minkowski space-time, just by taking a “wiggly” Cauchy hypersurface. In this case the focal point definitely does not signal any problem of space-time, but only of our coordinate system. We will discuss this point further in the next section.



### 3.4 Timelike geodesics and Hawking theorem

In 1966, Hawking used the equation of [Raychaudhuri] to derive a singularity theorem. The starting assumptions were a globally hyperbolic space-time and the condition  $\hat{T}_{00} > 0$ , known as the “strong energy condition”. A homogeneous and isotropic universe is described by the well-known FLRW <sup>13</sup> metric:

$$ds^2 = -dt^2 + a(t)^2 dx^i dx^i \quad (3.23)$$

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  $a(\vec{x}, t)$ . Further taking this function independent of  $\vec{x}$  is the assumption of homogeneity. In this metric the Hubble parameter  $H$  is defined as:

$$H(t) \equiv \frac{\dot{a}(t)}{a(t)} \quad (3.24)$$

Now we would like to consider the situation without the symmetry assumptions of homogeneity and isotropy.

For this, notice that Eq.(3.23) is just a special case of Eq.(3.4) with the identification:

$$g_{ij}(\vec{x}, t) = a(t)^2 \delta_{ij} \quad (3.25)$$

---

<sup>13</sup>Friedman-Lemaitre-Robertson-Walker

It follows that in this special case:

$$V \equiv \det^{\frac{1}{2}} g_{ij} = a^{D-1} \quad (3.26)$$

so that:

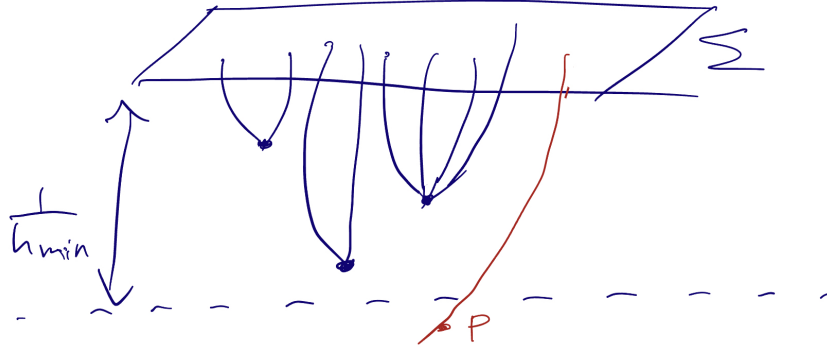
$$a = V^{\frac{1}{D-1}} = G \quad (3.27)$$

where  $G$  is the quantity used in [Raychaudhuri]. This motivates us to define a *local* Hubble parameter as:

$$H(\vec{x}, t) = \frac{\dot{G}(\vec{x}, t)}{G(\vec{x}, t)} = \frac{1}{D-1} \frac{\dot{V}(\vec{x}, t)}{V(\vec{x}, t)} \quad (3.28)$$

In what follows we will refer to this local Hubble parameter as  $H$  without always specifying that it depends on all space-time coordinates.

Now we pick a Cauchy hypersurface  $\Sigma$ , work in general the metric Eq.(3.4) (*not* in the highly symmetric FLRW metric!) and consider a point  $P$  in the causal past of this surface. From the assumption of globally hyperbolic space-time, there is a unique timelike geodesic of maximal proper time from  $P$  to  $\Sigma$  that moreover intersects  $\Sigma$  orthogonally. Let us choose  $H = h_{\min} > 0$  corresponding to  $\frac{G}{\dot{G}} = \frac{1}{h_{\min}}$ , on  $\Sigma$ . Then our discussion of the Raychaudhuri equation tells us that *all* past-directed timelike geodesics from  $\Sigma$  must reach a focal point at or before  $t = -\frac{1}{h_{\min}}$ . It follows that no past-directed timelike geodesic can be continued to the past beyond a time  $t = -\frac{1}{h_{\min}}$ . This property is called *geodesic incompleteness*. It means there is no point  $P$  beyond the focusing time – if there were such points  $P$ , they would not have a future-directed timelike geodesic reaching  $\Sigma$ , which contradicts global hyperbolicity.



This contradiction implies that *there are no points in space-time at a time  $t \lesssim -\frac{1}{h_{\min}}$* . This can be interpreted as the presence of a space-time singularity at a time in the past of this order. However, as emphasised by [Witten] among others, this is not really a fair conclusion – what are popularly known as “singularity theorems” are really statements of geodesic incompleteness.

It may seem that Hawking’s theorem has been reduced to a trivial consequence of the Raychaudhuri equation and global hyperbolicity. However many of the considerations governing focal points, that we discussed earlier, were also part of Hawking’s original discussion. In fact Hawking (and also Penrose) introduced and made use of the notion that they called “conjugate points”, pairs of points such that one has multiple geodesics (or approximate geodesics) passing through them. These are essentially our focal points. As we saw, these allow us to shorten the distance (in Euclidean signature) or maximise the proper time (in Lorentzian signature) by the “switching and smoothing” procedure we described <sup>14</sup>. These considerations have been crucial to obtain the present result <sup>15</sup>.

There is another relevant point here. Instead of settling for the “singularity” interpretation, one could try to resolve the above contradiction by saying that space-time is not globally hyperbolic. However, as explained nicely in [Wald] Theorem 9.5.2, Hawking was able to find a similar result with a weaker assumption. One assumes that the  $M$  is a *strongly causal* space-time as in Definition 2.11, and also that the analogue of the Cauchy hypersurface, an achronal hypersurface  $\Sigma$  with suitable conditions, is compact. Then it can be shown that there is at least *one* past-directed timelike geodesic that is incomplete, and ends at a time  $t \lesssim -\frac{1}{h_{min}}$  <sup>16</sup>. Note that we did not previously have to assume compactness of  $\Sigma$ . 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 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. We may note that these results give considerable physical importance to the Raychaudhuri equation, despite our earlier comments that they may only predict the breakdown of a coordinate system.

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

---

<sup>14</sup>A rigorous discussion of conjugate points can be found in [Wald].

<sup>15</sup>The concept of focal points is not mentioned in [Raychaudhuri] although there is a brief reference to the possibility of  $G = 0$  in the past, which is roughly the same thing.

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

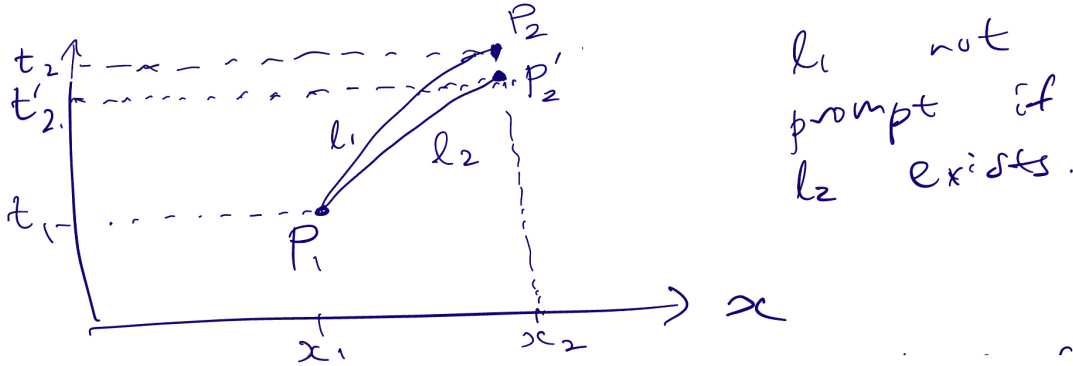
### 3.5 Null geodesics

As we have seen, null geodesics have zero elapsed proper time. This makes it difficult to talk about whether a geodesic maximises proper time. There is, however, a useful concept that will allow us to make analogous statements.

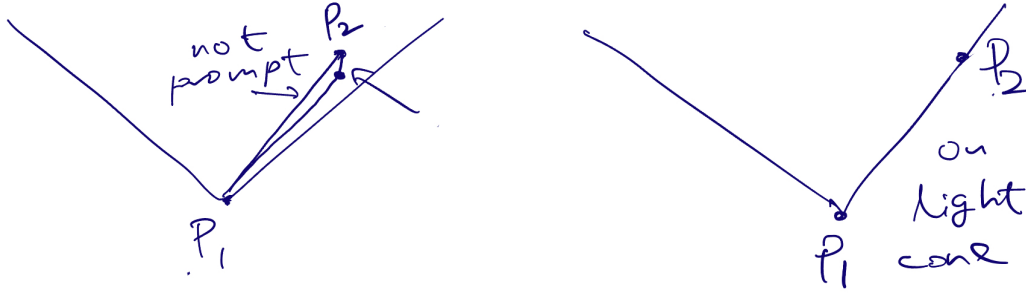
**Definition 3.3.** A causal path from  $P$  to  $Q$  is called **prompt** if there is no other causal path from  $P$  that reaches the spatial location of  $Q$  at an earlier time.

Note that this definition is over all causal paths, not just geodesics. However we will soon see that a prompt path is a null geodesic, while every null geodesic need not be prompt.

The definition can be made more explicit by assigning coordinates  $(\vec{x}_P, t_P), (\vec{x}_Q, t_Q)$  to the two points. 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  $(\vec{x}_Q, t' < t_Q)$ . Note that 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 is “more prompt” than the first, but – again – this does not mean the second path has the same endpoints in space-time as the first.

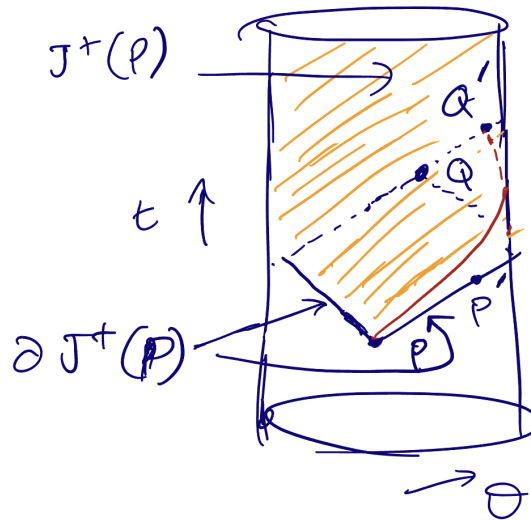


We can understand the situation better by recalling our discussions of the causal future of a point. For there to be any causal path from  $P$  to  $Q$  we must have  $Q \in J^+(P)$ . However if the path is prompt then there is no point  $Q'$  immediately to the past of  $Q$  (i.e. a point  $Q'$  with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



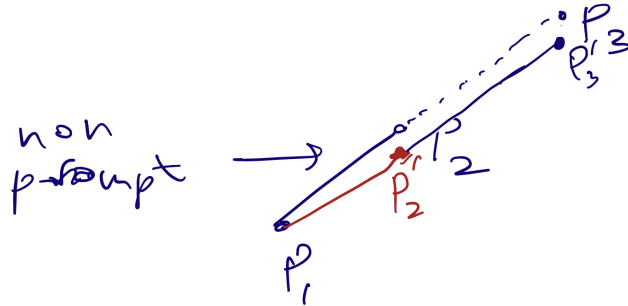
In Minkowski space-time this means  $Q$  lies on the future light-cone of  $P$ . In this case there is a unique null geodesic from  $P$  to every  $Q \in \partial J^+(P)$ , and this is a prompt geodesic. This shows us clearly that non-prompt geodesics can exist only in space-times having non-trivial 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.

A simple model of non-promptness 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 yellow in the figure. We also see that these two geodesics focus 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  $Q$ . In the future of  $Q$ , the entire cylinder is in  $J^+(P)$  and thus no part of the boundary extends to the future of  $Q$ .



Next, note that the null geodesic from  $P$  to  $P'$  is prompt and lies entirely in  $J^+(P)$ . But if we consider the point  $Q'$  beyond the focal point  $Q$ , things are different. There is also a null geodesic from  $P$  to  $Q'$ , going upward to 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 red curve. We also see that  $Q'$  is not on  $\partial J^+(P)$  since that boundary ends at  $Q$ . This tells us that no null geodesic from  $P$  to  $Q'$  can possibly be prompt.

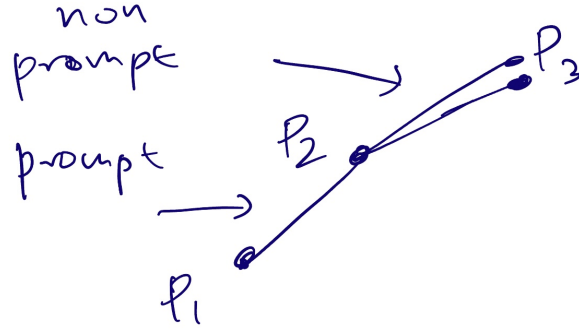
In general globally hyperbolic space-times, a short segment of a null geodesic is always prompt. Now suppose a null geodesic is non-prompt up to some point, can it become prompt by continuing it further? We can show that this is not possible, as follows. Take a null geodesic from  $P_1 = (\vec{x}_1, t_1)$  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  $P_1 P_2$  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  $(\vec{x}_3, t_3 - \epsilon')$  ( $\epsilon'$  does not have to be equal to  $\epsilon$ , 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.



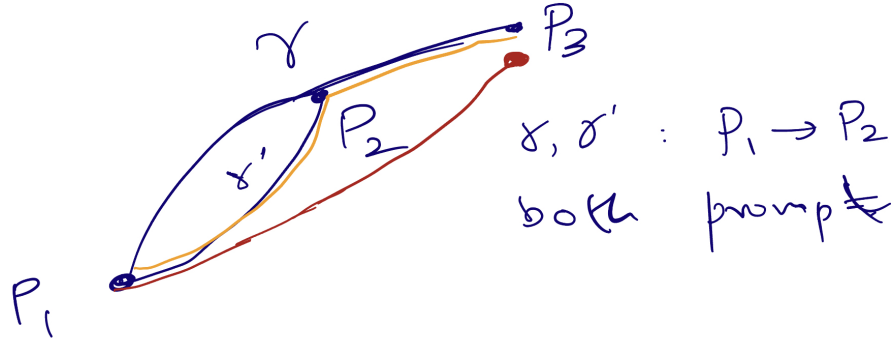
This reasoning also tells us that every prompt causal curve must be a null geodesic. For if there were any non-null (i.e. timelike) segment, we could “smooth it out” by a null curve just below it on a space-time diagram so the original curve would not be prompt.

However the converse is possible: if continued for a long enough distance, a null geodesic that is initially prompt may later become non-prompt. In fact, following [Witten] 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. The cylindrical space-time depicted above provides an example of a non-prompt null geodesic ( $P$  to  $Q'$ ) whose initial part ( $P$  to  $Q$ ) is prompt.

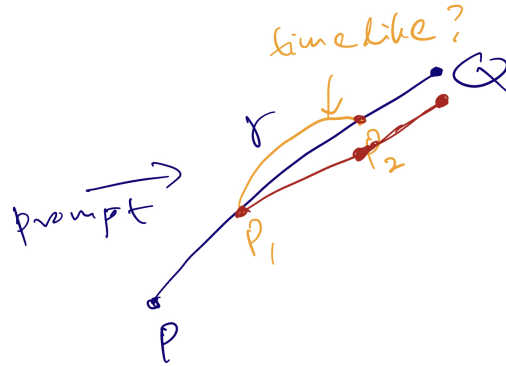


We can now consider focal points for null geodesics. The concept is the same as for timelike geodesics: two null geodesics starting from a common space-time point (or one null geodesic and a curve that is approximately a null geodesic up to first order) may meet again in the future. We now argue that if this happens, the original null geodesic when continued past the focal point is no longer prompt. This argument is analogous to the timelike 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. 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 [Witten], Chapter 8. The cylindrical space-time example above illustrates how promptness fails after a focal point.



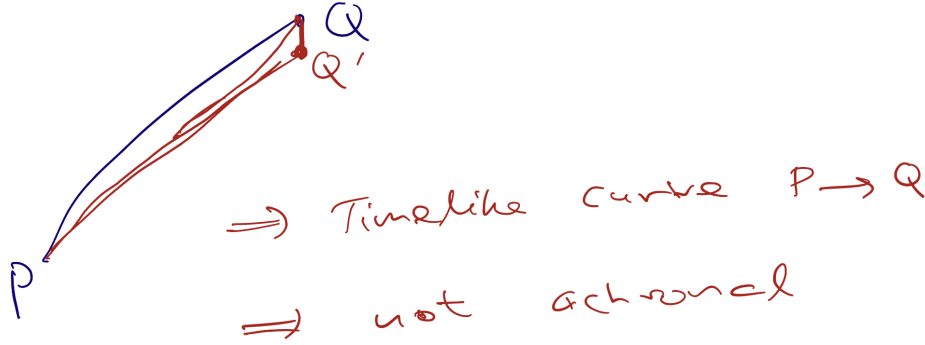
It should be mentioned that, while a focal point renders a null geodesic non-prompt, there could be other ways for it to become non-prompt without necessarily having a focal point.

It turns out that promptness and achronality of a null geodesic are equivalent. Recall that to be achronal, a subset of  $M$  had to be everywhere spacelike or null, and moreover have no timelike path in  $M$  connecting two points. Now take a prompt null geodesic and assume it is not achronal. Then it is possible to connect two points on it by a timelike path. This path is obviously non-prompt in its segment, and one can therefore find a more prompt “shadow” path below it. But now this shadow can be continued until the spatial endpoint of the original geodesic, all the time remaining below it, which contradicts promptness. Thus our assumption was inconsistent. Thus 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 at the end (which is timelike) gives us a causal path that is partly timelike. Such a path can always be deformed to be fully timelike. Then the geodesic is not achronal.





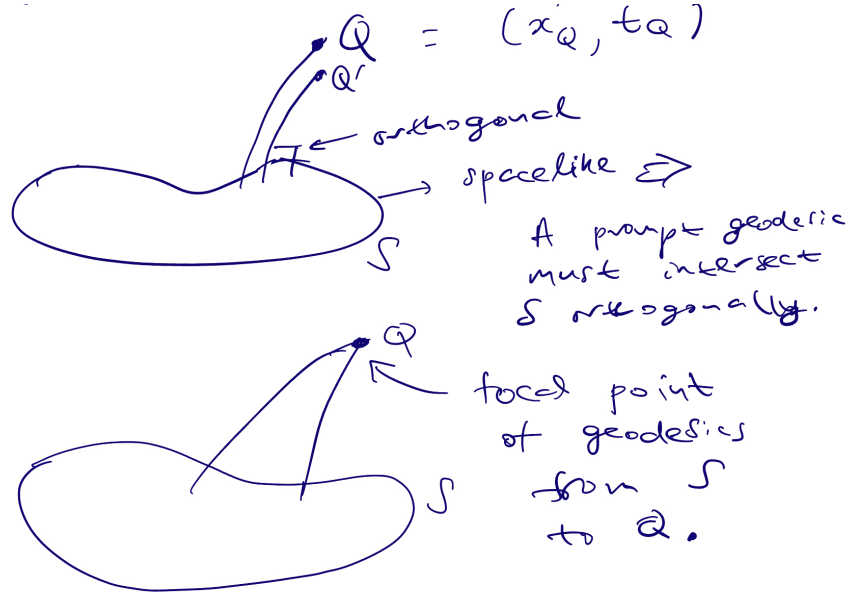
Thus we have proved:

**Theorem 3.1.** *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 above, 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 timelike 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 if there is no other path from anywhere on  $S$  that arrives at the spatial location of  $P$  sooner. A prompt causal path from the set  $S$  to the point  $P$  is necessarily an achronal null geodesic. Further, if  $S$  is everywhere spacelike 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 [Witten]).



Now we have an elementary but important result: if a spacelike sub-manifold is orthogonal to a null direction, it is orthogonal to *two independent* null directions. To see this, consider a vector  $W$  satisfying  $W \cdot W = 0$ . Now consider 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 (since  $W \cdot W = 0$ ). Hence the hypersurface we have defined cannot be spacelike. In fact it is called a *null hypersurface*. To get a spacelike 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:

$$W' = X - \frac{X \cdot X}{2W \cdot X} W \quad (3.29)$$

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, hence the sub-manifold is spacelike<sup>17</sup>.

**Exercise 3.4.** Show that if a spacelike 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

<sup>17</sup>A slightly cleverer way to do this, due to Sachs, is to define two sub-manifolds  $S_1, S_2$ , each of dimension  $(D - 1)$  but not spacelike, by the conditions that one is orthogonal to a null vector  $W_1$  and the other to an independent null vector  $W_2$ . Then  $S = S_1 \cap S_2$  is the desired  $(D - 2)$ -dimensional spacelike manifold.

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  $S$  consisting of the  $x^1 - x^2$  plane at 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 singularity.

**Theorem 3.2.** *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<sup>18</sup>.

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

**Theorem 3.3.** *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 timelike 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 [Witten] for the details.

**Theorem 3.4.** *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 spacelike surface, which we have denoted  $S$ .

---

<sup>18</sup>A sketch of the proof can be found in [Witten], end of Section 5.2, while in [Wald] this result is stated as Theorem 8.3.11 but the proof is left as an exercise.

### 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 timelike situation. We start with a compact spacelike sub-manifold  $S$  of dimension  $(D-2)$  in  $M$ , on which we assume a coordinate system  $x^a$ ,  $a = 1, 2, \dots, D-2$ . Now we want to consider null geodesics emanating from it. Such geodesics satisfy the usual geodesic equation:

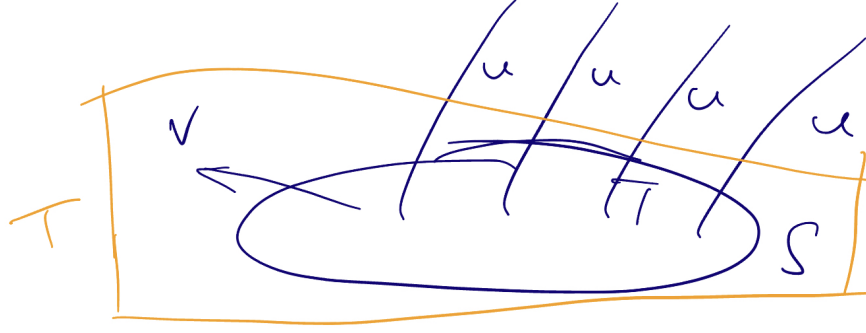
$$\frac{d^2 x^\alpha}{d\tau^2} + \Gamma_{\lambda\rho}^\alpha \frac{dx^\lambda}{d\tau} \frac{dx^\rho}{d\tau} = 0 \quad (3.30)$$

where  $\tau$  is an affine parameter, defined up to shifts and scaling  $\tau \rightarrow a\tau + b$ . Recall that for timelike 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  $\tau = u$ , with  $u$  a parameter such that the geodesic equation holds and such that  $u = 0$  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  $S$ . We pick one of these and continue it all over  $S$ <sup>19</sup>. These generate a  $(D-1)$ -dimensional space  $Y$ , on which we choose coordinates as follows: (i) each point has a spatial coordinate  $x^a$  obtained by following the null geodesic on which it lies back to  $S$  and choosing the coordinates of that intersection point, (ii) each point has a coordinate  $u$  that is the value of the affine parameter along the null geodesic, starting with  $u = 0$  on  $S$ . In these coordinates, the metric component  $g_{uu}$  vanishes since  $u$  is a null direction. Note that the ambiguity in scaling of  $u$  can be space-dependent: we are allowed to replace  $u$  by  $e^{f(x^a)}u$ , where we parametrised the factor as an exponential because it must be strictly positive (otherwise the null future and past would get exchanged).

---

<sup>19</sup>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 spacelike 3-manifold in 4d there is only one set of future-directed timelike geodesics.



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  $(0, +, \dots +)$ . 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 spacelike hypersurface  $T$  of dimension  $(D - 1)$  and choose a function  $v$  on  $T$  that vanishes along  $S$ , but whose normal derivative to  $S$  is zero everywhere. Then we treat the value of  $v$  as a coordinate transverse to  $S$ . Thus we can parametrise  $T$  in terms of “slices”  $S(v)$  that are all copies of  $S$ , where  $S(0) = S$ . So far,  $x^a$  were only defined on  $S$  but we can 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 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  $S(v)$ .



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  $g_{ab}$  among the  $x^a$  is arbitrary, this is the metric

of the  $(D - 2)$ -dimensional spacelike manifold  $S$ . 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 said 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  $\det g_{\mu\nu} = 0$ .

This is as far as we can get without using the geodesic equation Eq.(3.30). Imposing this equation (with  $\tau = u$ ) we get:

$$\Gamma_{uu}^\alpha = 0 \quad (3.31)$$

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

$$g_{u\beta,u} = 0 \quad (3.32)$$

for all  $\beta$ . 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 got no new constraints on  $g_{va}$  or  $g_{ab}$  so they are completely general. So our metric is:

$$ds^2 = -2e^q dudv + 2c_a dv dx^a + g_{ab} dx^a dx^b + g_{vv} dv^2 \quad (3.33)$$

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

$$dudv \rightarrow dudv + \partial_v h dv^2 + \partial_u h dudv + \partial_a h dv dx^a \quad (3.34)$$

This modifies the  $dudv$  and  $dv dx^a$  terms, but they are arbitrary anyway. Thus, at least locally, we can set  $g_{vv}$  to anything we like. We will make the choice  $g_{vv} = g_{ab} c^a c^b$  where  $c^a = g^{ab} c_b$ <sup>20</sup>, which turns out extremely convenient as we see below.

The metric now becomes<sup>21</sup>:

$$ds^2 = -e^q du^2 + g_{ab} (dx^a + c^a dv)(dx^b + c^b dv) \quad (3.35)$$

The arbitrary functions are  $q, g_{ab}, c_a$  with  $a = 1, \dots, D - 2$ , making up a total of  $\frac{D(D-1)}{2}$  functions. This is the same number of functions in the timelike case.

If now we compute the inverse of the above metric, we find that  $g^{uu} = g^{vv} = g^{av} = 0$

---

<sup>20</sup>At this stage  $g^{ab}$  just means the corresponding components of the inverse of the 4d metric.

<sup>21</sup>The entire derivation here has followed [Witten], but a direct and elegant derivation of this metric can be found in the short paper ‘‘On the Characteristic Initial Value Problem in Gravitational Theory’’ by J. Sachs, J. Math. Phys. 3 (1962), 908.

everywhere. In particular,  $g^{vv} = 0$  is an immediate consequence of the special form we chose for  $g_{vv}$ <sup>22</sup>.

**Exercise 3.5.** *For the metric in Eq.(3.35), calculate all components  $g^{\mu\nu}$  of the inverse metric and verify the above statements.*

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

$$R_{uu} = 8\pi G T_{uu} \quad (3.36)$$

of Einstein's equations. Important simplifications take place due to the fact that  $g_{uu} = 0$ : in the quantity  $\hat{T}$  defined in Eq.(3.8), the second and third terms vanish when we take the  $uu$  component. As a result the trace of  $T$  and also the cosmological constant drop out of this equation<sup>23</sup>. Thus the energy condition we need in order to get a bound is the *null energy condition*:

$$T_{\mu\nu} k^\mu k^\nu > 0 \quad (3.37)$$

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

Next we compute  $R_{uu}$  in terms of the functions  $q, c_a, g_{ab}$ . We have:

$$R_{uu} = -\partial_u \Gamma_{u\alpha}^\alpha - \Gamma_{u\beta}^\alpha \Gamma_{\alpha u}^\beta \quad (3.38)$$

where we dropped terms that vanish due to Eq.(3.31). We already know that  $\Gamma_{uu}^\alpha = 0$ , so we only need to evaluate  $\Gamma_{uv}^\alpha$  and  $\Gamma_{ub}^\alpha$ :

$$\begin{aligned} \Gamma_{uv}^\alpha &= \frac{1}{2} g^{\alpha\gamma} (g_{\gamma u, v} + g_{\gamma v, u} - g_{uv, \gamma}) \\ &= \frac{1}{2} g^{\alpha v} g_{vv, u} + \frac{1}{2} g^{\alpha a} (g_{av, u} - g_{uv, a}) \\ \Gamma_{ub}^\alpha &= \frac{1}{2} g^{\alpha\gamma} (g_{\gamma u, b} + g_{\gamma b, u} - g_{ub, \gamma}) \\ &= \frac{1}{2} g^{\alpha v} (g_{uv, b} + g_{vb, u}) + \frac{1}{2} g^{\alpha a} g_{ab, u} \end{aligned} \quad (3.39)$$

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

$$\begin{aligned} \Gamma_{u\alpha}^\alpha &= \frac{1}{2} (g^{vv} g_{vv, u} + g^{av} g_{av, u} - g^{av} g_{uv, a} + g^{av} g_{uv, a} + g^{av} g_{av, u} + g^{ab} g_{ab, u}) \\ &= \frac{1}{2} g^{ab} g_{ab, u} \end{aligned} \quad (3.40)$$

---

<sup>22</sup>A nice consequence of these properties is that if we take the inverse 4d metric  $g^{\mu\nu}$  and restrict to  $\mu = a, \nu = b$ , the resulting matrix  $g^{ab}$  is the inverse of  $g_{ab}$  in the 2d sense. Hence we also have  $c^a = g^{ab} c_b$  with no ambiguity in the meaning of  $g^{ab}$ .

<sup>23</sup>Of course,  $\Lambda$  still appears in the  $R_{uv}, R_{va}$  and  $R_{ab}$  components of Einstein's equations.

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

$$\Gamma_{u\beta}^\alpha \Gamma_{u\alpha}^\beta = \frac{1}{4} g^{ac} g_{bc,u} g^{bd} g_{ad,u} \quad (3.41)$$

It follows that:

$$R_{uu} = -\frac{1}{2} g^{ab} g_{ab,u} - \frac{1}{4} g^{ac} g_{bc,u} g^{bd} g_{ad,u} \quad (3.42)$$

Notice the similarity with Eq.(3.12), but with a  $(D-2)$ -dimensional metric.

**Exercise 3.6.** *Explicitly verify the second term in  $R_{uu}$  in Eq.(3.42).*

Next, denote:

$$A = \det^{\frac{1}{2}} g_{ab} \quad (3.43)$$

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

$$\begin{aligned} \text{Null expansion:} \quad \theta &\equiv \frac{\partial_u A}{A} = \frac{1}{2} g^{ab} \partial_u g_{ab} \\ \text{Null shear:} \quad \sigma_b^a &\equiv \frac{1}{2} \left( g^{ac} \partial_u g_{cb} - \frac{1}{D-2} \delta_b^a g^{cd} \partial_u g_{cd} \right) \end{aligned} \quad (3.44)$$

Again, these are very similar to their timelike versions Eq.(3.14) 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:

$$\dot{\theta} + \frac{1}{D-2} \theta^2 = -\text{tr } \sigma^2 - 8\pi G T_{uu} \quad (3.45)$$

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:

$$\hat{G} \equiv \det^{\frac{1}{D-2}} g_{ab} \quad (3.46)$$

which plays the role of  $G$  in [Raychaudhuri] for the timelike case. The bound then becomes:

$$\ddot{\hat{G}} < 0 \quad (3.47)$$

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



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. However the other cases are also possible in some situations.

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

The round 2-sphere in Minkowski space-time is not a trapped surface, because its null expansion is positive outside the sphere and negative inside. 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.

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  $ds^2 = r^2 d\theta^2 = (u - u')^2 d\theta^2$  where  $u = t + r, u' = t - r$ . Thus  $A = u - u'$

and we have  $\partial_u A = 1 > 0$  while  $\partial_{u'} A = -1 < 0$ , 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.

For now, instead of asking when trapped surfaces arise, we will ask what are the consequences if they do arise. Later we will discuss the physical context, and the results we are deriving will be applicable there. The consequences of a trapped surface are embodied in Penrose's famous theorem, which we now state as follows:

**Theorem 3.5.** *Let  $M$  be a globally hyperbolic space-time having a non-compact Cauchy surface  $\Sigma$  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  $\Sigma$ . Finally we recall that any achronal hypersurface must be homeomorphic (topologically equivalent) to part of a given Cauchy hypersurface— but, as  $\Sigma$  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  $u'$  similarly going to the future but along  $W'$  from  $S$ . We now assume  $S$  is a trapped surface, so  $\frac{\hat{G}}{\partial_u \hat{G}} = -\frac{1}{\alpha} < 0$  at  $u = 0$  (initial negative null expansion along  $W$ ) and along the second geodesic with  $\frac{\hat{G}}{\partial_{u'} \hat{G}} = -\frac{1}{\alpha'} < 0$  at  $u' = 0$  (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  $\hat{G} \rightarrow 0$  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  $\gamma$  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  $\gamma_{PQ}$  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  $\gamma$  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  $\gamma_{PQ}$  since  $\partial J^+(S)$  is closed in the space-time  $M$  (Theorem 3.2 of these notes). Since  $\gamma_{PQ}$  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  $\gamma'$  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 3.3 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  $\Sigma$  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:

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

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:

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

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

The coordinates  $\theta, \phi$  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:

$$dt = \pm \frac{dr}{1 - \frac{2GM}{r}} \quad (4.3)$$

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  $(t, r, \theta, \phi)$  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:

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

where:

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

Note that:

$$\frac{dr^*}{dr} = \frac{1}{1 - \frac{2GM}{r}} \quad (4.6)$$

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  $(u, r, \theta, \phi)$  or  $(v, r, \theta, \phi)$ <sup>24</sup>. These are called outgoing/ingoing Eddington-Finkelstein coordinates. Choosing the ingoing one, we find the metric to be:

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

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  $(t, r, \theta, \phi)$  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)

---

<sup>24</sup>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.

together with  $d\theta = d\phi = 0$  as before. We get:

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

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:

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

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.

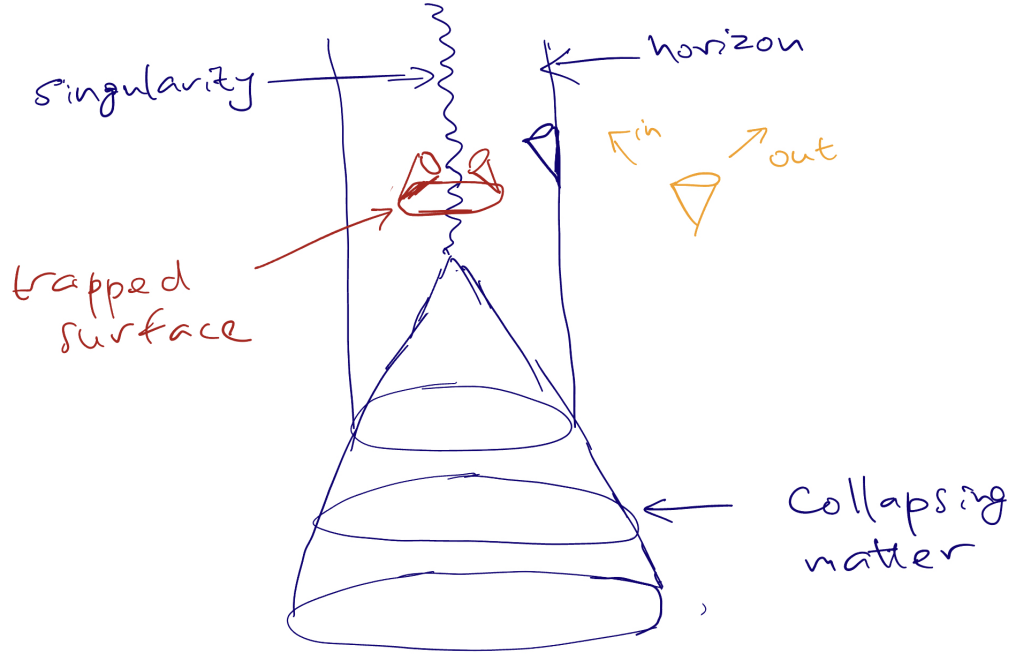
The first (originally ingoing) null ray  $dv = 0$  does not depend on  $r$  so it does not change as we vary  $r$ , and continues to correspond to an ingoing null ray. However for  $r < 2GM$  the second null ray becomes:

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

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  $r \rightarrow 0$ . Since this quantity is invariant under general coordinate transformations, the singularity as  $r \rightarrow 0$  is a genuine feature of the geometry that cannot be removed by any change of coordinates.

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.



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  $(\theta, \phi)$ , 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  $r \rightarrow 2GM$  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  $\Sigma$ . 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 to 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<sup>25</sup>. Instead of trying to be most general,

---

<sup>25</sup>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,



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 points from where causal (timelike 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”.

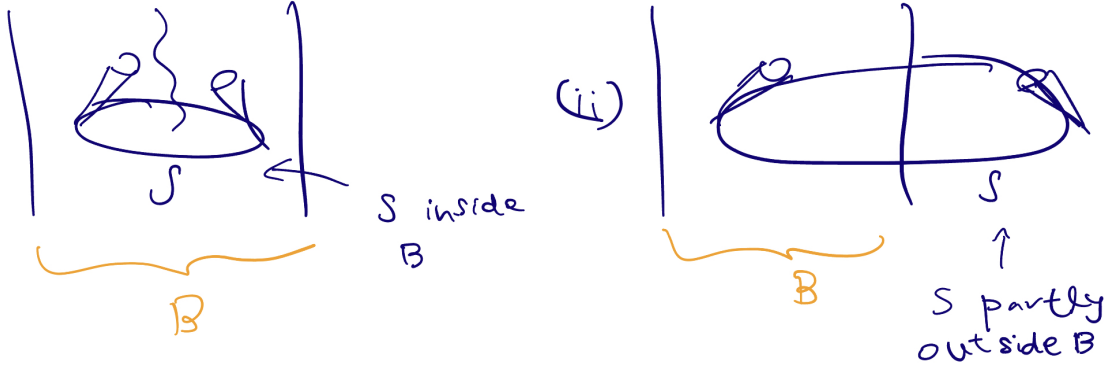
What topological information do we have about  $J^-(I)$  and  $B$ ? There is a result ([Wald], page 308) that  $J^-(I)$  is an open set. The formal proof relies on a corollary to our Theorem 2.1 on the existence of convex normal neighbourhoods. A more intuitive proof ([Witten], Section 6.2) says that 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

---

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.

completely inside  $B$ . A formal proof is in [Wald] 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 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  $P \in S \subset B$ .

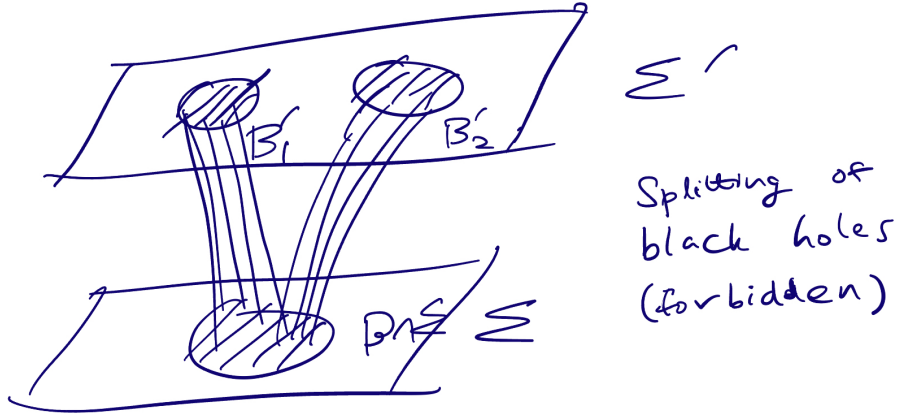
From this we can deduce the following important result:

**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  $\Sigma$  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  $\Sigma'$  after  $t_2$  intersects both black holes. Now consider a future-directed null geodesic coming out of

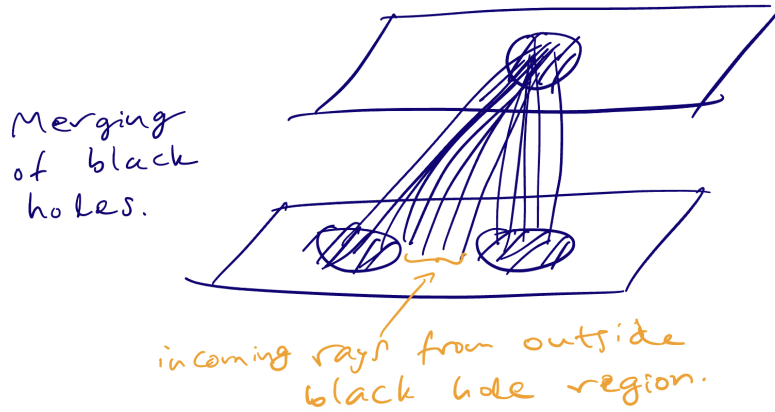
$B \cap \Sigma$ . We already showed that this geodesic must remain inside the black hole region of space-time forever. Therefore if it reaches  $\Sigma'$  then it must enter either  $B'_1 \cap \Sigma'$  or  $B'_2 \cap \Sigma'$ .

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 [Wald], 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.



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.



#### 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  $P_1, P_2, \dots$  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<sup>26</sup>. Thus each horizon generator just continues along the horizon forever towards the future.

Now let us take two Cauchy hypersurfaces  $\Sigma_1, \Sigma_2$  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  $\theta$  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  $\Sigma_2$  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  $\Sigma_1$  to  $\Sigma_2$ , 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:

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

which is Hawking’s area increase theorem. This proof relies on the fact that horizon generators are a “complete set”, an alternate without this assumption is sketched in [Witten]. A more formal proof is in [Wald] Theorem 12.2.6.

<sup>26</sup>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.

## 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”:

$$dE = TdS - PdV \quad (4.12)$$

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<sup>27</sup> 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:

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

The area of this black hole is:

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

In the limit  $J, Q \rightarrow 0$  these quantities reduce to  $r_+ = 2GM, r_- = 0$  and  $A = 16\pi G^2 M^2$

---

<sup>27</sup>A space-time is said to be stationary if it has a timelike Killing vector field.

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:

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

In the limit  $\vec{J} \rightarrow 0$  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:

$$\vec{\Omega} \equiv \frac{\frac{\vec{J}}{M}}{r_+^2 + \frac{\vec{J}^2}{M^2}} \quad (4.16)$$

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:

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

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:

$$dA = 8\pi G (r_+ dM + M dr_+ - Q dQ) \quad (4.18)$$

Using:

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

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

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

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

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

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 <sup>28</sup>, 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  $\Theta$  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 a more general class of stationary axisymmetric black holes than the Kerr-Newman vacuum solution, namely

---

<sup>28</sup>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.



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  $\Theta$  of  $dA$  in the thermodynamic relation is related to the surface gravity of the black hole:

$$\Theta = \frac{\kappa}{8\pi} \quad (4.25)$$

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  $\Theta$  so we cannot unambiguously fix the constant factor relating area to entropy. He tried to fix these problems in the same paper, obtaining (in units where  $c = 1$ ):

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

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 <sup>29</sup> This constant has since then turned out to be incorrect.

---

<sup>29</sup>For much of the paper, Bekenstein tries to interpret his thermodynamic analogy in the language of

Meanwhile BCH simply chose the constant to be 1 and quoted their “temperature” was  $\frac{\kappa}{8\pi}$ . 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  $\kappa$  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.

## A References, notation, conventions, identities

### A.1 References

These notes closely follow the ideas expounded in the following references:

- “General Relativity”, by Robert Wald.

---

classical information theory. Again, this was a far-reaching idea that had far-reaching consequences when it was implemented with quantum, rather than classical, information.

- “The Large-Scale Structure of Space-Time”, by Stephen Hawking and George Ellis.
- “Light Rays, Singularities, and All That”, arXiv: 1901.03928 by Edward Witten.
- “Relativistic Cosmology I”, Phys.Rev. 98 (1955), 1123 by Amalkumar Raychaudhuri.

Even when discussing the same results, the first three references take different approaches – so 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. All these sources are far too long and dense for a short course of 12 lectures, for which the present notes are designed. Moreover the books of Wald and Hawking-Ellis, and to a lesser extent the article of Witten, assume considerable familiarity with topology and differential geometry and I also did the same initially, but I have then gone into a few details as and when I felt they were needed.

In the notes I will refer to the sources above as [Wald], [Hawking-Ellis], [Witten] and [Raychaudhuri].

## A.2 Notation and conventions

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

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

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

Most often the arguments  $(t, \vec{x})$  will be suppressed.

Christoffel symbol:

$$\Gamma_{\mu\nu}^{\alpha} \equiv \frac{1}{2}g^{\alpha\beta}(g_{\beta\mu,\nu} + g_{\beta\nu,\mu} - g_{\mu\nu,\beta}) \quad (\text{A.1})$$

Riemann curvature tensor:

$$R^{\alpha}{}_{\lambda\mu\nu} \equiv \partial_{\mu}\Gamma_{\nu\lambda}^{\alpha} - \partial_{\nu}\Gamma_{\mu\lambda}^{\alpha} + \Gamma_{\mu\beta}^{\alpha}\Gamma_{\nu\lambda}^{\beta} - \Gamma_{\nu\beta}^{\alpha}\Gamma_{\mu\lambda}^{\beta} \quad (\text{A.2})$$

Ricci tensor:

$$R_{\mu\nu} \equiv R^{\alpha}{}_{\mu\alpha\nu} \quad (\text{A.3})$$

Ricci scalar:

$$R \equiv g^{\mu\nu}R_{\mu\nu} \quad (\text{A.4})$$

Energy-momentum tensor:

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

Einstein-Hilbert action:

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

Einstein equations:

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

### A.3 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  $|M|$ :

$$\begin{aligned} \delta|M| &= \delta \exp(\log|M|) \\ &= \delta \exp(\text{tr} \log M) \\ &= (\text{tr} M^{-1} \delta M) \exp(\text{tr} \log M) \\ &= |M|^{-1} (\text{tr} M^{-1} \delta M) \end{aligned} \quad (\text{A.8})$$

Hence,

$$|M|^{-1} \delta|M| = M^{ik} \delta M_{ik} \quad (\text{A.9})$$

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

Another helpful identity is:

$$\text{tr} \left( M - \frac{1}{d} (\text{tr} M) \right)^2 = \text{tr} (M^2) - \frac{1}{d} (\text{tr} M)^2 \quad (\text{A.10})$$