Relativity and Singularities – A Short Introduction for Mathematicians

José Natário

Abstract. We summarize the main ideas of General Relativity and Lorentzian geometry, leading to a proof of the simplest of the celebrated Hawking-Penrose singularity theorems. The reader is assumed to be familiar with Riemannian geometry and point set topology.

Introduction

Historically, much of the development of Riemannian geometry has been driven by General Relativity. This theory models spacetime as a Lorentzian manifold, which is analogous to a Riemannian manifold except that the positive definite metric is replaced by a metric with signature $(-, +, \ldots, +)$. Not only is Lorentzian geometry similar to Riemannian geometry in many respects but also Riemannian manifolds arise naturally as submanifolds of Lorentzian manifolds. Physical considerations then give rise to conjectures in Riemannian geometry. Recent examples of results inspired by such conjectures include the mass positivity theorem (Schoen and Yau, [SY79], [SY81]) and the Penrose inequality (Bray, [Bra01], Huisken and Ilmanen, [HI01]).

On the other hand, the effort involved in learning Lorentzian geometry is minimal once one has mastered Riemannian geometry. It therefore seems strange that many mathematicians (even geometers) choose not to do so. This may be in part due to the fact that most introductions to General Relativity start from first principles, developing the required differential geometry tools at length, and mostly focus on physical implications of the theory. A mathematician might prefer a shorter introduction to the subject from a more advanced starting point, focusing on interesting mathematical ideas. This paper aims to provide such an introduction, leading to a nontrivial result – the simplest of the Hawking-Penrose singularity theorems ([Pen65], [Haw67], [HP70]). These theorems basically state that physically reasonable Lorentzian manifolds (in a precise mathematical sense) must

---

be singular (i.e. geodesically incomplete). Since the motions of free-falling particles are represented by geodesics, this has the physical interpretation that General Relativity cannot be a complete description of Nature.

The paper is divided into three sections. The first section contains basic ideas of General Relativity and Lorentzian geometry: timelike, spacelike and null vectors and curves, matter models, the Einstein equation and its simplest solutions. Causality theory is developed in the second section, where we discuss time orientation, chronological and causal future and past sets, local causal structure, local maximizing properties of timelike geodesics (the Twin Paradox), the chronology condition, stable causality, domains of dependence and global hyperbolicity. The third section contains the proof of the singularity theorem. The proof has three ingredients: the first is that timelike geodesics cease to maximize the distance to a given time slice S once a conjugate point is reached. The second is that, under a physically reasonable condition (the strong energy condition), conjugate points always occur. The third is that a length maximizing geodesic connecting S to any given point p always exists in a globally hyperbolic Lorentzian manifold. This is the most difficult (and mathematically interesting) point to prove; the proof is achieved by showing that the set of timelike curves connecting S to p with the Hausdorff metric is a compact space where the length functional is upper semicontinuous.

We assume the reader to be familiar with elementary Riemannian geometry (as in [dC93]) and point set topology (as in [Mun00]). For the reader whose interest is aroused by this short introduction, there are many excellent texts on General Relativity, usually containing also the relevant differential and Lorentzian geometry. These range from introductory ([Sch02]) to more advanced ([Wal84]) to encyclopedic ([MTW73]). More mathematically oriented treatments can be found in [BEE96], [O'N83] ([GHL04] also contains a brief glance at pseudo-Riemannian geometry). Causality and the singularity theorems are treated in greater detail in [Pen87], [HE95], [Nab88].

1. General Relativity and Lorentzian Geometry

General Relativity is the physical theory of space, time and gravitation. It models spacetime (i.e. the set of all physical events) as a 4-dimensional Lorentzian manifold (spacetimes with different numbers of dimensions are also considered, e.g. in String Theory).

Definition 1.1. An n-dimensional **pseudo-Riemannian manifold** is a pair (M, g), where M is an n-dimensional differentiable manifold and g is a symmetric, nondegenerate 2-tensor field on M (called the **metric**). A

pseudo-Riemannian manifold is said to be **Riemannian** if g has signature $(+ \ldots +)$, and is said to be **Lorentzian** if g has signature $(- + \ldots +)$.

Example 1.2. The simplest Riemannian manifold is **Euclidean space**, which is $\mathbb{R}^n$ with the Riemannian metric

$$g = dx^1 \otimes dx^1 + \ldots + dx^n \otimes dx^n.$$

Analogously, the simplest Lorentzian manifold is **Minkowski space**, which is $\mathbb{R}^{n+1}$ with the Lorentzian metric

$$g = -dx^0 \otimes dx^0 + dx^1 \otimes dx^1 + \ldots + dx^n \otimes dx^n.$$

Lorentzian geometry is similar to Riemannian geometry in many respects. For instance, the Levi-Civita Theorem still holds (with the same proof) .

Theorem 1.3. (Levi-Civita) *Let (M, g) be a pseudo-Riemannian manifold. Then there exists a unique symmetric affine connection compatible with the metric.*

In particular, a Lorentzian manifold comes equipped with geodesics.

Example 1.4. The Levi-Civita connection of Minkowski space is the trivial connection, and its geodesics are straight lines.

On the other hand, the minus sign in the signature does introduce many novel features to Lorentzian geometry.

Definition 1.5. Let (M, g) be a Lorentzian manifold and $p \in M$. A vector $v \in T_p M$ is said to be

   (i) **Timelike** if $\langle v, v \rangle < 0$;
  (ii) **Spacelike** if $\langle v, v \rangle > 0$;
 (iii) **Null** if $\langle v, v \rangle = 0$.

The **length** of v is $|v| = |\langle v, v \rangle|^{\frac{1}{2}}$. (As usual one writes $\langle v, w \rangle$ for $g(v, w)$).

A curve $c : I \subset \mathbb{R} \to M$ is said to be timelike, spacelike or null if its tangent vector $\dot{c}(t)$ is timelike, spacelike or null for all $t \in I$. If c is a geodesic, then the nature of its tangent vector cannot change, as

$$\frac{d}{dt} \langle \dot{c}, \dot{c} \rangle = 2 \left\langle \frac{D\dot{c}}{dt}, \dot{c} \right\rangle = 0.$$

Example 1.6. Interpreting the x^0-coordinate of Minkowski space as the time measured in some inertial frame, we see that timelike curves represent motions of particles such that

$$\left(\frac{dx^1}{dx^0} \right)^2 + \ldots + \left(\frac{dx^n}{dx^0} \right)^2 < 1.$$

It is assumed that units have been chosen so that 1 is the maximum allowed velocity for a material particle (the speed of light). Therefore, timelike curves represent motions of material particles. Timelike geodesics, on the other hand, represent straight line motions with constant speed, i.e. motions of **free** particles. In addition, the length

$$\tau(c) = \int_a^b |\dot{c}|\, dt$$

of a timelike curve $c : [a, b] \to M$ is interpreted as the **proper time** measured by the particle between events $c(a)$ and $c(b)$.

Null curves, on the other hand, represent motions at the speed of light, and null geodesics represent the motions of light rays.
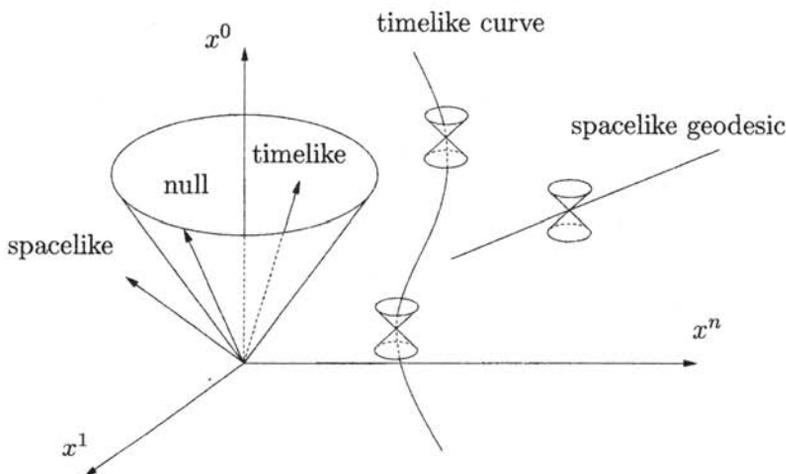

FIGURE 1. Minkowski space.

Einstein's great insight was realizing that to represent the gravitational field one must allow for general (curved) Lorentzian manifolds. The interpretations of timelike and null curves remain the same as in Minkowski space, except that timelike geodesics now represent **free-falling** material particles.

This first step, however, is just the easy half of the problem. To complete the theory, one needs to decide which Lorentzian manifolds represent actual (physical) gravitational fields. It took Einstein several years of hard work to find the answer.

Definition 1.7. A Lorentzian manifold (M, g) is said to satisfy the **Einstein equation** if its Ricci curvature tensor satisfies

$$Ric - \frac{S}{2}g = T,$$

where $S = \operatorname{tr} Ric$ is the scalar curvature and T is the **energy-momentum tensor** for a matter model in (M, g).

Example 1.8. Rather than describing in detail what a matter model is, we simply list the simplest examples:

 (i) **Vacuum:** Corresponds to taking $T = 0$.
 (ii) **Cosmological constant:** Corresponds to taking

$$T = -\Lambda g$$

 for some constant $\Lambda \in \mathbb{R}$.
 (iii) **Pressureless perfect fluid:** Is described by a **rest mass density function** $\rho \in C^\infty(M)$ and a timelike unit **velocity field** $U \in \mathfrak{X}(M)$ (whose integral lines are the motions of the fluid particles), and corresponds to the energy-momentum tensor

$$T = \rho U^\sharp \otimes U^\sharp$$

 ($U^\sharp$ is the 1-form associated to U by the metric).

Example 1.9. We now list the simplest solutions (M, g) of Einstein's field equation:

 (i) **Minkowski space (1907):** Is clearly a vacuum solution, describing an universe without gravitation.
 (ii) **Schwarzschild solution (1916):** Any vacuum solution admitting $O(n)$ as an isometry group is locally isometric to $M = \mathbb{R}^2 \times S^{n-1}$ with the metric

$$g = \left(1 - \left(\frac{r_s}{r}\right)^{n-2}\right) dt \otimes dt + \left(1 - \left(\frac{r_s}{r}\right)^{n-2}\right)^{-1} dr \otimes dr + r^2 h$$

 for some $r_s \in \mathbb{R}$, where h is the round metric in S^{n-1}. If $r_s > 0$ then r varies on either $(0, r_s)$ (**Schwarzschild interior**) or $(r_s, +\infty)$ (**Schwarzschild exterior**). These two regions can also be glued together by adding a **horizon**, where $r = r_s$, in which case one obtains a model of a **black hole**. Notice that in the Schwarzschild interior r is a time coordinate, and the integral lines of $\frac{\partial}{\partial r}$ are **incomplete** timelike geodesics, as as they cannot be continued past the curvature singularity at $r = 0$.

For $n \geq 3$ this metric admits circular orbits, i.e. geodesics of constant r coordinate traversing a great circle in S^{n-1}, and these satisfy

$$\left(\frac{d\varphi}{dt}\right)^2 = \frac{(n-2)r_s{}^{n-2}}{2r^n}$$

(φ being the angular coordinate in the great circle), which is exactly the condition for circular orbits around a point mass proportional to $r_s{}^{n-2}$ in Newtonian mechanics.

(iii) **de Sitter/Anti-de Sitter space (1917):** de Sitter space is simply the $(n+1)$-dimensional hyperboloid

$$-(x^0)^2 + (x^1)^2 + \ldots + x^{n+1} = \alpha^2$$

in $(n+2)$-dimensional Minkowski space, with the induced metric. Notice that this is the analogue of an Euclidean sphere, and is indeed a space of constant curvature. (As an aside, we remark that by reversing the sign of α^2 one obtains two copies of the $n+1$-dimensional hyperbolic space of radius α). de Sitter space is a solution of the Einstein equation with cosmological constant

$$\Lambda = \frac{n(n-1)}{2\alpha^2}.$$

The induced metric can be written as

$$g = \alpha^2 \left(-dt \otimes dt + \cosh t \, h\right),$$

where $x^0 = \alpha \sinh t$ and h is the round metric in S^n. Therefore one can think of de Sitter space as a spherical universe which contracts to a minimum radius α and then re-expands at an exponential rate. This cosmic repulsion is driven by the positive cosmological constant.

Anti-de Sitter space is the universal cover of the $(n+1)$-dimensional hyperboloid

$$(x^1)^2 + \ldots + (x^n)^2 - (x^{n+1})^2 - (x^{n+2})^2 = -\alpha^2$$

in $\mathbb{R}^{n+2}$ with the pseudo-Riemannian metric

$$dx^1 \otimes dx^1 + \ldots + dx^n \otimes dx^n - dx^{n+1} \otimes dx^{n+1} - dx^{n+2} \otimes dx^{n+2},$$

and is a solution of the Einstein equation with cosmological constant

$$\Lambda = -\frac{n(n-1)}{2\alpha^2}$$

(again being a space of constant curvature). The induced metric can be written as

$$g = \alpha^2 \left(-\cosh^2 \xi \, dt \otimes dt + d\xi \otimes d\xi + \sinh^2 \xi \, h \right)$$

$$= \frac{\alpha^2}{\cos^2 x} \left(-dt \otimes dt + dx \otimes dx + \sin^2 x \, h \right)$$

where $(t, \xi) \in \mathbb{R} \times (0, +\infty)$ are defined by

$$\begin{cases} x^{n+1} = \alpha \cosh \xi \cos t \\ x^{n+2} = \alpha \cosh \xi \sin t \end{cases},$$

$x \in \left(0, \frac{\pi}{2}\right)$ satisfies $\cos x = \frac{1}{\cosh \xi}$, and h is the round metric in S^{n-1}. Therefore one can think of de Sitter space as a static universe whose spatial sections are hyperbolic spaces (hence conformal to half spheres).

(iv) **Friedmann-Lemaître-Robertson-Walker models (1922):** If (Σ, h) is a n-dimensional manifold of constant curvature $k = -1, 0, 1$ (i.e. hyperbolic space, Euclidean space, S^n or quotients of these by discrete groups of isometries) then $(\mathbb{R} \times \Sigma, -dt \otimes dt + a^2(t)h)$ is a solution of the Einstein equation (representing an expanding or contracting universe) for a perfect fluid with velocity field $U = \frac{\partial}{\partial t}$ and rest mass density

$$\rho = \frac{n(n-1)\alpha}{a^n},$$

where the function $a(t)$ satisfies the first order ODE

$$\frac{\dot{a}^2}{2} - \frac{\alpha}{a^{n-2}} = -\frac{k}{2}$$

(for any constant $\alpha \in \mathbb{R}$). Notice that this is the equation of conservation of energy for a particle in the potential

$$V(a) = -\frac{\alpha}{a^{n-2}}.$$

Therefore, if $\alpha > 0 \Leftrightarrow \rho > 0$ then $a(t)$ blows up in finite time, corresponding to a curvature singularity (**Big Bang** or **Big Crunch**). In this case, the timelike geodesics given by the integral curves of $\frac{\partial}{\partial t}$ are **incomplete**.

Remark 1.10. Let (M, g) be a $(n+1)$-dimensional Lorentzian manifold. For $n = 1$ one has the identity

$$Ric - \frac{S}{2}g = 0,$$

and hence **any** 2-dimensional Lorentzian manifold is a vacuum solution of the Einstein equation (this being the only allowed matter model).

As is the case with all the fundamental equations of Physics, the Einstein equation can be derived from a variational principle. For instance, Lorentzian manifolds satisfying the vacuum Einstein equation are critical points of the **Einstein-Hilbert action**

$$\mathcal{A}[g] = \int_M S.$$

The fact that all 2-dimensional Lorentzian manifolds are vacuum solutions of the Einstein equation can thus be turned into a proof of the Gauss-Bonnet Theorem.

If $n \geq 2$ the trace of the Einstein equation yields,

$$S = -\frac{2\operatorname{tr} T}{n-1},$$

and hence the Einstein equation can be rewritten as

$$Ric = T - \frac{\operatorname{tr} T}{n-1}g.$$

In particular, any solution of the Einstein equation must be Ricci-flat at points where $T = 0$. This means that there is no gravitational field (i.e. curvature) in vacuum for $n = 2$.

2. Causality

We will now discuss the causal features of Lorentzian manifolds, a subject which has no parallel in Riemannian geometry.

Definition 2.1. Let (M, g) be a Lorentzian manifold and $p \in M$. Two timelike vectors $v, w \in T_pM$ are said to have the same (resp. opposite) **time orientation** if $\langle v, w \rangle < 0$ (resp. $\langle v, w \rangle > 0$).

Notice that $\langle v, w \rangle = 0$ cannot occur for timelike vectors.

Definition 2.2. A Lorentzian manifold (M, g) is said to be **time-orientable** if there exists a nonvanishing timelike vector field $T \in \mathfrak{X}(M)$.

This means that one can consistently time orient all tangent spaces T_pM. If (M, g) is time-orientable then it has exactly two time orientations: nonvanishing timelike vector fields $T, U \in \mathfrak{X}(M)$ define the same time orientation if and only if $\langle T, U \rangle < 0$. Timelike vectors $v \in T_pM$ with the same time orientation as T_p are said to be **future-pointing** for the time orientation determined by T. This easily extends to nonvanishing null vectors.

Example 2.3. The Möbius band $\mathbb{R} \times [0,1]/\sim$, where $\sim$ is the equivalence relation $(t,0) \sim (-t,1)$, admits the non-time-orientable Lorentzian metric

$$g = -dt \otimes dt + dx \otimes dx,$$

as well as the time-orientable Lorentzian metric $h = -g$.

The usual proof for the existence of an orientable double cover for a non-orientable manifold can be easily adapted to prove

Proposition 2.4. *Any non-time-orientable Lorentzian manifold* (M, g) *has a **time-orientable double cover**, i.e. a time-orientable Lorentzian manifold* $(\overline{M}, \overline{g})$ *and a local isometry* $\pi : \overline{M} \to M$ *such that every point in* M *has two preimages by* π.

A time-orientable Lorentzian manifold admits a nonvanishing vector field, and hence is either noncompact or has zero Euler characteristic. The same is true for a non-time-orientable Lorentzian manifold, for it must be true for its time-orientable double cover. On the other hand, if a differentiable manifold M is either noncompact or has zero Euler characteristic then it admits a nonvanishing vector field T, which we can assume to be of unit length for some Riemannian metric h on M. It is then easy to check that $g = -2T^\sharp \otimes T^\sharp + h$ is a Lorentzian metric on M. In other words:

Proposition 2.5. *A differentiable manifold* M *admits a Lorentzian metric if and only if is noncompact or has zero Euler characteristic.*

Definition 2.6. Let (M, g) be a time-oriented Lorentzian manifold.

  (i) A timelike curve $c : I \subset \mathbb{R} \to M$ is said to be **future-directed** if $\dot{c}$ is future-pointing.

 (ii) The **chronological future** of $p \in M$ is the set $I^+(p)$ of all points to which p can be connected by a future-directed timelike curve.

(iii) A **future-directed causal curve** is a curve $c : I \subset \mathbb{R} \to M$ such that $\dot{c}$ is non-spacelike and future-pointing (if nonzero).

(iv) The **causal future** of $p \in M$ is the set $J^+(p)$ of all points to which p can be connected by a future-directed causal curve.

One can make analogous definitions replacing "future" with "past". In general, the chronological and causal pasts and futures can be quite complicated sets, because of global features. Locally, however, causal properties are similar to those of Minkowski space. More precisely, we have the following statement:

Proposition 2.7. *Let* (M, g) *be a time-oriented Lorentzian manifold. Then each point* $p_0 \in M$ *has an open neighborhood* $V \subset M$ *such that the spacetime* (V, g) *obtained by restricting* g *to* V *satisfies:*

(1) If $p, q \in V$ then there exists a unique geodesic joining p to q (i.e. V is **geodesically convex**);

(2) $q \in I^+(p)$ iff there exists a future-directed timelike geodesic connecting p to q;

(3) $J^+(p) = \overline{I^+(p)}$;

(4) $q \in J^+(p)$ iff there exists a future-directed timelike or null geodesic connecting p to q.

Proof. The existence of geodesically convex neighborhoods holds for any affine connection (cf. [KN96]). Moreover, one can assume such neighborhoods to be **totally normal**, i.e. normal neighborhoods of all of their points.

To prove assertion (2), we start by noticing that if there exists a future-directed timelike geodesic connecting p to q then it is obvious that $q \in I^+(p)$. Suppose now that $q \in I^+(p)$; then there exists a future-directed timelike curve $c : [0, 1] \to V$ such that $c(0) = p$ and $c(1) = q$. Choose normal coordinates $(x^0, x^1, \ldots, x^n)$ given by the parametrization

$$\varphi(x^0, x^1, \ldots, x^n) = \exp_p(x^0 E_0 + x^1 E_1 + \ldots + x^n E_n),$$

where $\{E_0, E_1, \ldots, E_n\}$ is an orthonormal basis of $T_p M$ (with E_0 timelike and future-pointing). These are global coordinates in V, since V is totally normal. Defining

$$W_p(q) := -\left(x^0(q)\right)^2 + \left(x^1(q)\right)^2 + \ldots + \left(x^n(q)\right)^2$$

$$= \sum_{\mu,\nu=0}^{n} \eta_{\mu\nu} x^\mu(q) x^\nu(q)$$

(with $(\eta_{\mu\nu}) = \mathrm{diag}(-1, 1, \ldots, 1)$), we have to show that $W_p(q) < 0$. Let $W_p(t) := W_p(c(t))$. Since $x^\mu(p) = 0$ ($\mu = 0, 1, \ldots, n$), we have $W_p(0) = 0$. Setting $x^\mu(t) = x^\mu(c(t))$, we have

$$\dot{W}_p(t) = 2 \sum_{\mu,\nu=0}^{n} \eta_{\mu\nu} x^\mu(t) \dot{x}^\nu(t);$$

$$\ddot{W}_p(t) = 2 \sum_{\mu,\nu=0}^{n} \eta_{\mu\nu} x^\mu(t) \ddot{x}^\nu(t) + 2 \sum_{\mu,\nu=0}^{n} \eta_{\mu\nu} \dot{x}^\mu(t) \dot{x}^\nu(t),$$

and consequently (recalling that $\left(d\exp_p\right)_p = \mathrm{id}$)

$$\dot{W}_p(0) = 0;$$

$$\ddot{W}_p(0) = 2\langle \dot{c}(0), \dot{c}(0) \rangle < 0.$$

Therefore there exists $\varepsilon > 0$ such that $W_p(t) < 0$ for $t \in (0, \varepsilon)$.

The same proof as in Riemannian geometry shows that the unit tangent vector field to timelike geodesics through p is

$$X = -\operatorname{grad}(-W_p)^{\frac{1}{2}} = \frac{1}{2}\operatorname{grad} W_p\,(-W_p)^{-\frac{1}{2}}$$

(**Gauss Lemma**), where the gradient of a function is defined as in Riemannian geometry (notice however that in Lorentzian geometry a smooth function f **decreases** along the direction of $\operatorname{grad} f$ if $\operatorname{grad} f$ is timelike). Consequently $\operatorname{grad} W_p$ is tangent to timelike geodesics through p, being future-pointing on future-directed timelike geodesics.

Suppose that $W_p(t) < 0$. Then

$$\dot{W}(t) = \left\langle (\operatorname{grad} W_p)_{c(t)}, \dot{c}(t) \right\rangle < 0$$

as both $(\operatorname{grad} W_p)_{c(t)}$ and $\dot{c}(t)$ are timelike future-pointing. We conclude that we must have $W_p(t) < 0$ for all $t \in [0, 1]$. In particular, $W_p(q) = W_p(1) < 0$, and hence there exists a future-directed timelike geodesic connecting p to q.

Assertion (3) can be proved by using the global normal coordinates $(x^0, x^1, \ldots, x^n)$ of V to approximate causal curves by timelike curves. Once this is done, (4) is obvious from the fact that $\exp_p$ is a diffeomorphism onto V. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

A simple application of Proposition 2.7 is proving

Proposition 2.8. *Let (M, g) be a time-oriented Lorentzian manifold and $p \in M$. Then:*

(i) If $q \in I^+(p)$ and $r \in I^+(q)$ then $r \in I^+(p)$;
(ii) If $q \in J^+(p)$ and $r \in J^+(q)$ then $r \in J^+(p)$;
(iii) $I^+(p)$ is an open set.

In Lorentzian geometry there are no curves of minimal length, as any two points can be connected by piecewise null curves. However, there do exist curves with **maximal** length, and these are timelike geodesics. More precisely, we have the following statement:

Proposition 2.9. (Twin Paradox) *Let (M, g) be a time-oriented spacetime and $p_0 \in M$. Then there exists a geodesically convex open neighborhood $V \subset M$ of p_0 such that the spacetime (V, g) obtained by restricting g to V satisfies the following property: if $q \in I^+(p)$, c is the timelike geodesic connecting p to q and γ is any timelike curve connecting p to q, then $\tau(\gamma) \leq \tau(c)$, with equality iff γ is a is a reparametrization of c.*

Proof. Choose V as in the proof of Proposition 2.7. Any timelike curve $\gamma : [0,1] \to V$ satisfying $\gamma(0) = p$, $\gamma(1) = q$ can be written as

$$\gamma(t) = \exp_p(r(t)n(t)),$$

for $t \in [0,1]$, where $r(t) \geq 0$ and $\langle n(t), n(t) \rangle = -1$. We have

$$\dot{\gamma}(t) = (\exp_p)_* \left(\dot{r}(t)n(t) + r(t)\dot{n}(t) \right).$$

Since $\langle n(t), n(t) \rangle = -1$, we have $\langle \dot{n}(t), n(t) \rangle = 0$, and consequently $\dot{n}(t)$ is tangent to the level surfaces of the function $v \mapsto \langle v, v \rangle$. We conclude that

$$\dot{\gamma}(t) = \dot{r}(t)X_{\gamma(t)} + Y(t),$$

where X is the unit tangent vector field to timelike geodesics through p and $Y(t) = r(t)(\exp_p)_* \dot{n}(t)$ is tangent to the level surfaces of W_p (hence orthogonal to $X_{\gamma(t)}$). Consequently,

$$\tau(\gamma) = \int_0^1 \left| \langle \dot{r}(t)X_{\gamma(t)} + Y(t), \dot{r}(t)X_{\gamma(t)} + Y(t) \rangle \right|^{\frac{1}{2}} dt$$

$$= \int_0^1 \left(\dot{r}(t)^2 - |Y(t)|^2 \right)^{\frac{1}{2}} dt$$

$$\leq \int_0^1 \dot{r}(t)dt = r(1) = \tau(c),$$

(where we've used the facts that $\dot{r}(t) > 0$ for all $t \in [0,1]$, as $\dot{c}$ is future-pointing, and $\tau(c) = r(1)$, as $q = \exp_p(r(1)n(1))$. It should be clear that $\tau(\gamma) = \tau(c)$ if and only if $|Y(t)| \equiv 0 \Leftrightarrow Y(t) \equiv 0$ ($Y(t)$ is spacelike) for all $t \in [0,1]$, implying that n is constant. In this case, $\gamma(t) = \exp_p(r(t)n)$ is, up to reparametrization, the geodesic through p with initial condition $n \in T_pM$. $\qquad\Box$

Remark 2.10. Proposition 2.9 can be interpreted as follows: if two observers (e.g. two twins) meet at some event, are separated, and meet again at a later event, then the free-falling twin will always measure more time to have passed between the meetings.

For physical applications, it is important to demand that a Lorentzian manifold satisfies reasonable causality conditions. The simplest of these conditions excludes time travel, i.e. the possibility of a particle returning to an event in its past history.

Definition 2.11. A Lorentzian manifold (M, g) is said to satisfy the **chronology condition** if it does not contain closed timelike curves.

This condition is violated by compact Lorentzian manifolds:

Proposition 2.12. *Any compact Lorentzian manifold* (M, g) *contains closed timelike curves.*

Proof. Taking if necessary the time-orientable double cover, we can assume that (M, g) is time-oriented. It is easy to check that $\{I^+(p)\}_{p \in M}$ is an open cover of M. If M is compact, we can obtain a finite subcover $\{I^+(p_1), \ldots, I^+(p_N)\}$. Now if $p_1 \in I^+(p_i)$ for $i \neq 1$ then $I^+(p_1) \subset I^+(p_i)$, and we can exclude $I^+(p_1)$ from the subcover. Therefore, we can assume without loss of generality that $p_1 \in I^+(p_1)$, and hence there exists a closed timelike curve starting and ending at p_1. □

A stronger restriction on the causal behavior of the Lorentzian manifold is the following:

Definition 2.13. A Lorentzian manifold (M, g) is said to be **stably causal** if there exists a **global time function**, i.e. a smooth function $t : M \to \mathbb{R}$ such that $\mathrm{grad}(t)$ is timelike.

In particular, a stably causal Lorentzian manifold is time-orientable. We choose the time orientation defined by $-\mathrm{grad}(t)$, so that t increases along future-directed timelike curves. Notice that this implies that no closed timelike curves can exist, i.e. any stably causal Lorentzian manifold satisfies the chronology condition. In fact, any small perturbation of a causally stable Lorentzian manifold still satisfies the chronology condition.

Definition 2.14. Let (M, g) be a time-oriented Lorentzian manifold.

(i) A smooth future-directed causal curve $c : (a, b) \to M$ (with possibly $a = -\infty$ or $b = +\infty$) is said to be **future-inextendible** if $\lim_{t \to b} c(t)$ does not exist.

(ii) The **past domain of dependence** of $S \subset M$ is the set $D^-(S)$ of all points $p \in M$ such that any future-inextendible causal curve starting at p intersects S.

One can make analogous definitions replacing "future" with "past". The **domain of dependence** of S is simply the set $D(S) = D^+(S) \cup D^-(S)$.

Definition 2.15. A Lorentzian manifold (M, g) is said to be **globally hyperbolic** if it is stably causal and there exists a time function $t : M \to \mathbb{R}$ such that the time slices $S_a = t^{-1}(a)$ satisfy $D(S_a) = M$.

Example 2.16. The open set

$$U = \{(t, x) \in \mathbb{R}^2 \mid t \neq 0 \text{ or } x < 0\}$$

with the Minkowski metric

$$g = -dt \otimes dt + dx \otimes dx$$

is stably causal Lorentzian manifold (the coordinate $t : U \to \mathbb{R}$ is a global time function) which is not globally hyperbolic (cf. Figure 2). Notice that $J^+(-1, 0)$ is not closed. Moreover, the supremum of the lengths of timelike curves connecting $(-1, 0)$ to $(1, 0)$ is clearly 2, but it is not attained by any timelike curve.
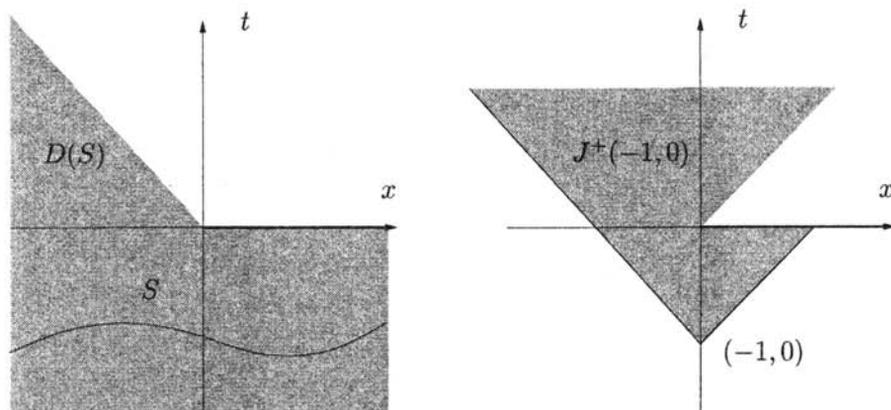

FIGURE 2. The open set U is not globally hyperbolic, and the causal future of $(-1, 0)$ is not closed.

3. Singularity Theorem

As we have seen in Example 1.9, both the Schwarzschild solution and the Friedmann-Lemaître-Robertson-Walker cosmological models display singularities, beyond which timelike geodesics cannot be continued.

Definition 3.1. A Lorentzian manifold (M, g) is **singular** if it is not geodesically complete.

One could think that the examples above are singular due to their high degree of symmetry, and that more realistic Lorentzian manifolds would be generically non-singular. We will show that this is not the case: any sufficiently small perturbation of these solutions will still be singular.

The question of whether a given Riemannian manifold is geodesically complete is settled by the Hopf-Rinow Theorem. Unfortunately, this theorem does not hold in Lorentzian geometry (essentially because one cannot use the metric to define a distance function).

Example 3.2.

(i) **Clifton-Pohl torus:** Consider the Lorentzian metric

$$\overline{g} = \frac{1}{u^2 + v^2}(du \otimes dv + dv \otimes du)$$

on $\overline{M} = \mathbb{R}^2 \setminus \{0\}$. The Lie group $\mathbb{Z}$ acts freely and properly on $\overline{M}$ by isometries through

$$n \cdot (u, v) = (2^n u, 2^n v),$$

and this determines a Lorentzian metric g on $M = \overline{M}/\mathbb{Z} \cong T^2$. A trivial calculation shows that there exist null geodesics satisfying $u \equiv 0$ and

$$\frac{d}{dt}\left(\frac{1}{v^2}\frac{dv}{dt}\right) = 0.$$

Since most solutions of this ODE blow up in finite time, we see that (M, g) is not geodesically complete (although M is compact).

(ii) **2-dimensional Anti-de Sitter space:** In spite of not solving the Einstein equation in 2 dimensions, we can still consider the unit 2-dimensional Anti-de Sitter space (M, g), i.e. the universal cover of the submanifold

$$\overline{M} = \{(u, v, w) \in \mathbb{R}^3 \mid u^2 + v^2 - w^2 = 1)\}$$

of $\mathbb{R}^3$ with the Lorentzian metric $\overline{g}$ induced by the pseudo-Riemannian metric

$$-du \otimes du - dv \otimes dv + dw \otimes dw.$$

As was seen in Example 1.9, there exist global coordinates $(t, x) \in \mathbb{R} \times (-\frac{\pi}{2}, \frac{\pi}{2})$ in M such that and

$$g = \frac{1}{\cos^2 x}(-dt \otimes dt + dx \otimes dx)$$

(notice the different range of the coordinate x, arising from the fact that $S^0 = \{-1, 1\}$). Therefore (M, g) is conformal to the open subset $\mathbb{R} \times (-\frac{\pi}{2}, \frac{\pi}{2})$ of Minkowski space, and its causality properties are the same. In particular, (M, g) is not globally hyperbolic.

As is the case with the Euclidean 2-sphere, the geodesics of $(\overline{M}, \overline{g})$ can be obtained by intersecting $\overline{M}$ with 2-planes through the origin. This fact can be used to prove that $(\overline{M}, \overline{g})$, and hence (M, g), are geodesically complete, and also that all timelike geodesics through the point p with coordinates $(0, 0)$ on (M, g) refocus on the points with coordinates $(\pm\pi, 0)$ (cf. Figure 3). On the other hand, spacelike geodesics through p are timelike geodesics of the Lorentzian metric $-g$, and hence are confined to the chronological future and past of

p in this metric. Therefore $\exp_p$ is not surjective, although (M, g) is geodesically complete.

Incidentally, if $\varepsilon > 0$ and q is the point with coordinates $(\pi + \varepsilon, 0)$ then there exist piecewise smooth causal curves connecting p to q with arbitrarily large length: simply take a future-directed null geodesic from p to the line $x = x_0$, a past-directed null geodesic from q the same line, and the portion of this line between the two geodesics. The resulting curve has length greater than $\frac{\varepsilon}{\cos(x_0)}$, and this can be made arbitrarily large by making $x_0 \to \frac{\pi}{2}$. These curves can be easily smoothed into timelike curves with arbitrarily large length.
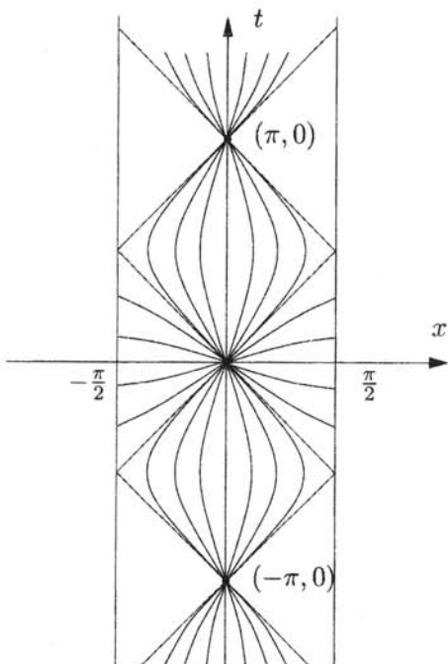

FIGURE 3. The exponential map is not surjective in 2-dimensional Anti-de Sitter spacetime.

We now proceed to show that geodesic incompleteness is a generic feature of Lorentzian manifolds satisfying the strong energy condition.

Definition 3.3. A Lorentzian manifold (M, g) is said to satisfy the **strong energy condition** if $Ric(V, V) \geq 0$ for any timelike vector field $V \in \mathfrak{X}(M)$.

If $n \geq 2$, this condition is equivalent (by the Einstein equation) to requiring that the energy-momentum tensor T satisfies

$$T(V,V) \geq \frac{\operatorname{tr} T}{n-1} \langle V, V \rangle$$

for any timelike vector field $V \in \mathfrak{X}(M)$, which turns out to be physically reasonable.

Example 3.4. Let us see the meaning of the strong energy condition for each of the matter models in Example 1.8:

(i) **Vacuum:** Trivially satisfied.
(ii) **Cosmological constant:** Equivalent to $\Lambda \leq 0$.
(iii) **Pressureless perfect fluid:** Becomes

$$\rho \left(\langle U, V \rangle^2 + \frac{1}{n-1} \langle V, V \rangle \right) \geq 0,$$

or, since the term in brackets is easily seen to be non-negative, simply $\rho \geq 0$.

Definition 3.5. Let (M, g) be a globally hyperbolic Lorentzian manifold and S a time slice with future-pointing normal vector field n. Let c_p be the timelike geodesic with initial condition n_p for each point $p \in S$. The **exponential map** $\exp : U \to M$ (defined on an open neighborhood U of $\{0\} \times S \subset \mathbb{R} \times S$) is the map $\exp(t, p) = c_p(t)$. The critical values of $\exp$ are said to be **conjugate points** to S.

If the geodesic c_p has no conjugate points between $c_p(0) = p$ and $c_p(t_0)$ then there exists an open neighborhood V of $c_p([0, t_0])$ which can be foliated by images of geodesics orthogonal to S. The tangent vectors to these geodesics yield a unit timelike vector field $X \in \mathfrak{X}(V)$, which by the Gauss Lemma satisfies $X = -\operatorname{grad} t$, where $t : V \to \mathbb{R}$ is the distance along the geodesics. Therefore $X^\sharp = -dt$, and the covariant derivative $K = \nabla X^\sharp$ is a symmetric tensor, which by the geodesic equation must satisfy $\iota_X \nabla X^\sharp = \nabla_X X^\sharp = 0$.

Definition 3.6. The divergence $\theta = \operatorname{div} X = \operatorname{tr} K$ of the vector field $X \in \mathfrak{X}(V)$ is called the **expansion** of the family of timelike geodesics on V.

Remark 3.7. K is actually the extrinsic curvature of the family of hypersurfaces of constant t, and θ gives the logarithmic rate of change of their volume element along X (as is easily seen from the Divergence Theorem).

Proposition 3.8. Let (M, g) be a globally hyperbolic Lorentzian manifold satisfying the strong energy condition, S a time slice and $p \in S$ a point

where $\theta = \theta_0 < 0$. If the geodesic c_p can be extended to a distance $t_0 = -\frac{n}{\theta_0}$ to the future of S, then it contains at least a point conjugate to S.

Proof. Suppose that there were no points in $c_p([0, t_0])$ conjugate to S. Then there would exist an open neighborhood V of $c_p([0, t_0])$ as above. An easy calculation shows that the **Raychaudhuri equation**

$$X(\theta) + \operatorname{tr} K^2 + Ric(X, X) = 0$$

holds. Since (M, g) satisfies the strong energy condition, we have $R(X, X) \geq 0$, and hence

$$X(\theta) + \operatorname{tr} K^2 \leq 0$$

The Cauchy-Schwarz inequality implies that

$$(\operatorname{tr} A)^2 \leq n \operatorname{tr}(A^t A),$$

for any $n \times n$ matrix A. Since K is symmetric and vanishes on X, we have

$$\operatorname{tr} K^2 \geq \frac{1}{n} \theta^2.$$

Setting $\theta(t) := \theta(c_p(t))$, we see that

$$\frac{d\theta}{dt} + \frac{1}{n} \theta^2 \leq 0.$$

Integrating this inequality one obtains

$$\frac{1}{\theta} \geq \frac{1}{\theta_0} + \frac{t}{n},$$

and hence θ must blow up at a value of t no greater than $t_0 = -\frac{n}{\theta_0}$. This yields a contradiction, as θ is smooth function on V. $\square$

Proposition 3.9. *Let (M, g) be a globally hyperbolic Lorentzian manifold, S a time slice, $p \in M$ and c a timelike geodesic through p orthogonal to S. If there exists a conjugate point between S and p then c does not maximize length (among the timelike curves connecting S to p).*

Proof. We will provide a sketch of the proof, which is similar to its analogue in Riemannian geometry. Let q be a conjugate point along c between S and p. Then there exists another geodesic $\tilde{c}$, orthogonal to S, with the same (approximate) length, which (approximately) intersects c at q. Let V be a geodesically convex neighborhood of q, $r \in V$ a point along $\tilde{c}$ between S and q, and $s \in V$ a point along c between q and p (cf. Figure 4). Then the piecewise smooth timelike curve obtained by following $\tilde{c}$ between S and r, the unique geodesic in V between r and s, and c between s and p connects S to p and has strictly bigger length than c (by the Twin Paradox). This curve can be easily smoothed while retaining bigger length than c. $\square$
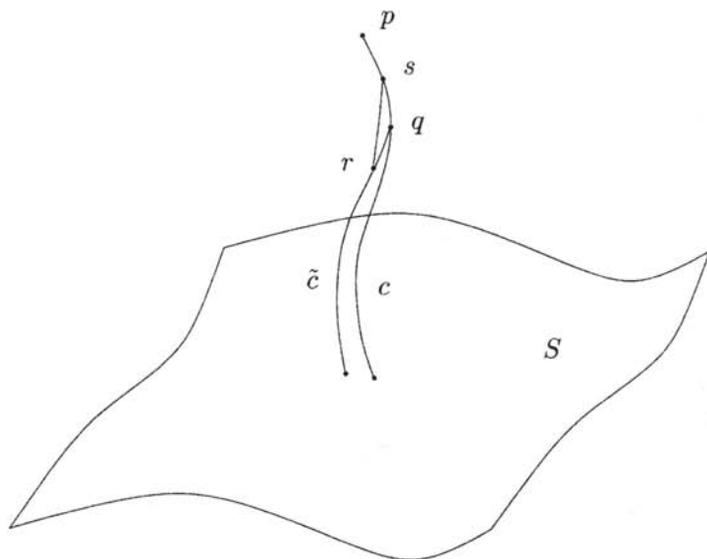
FIGURE 4. Proof of Proposition 3.9.

Remark 3.10. As in Riemannian geometry, a the critical values of $\exp_p$ are said to be **conjugate** to p. Essentially the same proof as that of Proposition 3.9 shows that if c is a timelike geodesic connecting p to q and there exists a conjugate point between p and q then c does not maximize length (among the timelike curves connecting p to q). An example is provided by the points p and q with coordinates $(0,0)$ and $(\pi + \varepsilon, 0)$ in 2-dimensional anti-de Sitter space (cf. Example 3.2(ii)). As another interesting example, consider circular motions with constant angular velocity in the Schwarzschild solution (not necessarily geodesics). From the condition that the tangent vector must be a unit timelike vector,

$$-\left(1 - \left(\frac{r_s}{r}\right)^{n-2}\right)\left(\frac{dt}{d\tau}\right)^2 + r^2\left(\frac{d\varphi}{d\tau}\right)^2 = -1,$$

one easily obtains

$$\frac{d\tau}{dt} = \left(\left(1 - \left(\frac{r_s}{r}\right)^{n-2}\right) - r^2\left(\frac{d\varphi}{dt}\right)^2\right)^{\frac{1}{2}}$$

and hence the proper time along one of these motions is

$$\Delta\tau = \left(\left(1 - \left(\frac{r_s}{r}\right)^{n-2}\right) - r^2\left(\frac{d\varphi}{dt}\right)^2\right)^{\frac{1}{2}}\Delta t.$$

Notice that $\Delta\tau$ decreases as $\left|\frac{d\varphi}{dt}\right|$ increases. A circular orbit is obtained by setting

$$\left(\frac{d\varphi}{dt}\right)^2 = \frac{(n-2)r_s{}^{n-2}}{2r^n}.$$

Consider two points on a circular orbit more than half an orbit away. Then the (non-geodesic) circular motion in the opposite direction passing through the same points has a smaller value of $\left|\frac{d\varphi}{dt}\right|$, and hence a larger value of $\Delta\tau$. This is to be expected, as the midway point of the circular orbit is clearly conjugate to the starting point: all circular geodesics through the starting point (in different orbital planes) have the same midway point.

Proposition 3.11. *Let (M, g) be a globally hyperbolic Lorentzian manifold, S a time slice and $p \in D^+(S)$. Then $D^+(S) \cap J^-(p)$ is compact.*

Proof. Let us define a **simple neighborhood** $U \subset M$ to be a geodesically convex open set diffeomorphic to an open ball whose boundary is a compact submanifold of a geodesically convex open set (therefore ∂U is diffeomorphic to S^n and $\overline{U}$ is compact). It is clear that simple neighborhoods form a basis for the topology of M. Also, it is easy to show that any open cover $\{V_\alpha\}_{\alpha \in A}$ has a countable, locally finite refinement $\{U_k\}_{k \in \mathbb{N}}$ by simple neighborhoods.

If $A = D^+(S) \cap J^-(p)$ were not compact, there would exist a countable, locally finite open cover $\{U_k\}_{k \in \mathbb{N}}$ of A by simple neighborhoods not admitting any finite subcover. Take $q_k \in A \cap U_k$ such that $q_k \neq q_l$ for $k \neq l$. The sequence $\{q_k\}_{k \in \mathbb{N}}$ cannot have accumulation points, for any point in M has a neighborhood intersecting only finite simple neighborhoods U_k. Consequently, each simple neighborhood U_k contains only finite points in the sequence (as $\overline{U_k}$ is compact).

Set $p_1 = p$. Since $p_1 \in A$, we have $p_1 \in U_{k_1}$ for some $k_1 \in \mathbb{N}$. Let $q_k \notin U_{k_1}$. Since $q_k \in J^-(p_1)$, there exists a future-directed causal curve c_k connecting q_k to p_1. This curve will necessarily intersect ∂U_{k_1}. Let $r_{1,k}$ be an intersection point. Since U_{k_1} contains only finite points in the sequence $\{q_k\}_{k \in \mathbb{N}}$, there will exist infinite intersection points $r_{1,k}$. As ∂U_{k_1} is compact, these will accumulate to some point $p_2 \in \partial U_{k_1}$.

Because $\overline{U_{k_1}}$ is contained in a geodesically convex open set, $p_2 \in J^-(p_1)$: if $\gamma_{1,k}$ is the unique causal geodesic connecting p_1 to $r_{1,k}$, parametrized by the global time function $t : M \to \mathbb{R}$ such that $S = t^{-1}(0)$, then the subsequence of $\{\gamma_{1,k}\}$ corresponding to a convergent subsequence of $\{r_{1,k}\}$ will converge (in the C^∞ topology) to a causal geodesic γ_1 connecting p_1 to p_2. Since $t(r_{1,k}) \geq 0$, we have $t(p_2) \geq 0$, and therefore $p_2 \in A$. Since $p_2 \notin U_{k_1}$, there must exist $k_2 \in \mathbb{N}$ such that $p_2 \in U_{k_2}$.

Since U_{k_2} contains only finite points in the sequence $\{q_k\}_{k \in \mathbb{N}}$, infinite curves c_k must intersect ∂U_{k_2} to the past of $r_{1,k}$. Let $r_{2,k}$ be the intersection points. As ∂U_{k_2} is compact, $\{r_{2,k}\}$ must accumulate to some point $p_3 \in \partial U_{k_2}$. Because $\overline{U_{k_2}}$ is contained in a geodesically convex open set, $p_3 \in J^-(p_2)$: if $\gamma_{2,n}$ is the unique causal geodesic connecting $r_{1,k}$ to $r_{2,k}$, parametrized by the global time function, then the subsequence of $\{\gamma_{2,k}\}$ corresponding to convergent subsequences of both $\{r_{1,k}\}$ and $\{r_{2,k}\}$ will converge to a causal geodesic connecting p_2 to p_3. Since $J^-(p_2) \subset J^-(p_1)$, $p_2 \in A$.

Iterating the procedure above, we can construct a sequence $\{p_i\}_{i \in \mathbb{N}}$ of points in A satisfying $p_i \in U_{n_i}$ with $n_i \neq n_j$ if $i \neq j$, such that p_i is connected p_{i+1} by a causal geodesic γ_i. It is clear that γ_i cannot intersect S, for $t(p_{i+1}) > t(p_{i+2}) \geq 0$. On the other hand, the piecewise smooth causal curve obtained by joining the curves γ_i can easily be smoothed into a past-directed causal curve starting at p_1 which does not intersect S. Finally, such curve is inextendible: it cannot converge to any point, as $\{p_i\}_{i \in \mathbb{N}}$ cannot accumulate. But since $p_1 \in D^+(S)$, this curve would have to intersect S. Therefore A must be compact. $\square$

Corollary 3.12. *Let (M,g) be a globally hyperbolic Lorentzian manifold and $p, q \in M$. Then:*

(i) $J^+(p)$ is closed;
(ii) $J^+(p) \cap J^-(q)$ is compact.

Proposition 3.11 is a key ingredient in establishing the following fundamental result:

Theorem 3.13. *Let (M,g) be a globally hyperbolic Lorentzian manifold with time slice S, and $p \in D^+(S)$. Then among all timelike curves connecting p to M there exists a timelike curve with maximal length. This curve is a timelike geodesic, orthogonal to S.*

Proof. Consider the set $T(S,p)$ of all timelike curves connecting S to p. Since we can always use the global time function $t : M \to \mathbb{R}$ such that $S = t^{-1}(0)$ as a parameter, these curves are determined by their images, which are compact subsets of the compact set $A = D^+(S) \cap J^-(p)$. As is well known (cf. [Mun00]), the set $C(A)$ of all compact subsets of A is a compact metric space for the **Hausdorff metric** d_H, defined as follows: if $d : M \times M \to \mathbb{R}$ is a metric yielding the topology of M,

$$d_H(K, L) = \inf\{\varepsilon > 0 \mid K \subset U_\varepsilon(L) \text{ and } L \subset U_\varepsilon(K)\},$$

where $U_\varepsilon(K)$ is a ε-neighborhood of K for the metric d. Therefore, the closure $C(S,p) = \overline{T(S,p)}$ is a compact subset of $C(A)$. It is not difficult
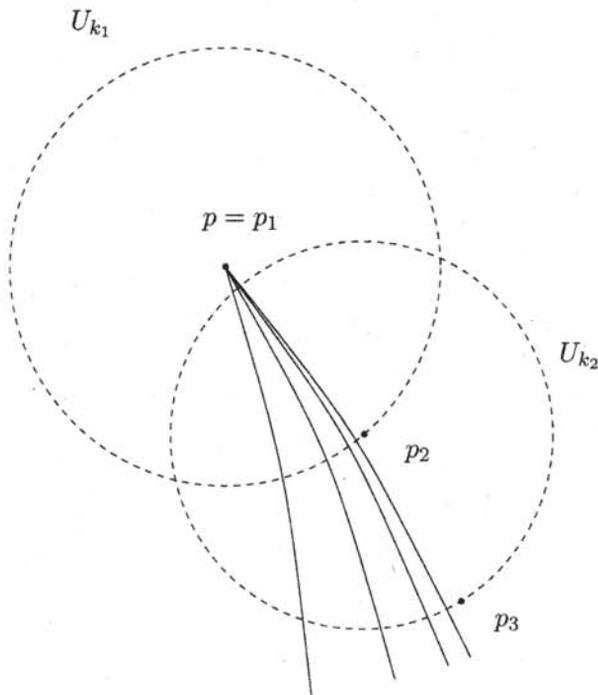
FIGURE 5. Proof of Proposition 3.11.

to show that $C(S, p)$ can be identified with the set of **continuous causal curves** connecting S to p (a continuous curve $c : [0, t(p)] \to M$ is said to be **causal** if $c(t_2) \in J^+(c(t_1))$ whenever $t_2 > t_1$).

The length function $\tau : T(S, p) \to \mathbb{R}$ is defined by

$$\tau(c) = \int_0^{t(p)} |\dot{c}(t)| dt.$$

This function is **upper semicontinuous**, i.e. continuous for the topology

$$\mathcal{O} = \{(-\infty, a) \mid -\infty \leq a \leq +\infty\}$$

in $\mathbb{R}$. Indeed, let $c \in T(S, p)$ be parametrized by its arclength $\mathcal{T}$. For sufficiently small $\varepsilon > 0$, the function $\mathcal{T}$ can be extended to the ε-neighborhood $U_\varepsilon(c)$ in such a way that its level hypersurfaces are spacelike and orthogonal to c (i.e. $-\operatorname{grad} \mathcal{T}$ is timelike and coincides with $\dot{c}$ on c), and $S \cap U_\varepsilon(c)$ is one of these surfaces. If $\gamma \in T(S, p)$ is in the open ball $B_\varepsilon(c) \subset C(A)$ then we can use $\mathcal{T}$ as a parameter, thus obtaining

$$\dot{\gamma} \cdot \mathcal{T} = 1 \Leftrightarrow \langle \dot{\gamma}, \operatorname{grad} \mathcal{T} \rangle = 1.$$

Therefore $\dot{\gamma}$ can be decomposed as

$$\dot{\gamma} = \frac{1}{\langle \operatorname{grad} T, \operatorname{grad} T \rangle} \operatorname{grad} T + X$$

where X is spacelike and orthogonal to $\operatorname{grad} T$. Consequently,

$$\tau(\gamma) = \int_0^{\tau(c)} |\dot{\gamma}| \, dT = \int_0^{\tau(c)} \left| \frac{1}{\langle \operatorname{grad} T, \operatorname{grad} T \rangle} + \langle X, X \rangle \right|^{\frac{1}{2}} dT.$$

Given $\delta > 0$, we can choose $\varepsilon > 0$ sufficiently small so that

$$-\frac{1}{\langle \operatorname{grad} T, \operatorname{grad} T \rangle} < \left(1 + \frac{\delta}{\tau(c)} \right)^2$$

on the ε-neighborhood $U_\varepsilon(c)$ (as $\langle \operatorname{grad} T, \operatorname{grad} T \rangle = -1$ on c). Consequently,

$$\tau(\gamma) = \int_0^{\tau(c)} \left| -\frac{1}{\langle \operatorname{grad} T, \operatorname{grad} T \rangle} - \langle X, X \rangle \right|^{\frac{1}{2}} dT$$

$$< \int_0^{\tau(c)} \left(1 + \frac{\delta}{\tau(c)} \right) dT = \tau(c) + \delta,$$

proving upper semicontinuity. As a consequence, the length function and can be extended to $C(S, p)$ through

$$\tau(c) = \lim_{\varepsilon \to 0} \sup \{ \tau(\gamma) \mid \gamma \in B_\varepsilon(c) \cap T(S, p) \}$$

(as for $\varepsilon > 0$ sufficiently small the supremum will be finite). Also, it is clear that if $c \in T(S, p)$ then the upper semicontinuity of the length forces the two definitions of $\tau(c)$ to coincide. The extension of the length function to $C(S, p)$ is trivially upper semicontinuous: given $c \in C(S, p)$ and $\delta > 0$, let $\varepsilon > 0$ be such that $\tau(\gamma) < \tau(c) + \frac{\delta}{2}$ for any $\gamma \in B_{2\varepsilon}(c) \cap T(S, p)$. Then it is clear that $\tau(c') < \tau(c) + \delta$ for any $c' \in B_\varepsilon(c)$.

Finally, we notice that the compact sets of $\mathbb{R}$ for the topology $\mathcal{O}$ are sets with maximum. Therefore, the length function attains a maximum at some point $c \in C(S, p)$. All that remains to be seen is that the maximum is also attained at a smooth timelike curve γ. To do so, cover c with finitely many geodesically convex neighborhoods and choose points $p_1, \ldots, p_m$ in c such that $p_1 \in S$, $p_m = p$ and the portion of c between p_{i-1} and p_i is contained in a geodesically convex neighborhood for all $i = 2, \ldots, m$. It is clear that there exists a sequence $c_k \in T(S, p)$ such that $c_k \to c$ and $\tau(c_k) \to \tau(c)$. Let $t_i = t(p_i)$ and $p_{i,k}$ be the intersection of c_k with $t^{-1}(t_i)$. Replace c_k by the sectionally geodesic curve γ_k obtained by joining $p_{i-1,k}$ to $p_{i,k}$ in the corresponding geodesically convex neighborhood. Then $\tau(\gamma_k) \geq \tau(c_k)$, and therefore $\tau(\gamma_k) \to \tau(c)$. Since each sequence $p_{i,k}$ converges to p_i, γ_k

converges to the sectionally geodesic curve γ obtained by joining p_{i-1} to p_i $(i = 2, \ldots, m)$, and it is clear that $\tau(\gamma_k) \to \tau(\gamma) = \tau(c)$. Therefore γ is a point of maximum for the length. Finally, we notice that γ must be smooth at the points p_i, for otherwise we could increase its length by using the Twin Paradox. Therefore γ must be a timelike geodesic. Using the Gauss Lemma, it is clear that γ must be orthogonal to S, for otherwise it would be possible to increase its length. $\qquad\Box$

Remark 3.14. Essentially the same proof can be used to show that if (M, g) is globally hyperbolic and $p, q \in M$ with $q \in I^+(M)$ then among all timelike curves connecting p to q there exists a timelike curve with maximal length, which is a timelike geodesic (note that the Lorentzian manifolds in Examples 2.16 and 3.2(ii) are not globally hyperbolic). Thus, in a way, global hyperbolicity is the Lorentzian analogue of completeness in Riemannian geometry.

We have now all the necessary ingredients to prove the singularity theorem:

Theorem 3.15. *Let (M, g) be a globally hyperbolic Lorentzian manifold satisfying the strong energy condition, and suppose that the expansion satisfies $\theta \leq \theta_0 < 0$ on a time slice S. Then (M, g) is singular.*

Proof. We will show that no future-directed timelike geodesic orthogonal to S can be extended to proper time greater than $\tau_0 = -\frac{n}{\theta_0}$ to the future of S. Suppose that this was not so. Then there would exist a future-directed timelike geodesic c orthogonal to S defined in an interval $[0, \tau_0 + \varepsilon]$ for some $\varepsilon > 0$. Let $p = c(\tau_0 + \varepsilon)$. According to Theorem 3.13, there would exist a timelike geodesic γ with maximal length connecting S to p, orthogonal to S. Because $\tau(c) = \tau_0 + \varepsilon$, we would necessarily have $\tau(\gamma) \geq \tau_0 + \varepsilon$. Proposition 3.8 guarantees that γ would develop a conjugate point at a distance of at most τ_0 to the future of S, and Proposition 3.9 states that γ would cease to be maximizing beyond this point. Therefore we arrive at a contradiction. $\qquad\Box$

Remark 3.16. It should be clear that (M, g) is singular if the condition $\theta \leq \theta_0 < 0$ on a time slice S is replaced by the condition $\theta \geq \theta_0 > 0$ on S. In this case, no **past-directed** timelike geodesic orthogonal to S can be extended to proper time greater than $\tau_0 = \frac{n}{\theta_0}$ to the **past** of S.

Remark 3.17. The proof of Theorem 3.15 does not hold in Riemannian geometry, where geodesics are length **minimizing** curves. Using similar ideas, one can prove the following: if (M, g) is a complete Riemannian manifold whose Ricci curvature satisfies $Ric \geq \varepsilon g$ for some $\varepsilon > 0$ then M is compact.

Example 3.18.

(1) Theorem 3.15 does not apply to Minkowski, de Sitter or anti-de Sitter spaces: the first does not contain a time slice with expansion bounded away from zero, the second does not satisfy the strong energy condition and the third is not globally hyperbolic. It does apply to the **Milne universe**, which is the open set U of Minkowski space defined by

$$x^0 > 0 \quad \text{and} \quad -\left(x^0\right)^2 + \left(x^1\right)^2 + \ldots + \left(x^n\right)^2 < 0$$

(cf. Figure 6). It is easily seen that the function

$$\tau\left(x^0, x^1, \ldots, x^n\right) = \sqrt{\left(x^0\right)^2 - \left(x^1\right)^2 - \ldots - \left(x^n\right)^2}$$

is a global time function in U whose time slices have domain of dependence U and constant positive expansion $n\tau$ (they are, in fact, isometric to n-dimensional hyperbolic spaces of radii τ). Here the geodesic incompleteness guaranteed by Theorem 3.15 is not due to curvature singularities.

(2) The Friedmann-Lemaître-Robertson-Walker models are globally hyperbolic, and for $\alpha > 0$ satisfy the strong energy condition (as $\rho > 0$, cf. Example 1.9). Furthermore, one easily checks that the expansion of the time slices is

$$\theta = \frac{n\dot{a}}{a}.$$

Assume that the model is expanding at time t_0. Then $\theta = \theta_0 = \frac{n\dot{a}(t_0)}{a(t_0)}$ on the time slice $S = \{t = t_0\}$, and hence Theorem 3.15 guarantees that this model is singular to the past of S (i.e. there exists a Big Bang). Furthermore, Theorem 3.15 implies that this singularity is generic: any sufficiently small perturbation of an expanding Friedmann-Robertson-Walker model satisfying the strong energy condition will also be singular. Loosely speaking, any expanding universe must have begun at a Big Bang.

(3) The region $\{r < r_s\}$ of the Schwarzschild solution is globally hyperbolic, and satisfies the strong energy condition (as $Ric = 0$). The metric can be written is this region as

$$g = -d\tau \otimes d\tau + \left(\left(\frac{r_s}{r}\right)^{n-2} - 1\right) dt \otimes dt + r^2 h$$

where h is the round metric in S^{n-1} and

$$\tau = \int_r^{r_s} \left(\left(\frac{r_s}{u}\right)^{n-2} - 1\right)^{-\frac{1}{2}} du.$$

Therefore the inside of a black hole can be pictured as a cylinder $\mathbb{R} \times S^{n-1}$ whose shape is evolving in time: as $r \to 0$, the sphere S^{n-1} contracts to a singularity, with the t-direction expanding. The expansion of the time slice $S = \{\tau = \tau_0\} = \{r = r_0\}$ can be computed to be

$$\theta = \left(\left(\frac{r_s}{r} \right)^{n-2} - 1 \right)^{-\frac{1}{2}} \left(\frac{n-1}{r} - \frac{n}{2r} \left(\frac{r_s}{r} \right)^{n-2} \right).$$

Therefore $\theta = \theta_0 < 0$ for r_0 sufficiently small, and hence Theorem 3.15 guarantees that the Schwarzschild solution is singular to the future of S. Moreover, Theorem 3.15 implies that this singularity is generic: any sufficiently small perturbation of the Schwarzschild solution satisfying the strong energy condition will also be singular. Loosely speaking, once the collapse has advanced long enough, nothing can prevent the formation of a singularity.
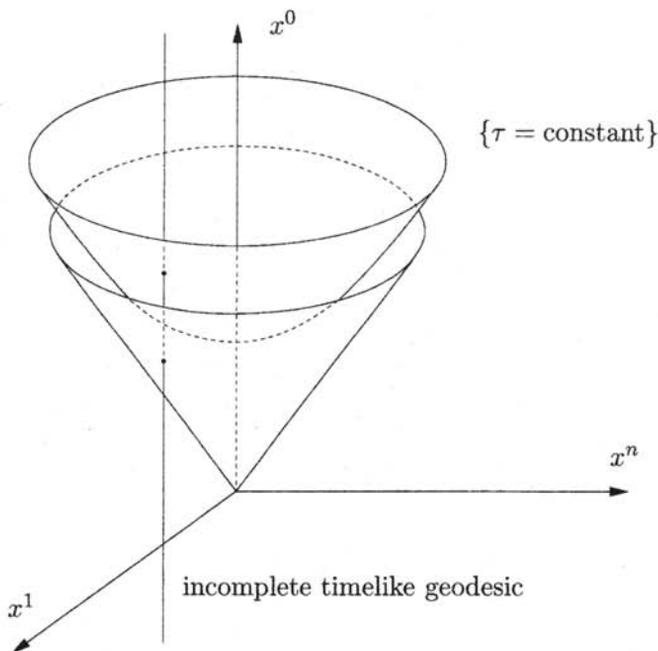

FIGURE 6. Milne universe.

References

[BEE96] J. Beem, P. Ehrlich, and K. Easley, *Global lorentzian geometry*, Marcel Dekker, 1996.

[Bra01] H Bray, *Proof of the riemannian penrose inequality using the positive mass theorem*, J. Diff. Geom. **59** (2001), 177–267.

[dC93] M. do Carmo, *Riemannian geometry*, Birkhäuser, 1993.

[GHL04] S. Gallot, D. Hulin, and J. Lafontaine, *Riemannian geometry*, Springer, 2004.

[Haw67] S. Hawking, *The occurrence of singularities in cosmology. iii. causality and singularities*, Proc. Roy. Soc. Lon. A **300** (1967), 187–201.

[HE95] S. Hawking and G. Ellis, *The large scale structure of space-time*, Cambridge University Press, 1995.

[HI01] G. Huisken and T. Ilmanen, *The inverse mean curvature flow and the riemannian penrose inequality*, J. Diff. Geom. **59** (2001), 353–437.

[HP70] S. Hawking and R. Penrose, *The singularities of gravitational collapse and cosmology*, Proc. Roy. Soc. Lon. A **314** (1970), 529–548.

[KN96] S. Kobayashi and K. Nomizu, *Foundations of differential geometry*, vol. I and II, Wiley, 1996.

[MTW73] C. Misner, K. Thorne, and J. A. Wheeler, *Gravitation*, Freeman, 1973.

[Mun00] J. Munkres, *Topology*, Prentice-Hall, 2000.

[Nab88] G. Naber, *Spacetime and singularities – an introduction*, Cambridge University Press, 1988.

[O'N83] B. O'Neill, *Semi-riemannian geometry*, Academic Press, 1983.

[Pen65] R. Penrose, *Gravitational collapse and space-time singularities*, Phys. Rev. Lett. **14** (1965), 57–59.

[Pen87] ———, *Techniques of differential topology in relativity*, Society for Industrial and Applied Mathematics, 1987.

[Sch02] B. Schutz, *A first course in general relativity*, Cambridge University Press, 2002.

[SY79] R. Schoen and S.-T. Yao, *Proof of the positive mass theorem*, Comm. Math. Phys. **65** (1979), 45–76.

[SY81] ———, *Proof of the positive mass theorem ii*, Comm. Math. Phys. **79** (1981), 231–260.

[Wal84] R. Wald, *General relativity*, University of Chicago Press, 1984.

DEPARTMENT OF MATHEMATICS, INSTITUTO SUPERIOR TÉCNICO, PORTUGAL