University of Helsinki
Faculty of Science
Department of Mathematics and Statistics

Master's Thesis

Penrose's singularity theorem

Miika Sarkkinen

Master's Programme in Mathematics and Statistics

Supervisor: Lauri Oksanen October 2023

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Mathematics and Statistics

Tekijä — Författare — Author
Miika Sarkkinen

Työn nimi — Arbetets titel — Title

Penrose's singularity theorem

Oppiaine — Läroämne — Subject		
Mathematics		
Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's Thesis	October 2023	30 s.

Tiivistelmä — Referat — Abstract

In this thesis we present and prove Roger Penrose's singularity theorem, which is a fundamental result in mathematical general relativity. In 1965 Penrose showed that in Einstein's theory of general relativity, under certain general assumptions on the topology, curvature, and causal structure of a Lorentzian spacetime manifold, the spacetime manifold is null geodesically incomplete. At the time, Penrose's theorem was highly topical in a longstanding debate on the question whether singularities are formed in the process of gravitational collapse. In the proof of the theorem, novel mathematical techniques were introduced in the study of Einstein's theory of gravity, leading to further important developments in the mathematics of general relativity.

Penrose's theorem is built on the methods of semi-Riemannian geometry, in particular Lorentzian geometry. To lay the basis for later constructions, we therefore review the basic concepts and results of semi-Riemannian geometry needed in order to understand Penrose's theorem. The discussion includes semi-Riemannian metrics, connection, curvature, geodesics, and semi-Riemannian submanifolds.

Second, calculus of variations on semi-Riemannian manifolds is introduced and a set of results pertinent to Penrose's theorem is given. The notion of focal point of a spacelike submanifold is defined and a proposition stating sufficient conditions for the existence of focal points is presented. Furthermore, we give a series of results that establish a relation between focal points of spacelike submanifolds and causality on a Lorentzian manifold.

In the last chapter, we define a family of concepts that can be used to analyze the causal structure of Lorentzian manifolds. In particular, we define the notions of global hyperbolicity, Cauchy hypersurface, and trapped surface, which are central to Penrose's theorem, and show some important properties thereof. Finally, Penrose's theorem is stated and proved in detail.

Avainsanat — Nyckelord — Keywords
Semi-Riemannian geometry, general relativity, trapped surface, null geodesic incompleteness

Säilytyspaikka — Förvaringsställe — Where deposited
E-thesis, Kumpula Campus Library

Muita tietoja — Övriga uppgifter — Additional information

Contents

Chapter 1

Introduction

In general relativity, gravitation is described in terms of curvature of spacetime given by a metric tensor on a smooth manifold. In contrast to Newtonian gravity, there is no *a priori* geometry that forms a fixed background for gravitating objects; instead, the geometry of spacetime is determined in a dynamical interplay between matter and curvature, governed by Einstein's field equations. One of the drastic consequences of spacetime being dynamical is the inevitability of gravitational collapse in general relativity: if energy density surpasses certain critical value, matter collapses under its own gravitational pull and curvature grows unboundedly. As a result, spacetime becomes singular, with ordinary laws of physics breaking down at the singularity. This is expected to happen, for example, in the gravitational collapse of a sufficiently large dying star. A central mathematical result that makes this scenario a robust prediction of general relativity is Penrose's singularity theorem, which is the topic of this thesis.

Singular black hole geometries have been known since the early days of general relativity. One of the first exact solutions found for the vacuum Einstein equations was the Schwarzschild metric, which describes a spherically symmetric empty space around a central mass. As the radial coordinate tends to zero, certain curvature invariants built from the Schwarzschild metric (*e.g.*, the Kretschmann scalar) grow without limit. The Schwarzschild black hole thus has a curvature singularity at its core. However, as a static, time-symmetric solution of Einstein's equations, it is not a full description of spacetime formed in a dynamical process, like gravitational collapse.

In 1939, J.R. Oppenheimer and H. Snyder [OS39] gave a model of gravitational collapse where, put in current terminology, they glued together patches of two exact solutions of Einstein's equations: the Schwarzschild metric and the spatially closed Friedmann–Lemaître–Robertson–Walker metric (for a modern exposition, see [Poi09], Sec. 3.8). The latter describes a fluid collapsing to a singular point and forms the interior patch of the Oppenheimer–Snyder model; the exterior patch is given by the Schwarzschild metric. An outside observer sees the fluid contracting asymptotically towards the Schwarzschild ra-

dius given by the total mass of the fluid, with the process never actually complete from the observer's point of view. An observer comoving with the fluid, on the other hand, sees the fluid continually compressing until a point of infinite density is reached within a finite proper time experienced by the observer. The observer worldline ends abruptly, as it meets a curvature singularity inside the newly formed black hole.

The Oppenheimer–Snyder collapse model was far from establishing a convincing argument for singularities as a generic outcome of physical processes in general relativity. Both solutions employed in the model are highly symmetric and hence do not exactly describe actual physical configurations where perturbations, at least small, are always present. For instance, unlike in the model, the collapse of a dying star is never perfectly spherically symmetric. This left open the possibility that in generic situations lacking symmetries singularities could be avoided – along with the breakdown of laws of physics that would ensue. This was the expectation of many physicists, even though there was no proof along these lines either [SG15].

The question of singularities in general relativity remained unresolved for half a century until in 1965 Roger Penrose published the first modern singularity theorem [Pen65]. Penrose proved, in a mathematically rigorous manner, that given a certain physically plausible energy condition, reasonable assumptions on the causal structure and topology of spacetime, and physically realistic geometric substructures of spacetime, called trapped surfaces, singularities would be unavoidable. The concept of a trapped surface, a compact two-dimensional submanifold with all light rays initially perpendicular to it having the property of converging in the future direction, played a quintessential role in generalizing the earlier arguments for the singular end of a gravitational collapse. The concept is an abstraction of sufficient generality that removes the typical symmetry assumptions needed in more rudimentary singularity theorems but maintains the properties of compactness and focusing of light rays that prove to be pivotal in deduction of singularities.

In addition to the notion of a trapped surface, key to Penrose's theorem was introduction of novel mathematical techniques of differential topology, unforeseen in physics at the time. By then, most of the research in general relativity had consisted of the study of exact solutions of Einstein's equations and perturbation theory in the weak field limit, with a predominantly local point of view on the spacetime geometry. Although this was and still is an adequate approach to many problems, it typically falls short when considering highly complex and nonlinear situations like a gravitational collapse without any symmetry assumptions. Instead of trying to solve Einstein's equation exactly or in some approximation, one can instead consider curvature tensor inequalities that capture the physically sensible statement that (for ordinary matter) gravity is always attractive. A central insight in Penrose's proof was that by applying these curvature inequalities to light rays emanating from a trapped surface, one is led to consider the topology of the

surface generated by following the light rays into the future. In particular, one finds that under the assumption of geodesic completeness, such a surface has to be compact. After a series of topological arguments this in turn leads to a contradiction with assumptions on the causal structure and topology of spacetime.

In physics, Penrose's theorem was groundbreaking for the reasons already explained above. Furthermore, it paved the way to the development of Hawking's cosmological singularity theorems, which state the existence of a past singularity in expanding universe models and thereby advance our understanding of the implications of general relativity to cosmology. Together, the Hawking–Penrose singularity theorems indicate that evolution of spacetime geometry in general relativity typically converges to a state where the theory no longer is an adequate description of what happens, physically.

Mathematically, the significance of Penrose's theorem lies in the fact that it elucidates the global properties of Lorentzian manifolds, which are the basic mathematical structure underlying general relativity. Unlike for Riemannian manifolds, incompleteness is a generic feature of Lorentzian manifolds. Thus, completeness is not a safe assumption on a Lorentzian manifold, which is one remarkable difference between Riemannian and Lorentzian geometry.

This thesis is structured as follows. In Chapter 2, we lay out some basic definitions and results in semi-Riemannian geometry. We assume that the reader is already familiar with elementary differential geometry and Riemannian geometry. Therefore, our review of semi-Riemannian geometry, where much of the results are directly transferred from Riemannian geometry, will be rather quick. For a more comprehensive treatment of differential and Riemannian geometry, see *e.g.* [Lee97,Lee03]. In semi-Riemannian geometry and its application in Penrose's theorem, our basic reference will be [O'N83].

In Chapter 3, we introduce calculus of variations on semi-Riemannian manifolds and review some results therein that will be necessary in the proof of Penrose's theorem. Here, again, much is already familiar from the Riemannian context. We will concentrate on results that are not typically encountered on a basic Riemannian geometry course. In particular, we will review geometric conditions relevant for light ray focusing due to curvature, and how light ray focusing is related to causality on a Lorentzian manifold.

In Chapter 4, we present a set of definitions that can be used to analyze causal structure of Lorentzian manifolds. We will state results that establish a connection between causality and topology of Lorentzian manifolds. A precise definition of a trapped surface will be given and its causal–topological properties will be studied. Finally, we will state and prove Penrose's theorem. We conclude with a few remarks on the theorem.

Chapter 2

Semi-Riemannian manifolds

2.1 Smooth manifolds

We start by summarizing some basic facts on smooth manifolds. First, recall that an n-dimensional topological manifold is a second countable Haussdorff space in which each point has a neighborhood homeomorphic to an open subset of $\mathbb{R}^n$. A subset S of a topological n-manifold is a topological hypersurface if for every $p \in S$ there exists a neighborhood $U \subset M$ of p and a homeomorphism $\phi : U \to V \subset \mathbb{R}^n$ s.t. $\phi(U \cap S) = \phi(U) \cap \Pi$, where V is open and Π a hyperplane in $\mathbb{R}^n$.

A smooth manifold M is a topological manifold equipped with a complete atlas (a collection of smoothly compatible coordinate charts that covers all of M and includes any coordinate chart smoothly compatible with every coordinate chart in the collection). A submanifold of a smooth manifold M is a subset $S \subset M$ s.t.

1. S has the induced topology, and
2. The inclusion $j : S \to M$ is an immersion (*i.e.*, j is smooth and at each $p \in S$ the pushforward j_{*p} is injective).

We use the standard notation $C^\infty(M)$ for the set of smooth functions on M, $\mathcal{T}_k(M)$ for the set of smooth $(0,k)$ tensor fields on M, $\mathcal{T}^l(M)$ for the set of smooth $(l,0)$ tensor fields on M, and $\mathcal{T}^l_k(M)$ for the set of smooth (l,k) tensor fields on M. As usual in differential geometry, we employ the Einstein summation convention where a sum over a repeated index is implied.

2.2 Semi-Riemannian metric

Let V be a real vector space. In our context, it suffices to consider a finite-dimensional V. A symmetric bilinear form on V is an $\mathbb{R}$-bilinear function $f : V \times V \to \mathbb{R}$ s.t. $f(v,w) = f(w,v)\ \forall v, w \in V$. We say that f is *positive (negative) definite* if $v \neq 0$ entails $f(v,v) > 0$ ($f(v,v) < 0$). If for all $v \in V$ $f(v,v) \geq 0$ ($f(v,v) \leq 0$, f is *positive (negative)*

semidefinite. If $f(v,w) = 0\ \forall w \in V$ implies $v = 0$, we say that f is *nondegenerate*.

Definition 2.2.1. The index ν of a symmetric bilinear form f on V is $\max\{\dim W : W \subset V, f|_{W\times W}$ is negative definite$\}$.

Definition 2.2.2. A scalar product g on V is a nondegenerate symmetric bilinear form on V.

From now on we suppose that the vector space V is furnished with a scalar product g, which makes it a *scalar product space*. We say that vectors $v, w \in V$ are *orthogonal* if $g(v,w) = 0$, and denote this by $v \perp w$. Given a subspace $W \subset V$, we denote

$$W^{\perp} = \{v \in V : v \perp w\ \ \forall w \in W\}. \tag{2.2.1}$$

Furthermore, we say that a subspace $W \subset V$ is nondegenerate provided $g|_{W\times W}$ is nondegenerate.

The metric tensor is defined almost as in Riemannian geometry but positive definiteness is now replaced by nondegeneracy:

Definition 2.2.3. Let M be a smooth manifold. A metric tensor $g \in \mathcal{T}_2(M)$ is a symmetric nondegenerate tensor field of constant index.

Definition 2.2.4. A semi-Riemannian manifold (M, g) is a smooth manifold M equipped with a metric tensor g.

As usual, we simplify our notation by calling M a semi-Riemannian manifold and write (M, g) only when it is necessary to distinguish this from some other semi-Riemannian structure defined on the same smooth manifold. Also, we use $\langle\cdot,\cdot\rangle$ as another notation for the metric.

Semi-Riemannian manifolds pertinent to relativity are those whose metric tensors have an index $\nu = 1$. With $n = \dim M \geq 2$ and $\nu = 1$, we call M a *Lorentzian manifold*. The simplest Lorentzian manifold is the *Minkowski n-space* $\mathbb{R}^n_1 = (\mathbb{R}^n, g)$ where the metric is given in the canonical coordinates $(x^0, ..., x^{n-1})$ of $\mathbb{R}^n$ as follows: for $v_p = v^i\partial_i, w_p = w^i\partial_i \in T_p(\mathbb{R}^n)$, we define

$$\langle v_p, w_p\rangle = -v^0w^0 + \sum_{i=1}^{n-1} v^iw^i. \tag{2.2.2}$$

The components of the Minkowski metric are denoted by η_{ij} and in the above coordinate frame they form the matrix $[\eta_{ij}] = \mathrm{diag}(-1, 1, ..., 1)$ with the number of 1's on the diagonal equal to $n-1$. The inverse matrix of $[\eta_{ij}]$ we write as $[\eta^{ij}]$. Observe that above we used the relativistic indexing of coordinates starting with x^0. We shall use this convention henceforth.

Let $p \in M$. A tangent vector $v \in T_p(M)$ can be classified into three different types that specify the *causal character* of a vector: If $\langle v, v\rangle > 0$ or $v = 0$, we say that v is a *spacelike vector*. If $\langle v, v\rangle = 0$ for $v \neq 0$, we say that v is a *null vector*. If $\langle v, v\rangle < 0$, then v is called a *timelike vector*. Finally, if a vector is nonspacelike, we say it is a *causal vector*. If each tangent vector in a subset $W \subset T_p(M)$ has the same causal character, we say that W itself has that character. In particular, the null subset of $T_p(M)$ consisting of all the null vectors in $T_p(M)$ is called the *nullcone* at $p \in M$.

Let then M be a Lorentzian manifold and let Υ_p be the set of all timelike vectors in $T_p(M)$. We define, for $u \in \Upsilon_p$, the *timecone* of $T_p(M)$ containing u:

$$\tau_p(u) = \{v \in \Upsilon_p : \langle u, v\rangle < 0\}, \tag{2.2.3}$$

and the *opposite* timecone:

$$\tau_p(-u) = \{v \in \Upsilon_p : \langle u, v\rangle > 0\}. \tag{2.2.4}$$

It can be shown (see [O'N83], Lemma 5.29) that any two timelike vectors $v, w \in T_p(M)$ belong to the same timecone if and only if $\langle v, w\rangle < 0$. It then follows that in each tangent space of the manifold, there are exactly two timecones. Smoothly assigning to each point p a timecone τ_p defines a *time-orientation* of M. If there exists a smooth function that defines a time-orientation for M, we say that M is *time-orientable*. A chosen time-orientation of M we call the *future* and the corresponding opposite orientation the *past*. A tangent vector $v \in T_p(M)$ that either lies in τ_p or is null and satisfies $\langle v, u\rangle < 0$ for some (and, hence, for all) $u \in \tau_p$ is said to be *future-pointing*. A *past-pointing* tangent vector is defined analogously.

2.3 Connection and curvature

On a semi-Riemannian manifold M, besides the metric we also want to introduce a further structure that generalizes the notion of a directional rate of change of a vector field, a *connection*:

Definition 2.3.1. A connection ∇ on M is a map $\nabla : \mathcal{T}(M) \times \mathcal{T}(M) \to \mathcal{T}(M)$ that satisfies the following conditions:

1. $\nabla_X Y$ is linear over $C^\infty(M)$ in X,
2. $\nabla_X Y$ is linear over $\mathbb{R}$ in Y,
3. $\nabla_X(fY) = (Xf)Y + f\nabla_X Y$ for all $f \in C^\infty(M)$,

for all $X, Y \in \mathcal{T}(M)$.

We say that $\nabla_X Y$ is the *covariant derivative* of Y in the direction of X. This is a

generalization of the directional derivative of a vector field with respect to some vector field.

A well-known fundamental theorem states that there is a unique, natural connection on a semi-Riemannian manifold:

Theorem 2.1. Let M be a semi-Riemannian manifold. Then there exists a unique connection ∇, called the *Levi-Civita connection* of M, satisfying

1. $[X, Y] = \nabla_X Y - \nabla_Y X$,
2. $X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Y\rangle$

for all $X, Y, Z \in \mathcal{T}(M)$.

Remark. The first condition is often stated in terms of the vanishing of the *torsion tensor* $T : \mathcal{T}(M) \times \mathcal{T}(M) \to \mathcal{T}(M)$ defined by

$$T(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y]. \tag{2.3.1}$$

The second property is called *metric compatibility*.

The concept of curvature remains essentially the same as well when we move from the Riemannian context to semi-Riemannian geometry. Curvature measures the extent to which covariant derivatives in different directions fail to commute:

Definition 2.3.2. Let M be a semi-Riemannian manifold with the Levi-Civita connection ∇. The *Riemann curvature tensor* of ∇ is the map $R : \mathcal{T}(M) \times \mathcal{T}(M) \times \mathcal{T}(M) \to \mathcal{T}(M)$ given by

$$R(X, Y)Z = \nabla_{[X,Y]} Z - \nabla_X \nabla_Y Z + \nabla_Y \nabla_X Z \tag{2.3.2}$$

The Riemann tensor possesses the usual algebraic symmetries and satisfies the Bianchi identities. These properties are discussed at length in, for instance, [Lee97]; we shall not write them down in detail here.

With a metric contraction of the Riemann tensor, we can define the *Ricci curvature tensor* and the *Ricci scalar curvature* of ∇:

Definition 2.3.3. The Ricci curvature tensor of M is the tensor field $\mathrm{Ric} \in \mathcal{T}_2(M)$ defined pointwise as $\mathrm{Ric}(v, w) = \mathrm{Tr}(z \mapsto R(v, z)w)$ for all $v, w \in T_p(M)$ for $p \in M$.

Definition 2.3.4. The Ricci scalar curvature $S \in C^\infty(M)$ of M is defined locally as follows: let (U, x) be a chart on M, let $[g^{ij}]$ be the components of the inverse metric and $[R_{ij}]$ the components of the Ricci tensor in (U, x). Then S is given by

$$S = g^{ij} R_{ij}. \tag{2.3.3}$$

Finally, in view of general relativity we define the *Einstein tensor*:

Definition 2.3.5. The Einstein tensor $G \in \mathcal{T}_2(M)$ is defined as $G = \mathrm{Ric} - \frac{1}{2}Sg$.

In general relativity, M is a connected and time-oriented Lorentzian manifold with $n = 4$, called a *spacetime* [HE23]. Matter and energy content of a spacetime is described by a symmetric tensor $T \in \mathcal{T}_2(M)$ called the *stress-energy tensor* of M. The equation of motion of general relativity is *Einstein's equation*

$$G = 8\pi T. \tag{2.3.4}$$

Here we expressed Einstein's equation in the geometric units where the speed of light $c = 1$ and Newton's gravitational constant $G_{\mathrm{N}} = 1$. In case $T = 0$, *i.e.*, if the spacetime is devoid of matter content, the spacetime geometry satisfies the *vacuum Einstein equation*:

$$\mathrm{Ric} = 0, \tag{2.3.5}$$

that is, M is *Ricci-flat*.

2.4 Geodesics and normal neighborhoods

On a smooth manifold M, a *curve* is a smooth mapping $\alpha : I \to M$ with $I \subset \mathbb{R}$ an open subset of real numbers. Given a closed interval $[a, b] \subset \mathbb{R}$, we say that a mapping $\alpha : [a, b] \to M$ is a *curve segment* if there is a smooth mapping $\tilde{\alpha} : (a', b') \to M$ s.t. $[a, b] \subset (a', b')$ and $\tilde{\alpha}|_{[a,b]} = \alpha$. We say that a map $\alpha : [a, b] \to M$ is a *piecewise smooth curve segment* if there exists a partition $a < t_1 < ... < t_k < b$ of $[a, b]$ s.t. the restriction $\alpha|_{[t_i,t_{i+1}]}$ is a curve segment for all i. Finally, we consider a half-open interval $[0, b)$ and define that a piecewise smooth curve $\alpha : [0, b) \to M$ is *extendible* if it has a continuous (not necessarily piecewise smooth) extension $\tilde{\alpha} : [0, b] \to M$. We say that the point $\tilde{\alpha}(b)$ is an *endpoint* of α.

Let now M be a semi-Riemannian manifold with the Levi-Civita connection ∇ and let $\alpha : I \to M$ be a smooth curve. A smooth map $X : I \to TM$ is a *smooth vector field along* α if $X_t = X_{\alpha(t)} \in T_{\alpha(t)}(M)$ $\forall t \in I$, *i.e.* X smoothly assigns to every $t \in I$ a tangent vector in the tangent space at the point $\alpha(t)$. We denote the set of all such vector fields by $\mathcal{T}(\alpha)$. The connection ∇ induces, in a natural way, a unique map $D_t : \mathcal{T}(\alpha) \to \mathcal{T}(\alpha)$, which we call the *covariant derivative along* α. Its basic properties are encapsulated in the following proposition whose proof can be found, for instance, in [Lee97]:

Proposition 2.1. Let M and α be as above. Then there is a unique map $D_t : \mathcal{T}(\alpha) \to \mathcal{T}(\alpha)$ that satisfies, for all $V, W \in \mathcal{T}(\alpha)$:

1. $D_t(aV + bW) = aD_tV + bD_tW \quad \forall\, a, b \in \mathbb{R}$,

2. $D_t(fV) = \dot{f}V + fD_tV \quad \forall\, f \in C^\infty(I)$,
3. If $Y \in \mathcal{T}(M)$ is a vector field s.t. $V_t = Y_{\alpha(t)}$, then $D_tV = \nabla_{\dot{\alpha}}Y$,
4. D_t inherits metric compatibility from ∇:

$$\frac{d}{dt}\langle V, W\rangle = \langle D_tV, W\rangle + \langle V, D_tW\rangle. \tag{2.4.1}$$

Given a $V \in \mathcal{T}(\alpha)$, we say that V is *parallel* if $D_tV = 0$. Expressed in local coordinates, $D_tV = 0$ reduces to a system of linear ordinary differential equations, which has a unique solution given some initial conditions. Hence, given a tangent vector $v \in T_{\alpha(a)}(M)$ with $a \in I$, a parallel vector field $V \in \mathcal{T}(\alpha)$ satisfying $v = V(a)$ is uniquely determined. Therefore, for any $p, q \in \alpha(I)$ we may define a map $T_p(M) \to T_q(M)$ called *parallel translation along* α. The unique parallel vector field V satisfying $v = V(a)$ we call the *parallel translate of* v *along* α.

In semi-Riemannian geometry, the Euclidean straight lines are generalized by the concept of a *geodesic*:

Definition 2.4.1. Let $\gamma : I \to M$ be a smooth curve on a semi-Riemannian manifold with a connection ∇. If $D_t\dot{\gamma} = 0$, we say that γ is a geodesic.

In case γ is a piecewise smooth curve segment s.t. each of its smooth subsegments are geodesics, γ is called a *broken geodesic*. If a curve $\alpha : I \to M$ has a *reparametrization* $\gamma = \alpha \circ h : J \to M$, with $h : J \to I$ a smooth function on an interval J, s.t. γ is a geodesic, then we call α a *pregeodesic*. We say that a curve α in M is *spacelike, timelike,* or *null* if for all t its tangent vector $\dot{\alpha}(t)$ is spacelike, timelike, or null, respectively. Given a geodesic γ, its causal character is always fixed to one of these since $(d/dt)\langle\dot{\gamma}, \dot{\gamma}\rangle = 2\langle\dot{\gamma}, D_t\dot{\gamma}\rangle = 0$. An arbitrary curve does not have to fall into any of these three classes. A curve whose tangent vector is never spacelike we call a *causal curve*. A causal curve whose tangent vector is always future-pointing is said to be *future-pointing*. A *past-pointing* causal curve is defined analogously.

The standard existence and uniqueness result (see *e.g.* [Lee97], Theorem 4.10) states that given a tangent vector $v \in T_p(M)$, there exists a unique geodesic $\gamma : I \to M$ with $0 \in I, \gamma(0) = p$, and $\dot{\gamma}(0) = v$. Such a geodesic is called a *geodesic starting at* p *with initial velocity* v. In fact, a stronger statement holds:

Proposition 2.2. Let $v \in T_p(M)$. Then there exists a unique geodesic $\gamma_v : I_v \to M$ satisfying:

1. v is the initial velocity of γ_v,
2. The domain of definition I_v is as large as possible: if $\tilde{\gamma} : I \to M$ is another geodesic with initial velocity v, then $I \subset I_v$ and $\tilde{\gamma} = \gamma|_I$.

Proof. See [O'N83], Proposition 3.24. □

The geodesic that satisfies the 2nd condition above is called *maximal* or *geodesically inextendible*. The following related notion is central to Penrose's theorem:

Definition 2.4.2. We say that a semi-Riemannian manifold M is *geodesically compelete* (or, simply, *complete*) if each maximal geodesic in M is defined on all of $\mathbb{R}$.

Restricting the notion to timelike, null, or spacelike geodesics only, we also say that M is *timelike, null*, or *spacelike geodesically compelete*, respectively.

For a point $p \in M$, consider all the tangent vectors to M at p and all the geodesics emanating from p. For each such geodesic there is a unique $v \in T_p(M)$ that is the tangent vector to the geodesic at p. Under certain conditions, we can thus map a tangent vector v at p to a point $q \in M$ reached by following the geodesic γ_v to some predefined affine parameter value. This map is called the *exponential map*, formally defined as follows:

Definition 2.4.3. Let $p \in M$ and let $D \subset T_p(M)$ be the set of tangent vectors v s.t. the inextendible geodesic γ_v is defined at least on $[0, 1]$. The exponential map at p is then the map $\exp_p : D \to M$ s.t. $\exp_p(v) = \gamma_v(1)\ \forall v \in D$.

As is known from Riemannian geometry, the exponential map becomes a diffeomorphism when restricted to small enough subsets of the tangent space:

Proposition 2.3. Let $p \in M$. Then there is a neighborhood U of $0 \in T_p(M)$ s.t. $\exp_p|_U : U \to V$ is a diffeomorphism onto some neighborhood V of p.

Let U and V be as above. If in addition U is starshaped about 0 (*i.e.*, if $v \in T_p(M)$ implies $tv \in T_p(M)\ \forall t \in [0, 1]$), then V is a *normal neighborhood* of p.

Definition 2.4.4. Let $C \subset M$ be open. If C is a normal neighborhood of each $p \in C$, we say that C is convex.

An important property of convex open sets is that given any $p, q \in C$ there exists a unique geodesic segment $\gamma : [0, 1] \to M$ with $\gamma(0) = p, \gamma(1) = q$, and $\gamma([0, 1]) \subset C$.

By a *convex covering* $\mathcal{C}$ of M we mean an open covering of M s.t. for any $U, V \in \mathcal{C}$, $U \cap V \neq \varnothing$ implies that $U \cap V$ is convex. The following lemma establishes the existence of convex coverings subordinate to arbitrary open coverings of a semi-Riemannian manifold:

Lemma 2.1. Let $\mathcal{D}$ be an open covering of M. Then there is a convex covering $\mathcal{C}$ s.t. each element of $\mathcal{C}$ is contained in some $D \in \mathcal{D}$.

Proof. See [O'N83], Lemma 5.10. □

2.5 Semi-Riemannian submanifolds

Definition 2.5.1. Let $S \subset M$ be a submanifold of a semi-Riemannian manifold M. S is a semi-Riemannian submanifold of M provided the pullback by inclusion $j^*(g)$ is a metric tensor on S.

Let then $S \subset M$ be a submanifold of a Lorentzian manifold M. If the vector subspace $T_p(S)$ has a fixed causal character in $T_p(M)$ for all $p \in S$, we say that the submanifold possesses that character, *e.g.* we say that a submanifold is spacelike.

Remark. Given a point $p \in M$, the singleton $\{p\}$ has a trivial tangent space $T_p(\{p\}) = \{0\}$ and 0 is a spacelike vector by definition. Hence, $\{p\}$ is a spacelike submanifold of M.

For a smooth submanifold $M \subset \widetilde{M}$, let $\widetilde{\mathcal{T}}(M)$ be the set of smooth $\widetilde{M}$ vector fields on M.

Let now $M \subset \widetilde{M}$ be a semi-Riemannian submanifold of a semi-Riemannian manifold $\widetilde{M}$ with dimensions $\dim M = n$ and $\dim \widetilde{M} = n + k$. At each $p \in M$ the tangent space $T_p(\widetilde{M})$ can be decomposed as

$$T_p(\widetilde{M}) = T_p(M) + T_p(M)^{\perp}. \tag{2.5.1}$$

Given that a connection $\widetilde{\nabla}$ is defined on $\widetilde{M}$, the induced connection $\widetilde{\nabla} : \mathcal{T}(M) \times \widetilde{\mathcal{T}}(M) \to \widetilde{\mathcal{T}}(M)$ can be defined as follows. Given any smooth vector fields $X, Y \in \mathcal{T}(M)$, a point $p \in M$, and a coordinate neighborhood $U \subset \widetilde{M}$ of p, let $\widetilde{X}, \widetilde{Y}$ be any smooth extensions of X, Y over U. The induced connection is then $\widetilde{\nabla}_X Y = \widetilde{\nabla}_{\widetilde{X}} \widetilde{Y}|_{U \cap M}$.

Remark. The induced connection is independent of the choice of smooth extensions of the vector fields because $(\widetilde{\nabla}_{\widetilde{X}} \widetilde{Y})_p$ only depends on the connection $\widetilde{\nabla}$, the tangent vector X_p, and the directional derivative $X_p b^j$ where $b^j \partial_j = Y$ is the local representation of Y in (U, x).

The normal projection of a covariant derivative of a tangent vector field on M in the direction of another tangent vector field on M turns out to be useful in characterizing the extrinsic geometry of M. The following lemma, the proof of which can be found, *e.g.*, in [O'N83], states a couple of fundamental properties of the normal projection:

Lemma 2.2 (Second fundamental form)**.** The map $II : \mathcal{T}(M) \times \mathcal{T}(M) \to \mathcal{T}(M)^{\perp}$ defined by

$$II(X, Y) = (\widetilde{\nabla}_X Y)^{\perp} \tag{2.5.2}$$

is bilinear and symmetric, and it is called the *second fundamental form* of $M \subset \widetilde{M}$.

Metric contraction of the second fundamental form allows us to introduce an averaged

version of II: the *mean curvature vector field* H of M is defined, pointwise at any $p \in M$, as

$$H_p = \frac{1}{n}\eta^{ij} II(e_i, e_j), \tag{2.5.3}$$

where $n = \dim M$ and $e_1, ..., e_n$ is any orthonormal basis in $T_p(M)$.

For $X \in \mathcal{T}(M), Y \in \mathcal{T}(M)^\perp$ we denote $\widetilde{II}(X,Y) = (\widetilde{\nabla}_X Y)^\top$. This provides another object that characterizes the extrinsic geometry of the submanifold:

Definition 2.5.2 (Shape operator)**.** Let $Y \in \mathcal{T}(M)^\perp$. The shape operator of M with respect to Y is the linear map $S_Y : \mathcal{T}(M) \to \mathcal{T}(M)$ defined by

$$S_Y(X) = -\widetilde{II}(X,Y). \tag{2.5.4}$$

As for the tangent vectors to the manifold we can define the tangent bundle $T\widetilde{M}$, for a semi-Riemannian submanifold M we can also define the *normal bundle* (NM, π) where $NM = \bigsqcup_{p\in M}(T_p(M))^\perp$ and $\pi : NM \to M$ is the natural projection. Vector fields in $\mathcal{T}(M)^\perp$ are smooth sections of NM.

Definition 2.5.3. Given a semi-Riemannian submanifold M of $\widetilde{M}$ with mean curvature vector H, the *convergence* of M is the function $\mathbf{k} : NM \to \mathbb{R}$ defined by

$$\mathbf{k}(n) = \langle n, H_p\rangle = \frac{1}{\dim M}\operatorname{Tr} S_n \quad \text{for } n \in T_p(M)^\perp.$$

The convergence of a submanifold will be used in the definition of a trapped surface in Chapter 4.

Chapter 3

Calculus of variations

The purpose of this chapter is to introduce results on the effect of curvature on geodesics, in particular focusing of null geodesics due to curvature, that tell us when it is possible to connect a point in a Lorentzian manifold to a spacelike submanifold thereof by a timelike path. In the next chapter, these results can be used to relate such causal notions to the topology of a Lorentzian manifold and certain causally defined subsets thereof.

3.1 Jacobi fields

The concepts of variation of a curve, variation vector fields, and Jacobi fields are defined as in Riemannian geometry.

Definition 3.1.1. A variation of a curve segment $\alpha : [a, b] \to M$ is a mapping $\Gamma : [a, b] \times (-\epsilon, \epsilon) \to M$ s.t. $\alpha(t) = \Gamma(t, 0)\ \forall\, t \in [a, b]$.

The curves $\Gamma_s : [a, b] \to M, \Gamma_s(t) = \Gamma(t, s)$ are called *longitudinal curves* or *main curves*, and the curves $\Gamma^t : (-\epsilon, \epsilon) \to M, \Gamma^t(s) = \Gamma(t, s)$ are called *transverse curves*. We also denote

$$\partial_t\Gamma(t, s) := \frac{d}{dt}\Gamma_s(t), \quad \partial_s\Gamma(t, s) := \frac{d}{ds}\Gamma^t(s) \quad \forall\,(t, s) \in [a, b] \times (-\epsilon, \epsilon).$$

The vector field V along α defined by $V(t) = \partial_s\Gamma(t, 0)$ is the *variation vector field* of Γ. The *transverse acceleration vector* A is defined by $A(t) = D_s\partial_s\Gamma(t, 0)$. Γ is a *fixed endpoint variation* provided $\Gamma_s(a) = \alpha(a), \Gamma_s(b) = \alpha(b)$ for all s. If each main curve is a geodesic, Γ is a *geodesic variation*.

Definition 3.1.2. Let γ be a geodesic. If a vector field J on γ satisfies the Jacobi equation

$$D_t^2 J + R(J, \dot\gamma)\dot\gamma = 0, \tag{3.1.1}$$

we call it a *Jacobi field*.

Geodesic variations and Jacobi fields are closely connected, as shown by this lemma:

Lemma 3.1. Let Γ be a geodesic variation of a geodesic γ. Then its variation vector field is a Jacobi field.

Proof. See [O'N83], Lemma 8.3. □

Definition 3.1.3. Let γ be a geodesic. Points $\gamma(a)$ and $\gamma(b)$, with $a \neq b$, are *conjugate along* γ if there is a nonzero Jacobi field J on γ s.t. $J(a) = J(b) = 0$.

3.2 Focal points and curvature

A basic problem in Riemannian geometry is to study the arc length of curves with fixed endpoints p and q. The first variation of the arc length functional

$$L(\alpha) = \int_a^b \sqrt{|\langle \dot{\alpha}(t), \dot{\alpha}(t) \rangle|} dt \tag{3.2.1}$$

yields the well-known result that the length-minimizing curves are exactly those that satisfy the geodesic equation. The second variation of arc length establishes a relation between the critical points of the length functional and curvature of the manifold. This in turn can be related to global properties of a Riemannian manifold by analyzing the *index form*, which can be used to study conjugacy of points along a non-null geodesic and singular points of the exponential map.

In semi-Riemannian geometry, and in Lorentzian geometry in particular, we also need to include null geodesics in our analysis. Here we take a slightly more general point of view than in a typical Riemannian discussion by letting one of the points be instead a submanifold, which we call the *endmanifold*. Letting $q \in M$ be a point and $S \subset M$ be a semi-Riemannian submanifold of M, and $[0, b]$ an interval, we denote by $\Omega(S, q)$ the set of all piecewise smooth curves $\alpha : [0, b] \to M$ with $\alpha(0) \in S$ and $\alpha(b) = q$. Then we define $T_\gamma(\Omega)$ as the set of all piecewise smooth vector fields V on γ s.t $V(0) \in T_{\gamma(0)}(S)$ and $V(b) = 0$.

Definition 3.2.1. Let γ be a geodesic normal to S, meaning that $\gamma(0) \in S$ and $\gamma'(0) \perp v$ for all $v \in T_{\gamma(0)}(S)$, and let V be a Jacobi field on γ. We say that V is an S-Jacobi field on γ if $V(0) \in T_{\gamma(0)}(S)$ and $(D_t V(0))^\top = \widetilde{II}(V(0), \dot{\gamma}(0))$.

We get an equivalent characterization of an S-Jacobi field in terms of geodesic variations:

Proposition 3.1. A Jacobi field V on a geodesic γ normal to S is an S-Jacobi field if and only if V is the variation vector field of a variation Γ of γ through geodesics normal to S.

Proof. See [O'N83], Proposition 10.28. □

In consequence, we get a corollary for S-normal null geodesics in the Lorentzian case, which we are interested in, in particular:

Corollary 3.1. Let $S \subset M$ be a semi-Riemannian submanifold of a Lorentzian manifold M. Given a null geodesic γ normal to S, an S-Jacobi field on γ is the variation vector field of a variation of γ through null geodesics if and only if $V \perp \gamma$.

Proof. See [O'N83], Corollary 10.40. □

The concept of conjugate point along a geodesic is generalized by defining a *focal point* of S along a geodesic:

Definition 3.2.2. Let $S \subset M$ and let γ be a geodesic normal to S. The point $\gamma(s)$ with $s \neq 0$ is a focal point of S along γ if there exists a nonzero S-Jacobi field V on γ with $V(s) = 0$.

When S consists of a single point, this naturally reduces to the conventional notion of conjugacy along a geodesic.

Let $\alpha : [0, b] \to M$ be a curve segment in a semi-Riemannian manifold M and let $\Gamma : [0, b] \times (-\epsilon, \epsilon) \to M$ be a piecewise smooth variation of α. Instead of the arc length, we study the energy functional

$$E(\alpha) = \frac{1}{2} \int_0^b \langle \dot{\alpha}, \dot{\alpha} \rangle dt. \tag{3.2.2}$$

The function $E_\Gamma : (-\epsilon, \epsilon) \to \mathbb{R}$ defined by

$$E_\Gamma(s) = \frac{1}{2} \int_0^b \langle \partial_t \Gamma(t, s), \partial_t \Gamma(t, s) \rangle dt \tag{3.2.3}$$

is smooth irrespective of the choice of variation Γ, in contrast to the arc length whose integrand is a composition of smooth and non-smooth functions. This allows for study of variations of null curves using E_Γ. With the arc length, such analysis is constrained, for instance, by the fact that variations whose main curves have different causal character for $s \in (-\epsilon, 0)$ and $s \in (0, \epsilon)$ do not define, in general, a smooth length functional.

The first and second variation formulas for E_Γ are given in the following proposition:

Proposition 3.2. Let $\alpha : [0, b] \to M$ be a piecewise smooth curve with breaks $t_1 < ... < t_k$. Given a piecewise smooth variation $\Gamma : [0, b] \times (-\epsilon, \epsilon) \to M$ of α with the variation and

transverse acceleration vector fields V and A, the first variation of E gives

$$E'_\Gamma(0) = \int_0^b \langle D_t V, \dot{\alpha}\rangle dt = \langle V, \dot{\alpha}\rangle\Big|_0^b - \sum_{i=1}^{k} \langle V(t_i), \Delta\dot{\alpha}(t_i)\rangle - \int_0^b \langle V, \ddot{\alpha}\rangle dt. \tag{3.2.4}$$

Moreover, in case α is a geodesic, the second variation yields

$$E''_\Gamma(0) = \langle A, \dot{\alpha}\rangle\Big|_0^b + \int_0^b \left(\langle D_t V, D_t V\rangle - \langle R(V, \dot{\alpha})V, \dot{\alpha}\rangle\right) dt. \tag{3.2.5}$$

Proof. See [O'N83], Proposition 10.39. □

Letting now $q \in M, S \subset M$ be as above, we may consider E as a function $\Omega(S, q) \to \mathbb{R}$. Completely analogously to the variation of the arc length, the first variation formula above implies that the critical points of E are geodesics from S to q that are normal to S. Continuing with the analogy, given a geodesic γ and any variation $\Gamma(t, s)$ of γ with $V = \partial_s\Gamma$ and each main curve $\Gamma_s \in \Omega(S, q)$, we define a bilinear form $\mathcal{H}_\gamma : T_\gamma(\Omega) \times T_\gamma(\Omega) \to \mathbb{R}$ by setting

$$\mathcal{H}_\gamma(V, V) = E''_\Gamma(0). \tag{3.2.6}$$

The form $\mathcal{H}_\gamma$, which is analogous to the index form of the arc length functional, we call the *Hessian* of E. It is the unique bilinear form satisfying the definition and does not depend on the full variation Γ but only its variation vector field V; any variation of α having the same V gives the same Hessian for E. As a corollary to Proposition 3.2, we get an explicit formula for the Hessian:

Corollary 3.2. Let $\gamma \in \Omega(S, q)$ be a normal geodesic. Then

$$\mathcal{H}_\gamma(V, W) = \int_0^b \left(\langle D_t V, D_t W\rangle - \langle R(V, \dot{\gamma})W, \dot{\gamma}\rangle\right) dt - \langle \dot{\gamma}(0), II(V(0), W(0))\rangle, \tag{3.2.7}$$

for all $V, W \in T_\gamma(\Omega)$.

Proof. The expression on the right hand side is clearly bilinear and because of symmetries of the Riemann tensor and the second fundamental form, it is also a symmetric form. Setting $V = W$, we compare the right hand side of the formula above to the second variation formula of Proposition 3.2. Observe that the transverse acceleration vanishes at $\gamma(b)$ and, at the other endpoint, $A(0) = D_s V(0) = (\nabla_{\widetilde{V}}\widetilde{V})_{\gamma(0)}$, where $\widetilde{V}$ is any smooth extension of V to a neighborhood of $\gamma(0)$. Thus, $\langle \dot{\gamma}(0), A(0)\rangle = \langle \dot{\gamma}(0), (\nabla_{\widetilde{V}}\widetilde{V})^\perp_{\gamma(0)}\rangle = \langle \dot{\gamma}(0), II(V(0), V(0))\rangle$ and we see that the definition $\mathcal{H}_\gamma(V, V) = E''_\Gamma(0)$ is satisfied. By straightforward linear algebra, it follows that any other symmetric bilinear form satisfying this has to be identical to $\mathcal{H}_\gamma$. □

The above formula for the Hessian is useful when studying the focal points of a null

geodesic normal to a submanifold, analogously to the way the index form can be used to investigate conjugacy of points along a non-null geodesic. In particular, the Hessian can be used to prove a proposition that acts as a crucial intermediate step towards Penrose's theorem:

Proposition 3.3. Let S be a spacelike codimension 2 submanifold of a Lorentzian manifold M and H be the mean normal curvature vector field and $\mathbf{k}$ the convergence of S. Given a null geodesic σ normal to S at $\sigma(0)$ s.t.

1. $\mathbf{k}(\dot{\sigma}(0)) = \langle \dot{\sigma}(0), H_{\sigma(0)} \rangle > 0$,
2. $\mathrm{Ric}(\dot{\sigma}, \dot{\sigma}) \geq 0$,

there exists a focal point $\sigma(s)$ of S along σ with $0 < s \leq 1/k$, where $k = \mathbf{k}(\dot{\sigma}(0))$, assuming σ is defined on this interval.

Proof. See [O'N83], Proposition 10.43. □

Proposition 3.3 tells us that given a surface that makes a normal null geodesic re-converge and Ricci curvature that bends null geodesics towards each other, there has to be a future point on the geodesic where nearby geodesics tend to focus. This by itself does not mean that a singularity has to occur in the future of the null geodesic. However, in the next chapter we will see that if the shape of the surface is such that this holds for all of its normal null geodesics and, moreover, the surface is compact as a topological space, we are not far from concluding that a singularity is inevitable. Before that, there are still a few necessary results that can be learned from the calculus of variations of curves in a Lorentzian manifold.

3.3 Causality from calculus of variations

Given a spacelike submanifold of a Lorentzian manifold, we would like to know when it is possible to connect a point to the submanifold by a causal curve. For the present discussion it is particularly important to figure out when this cannot be done by a timelike curve even though it can be done by some causal curve. Furthermore, one can ask when it is possible to do this by a null geodesic while it is also possible with a timelike curve.

We start by considering a special case where the endmanifold is consists of a single point p and ask when it is possible to deform a causal curve from p to q into a timelike curve by a fixed endpoint variation. The following proposition gives a sufficient condition for this:

Proposition 3.4. Let M be a Lorentzian manifold. Given a causal curve α from p to q that is not a null pregeodesic, there exists a timelike curve from p to q arbitrarily close to α.

Proof. See [O'N83], Proposition 10.46. □

Thus, any causal curve that cannot be turned into a null geodesic by a reparametrization can be deformed into a timelike curve. Then the question remains, when is this possible for a causal curve that is a null geodesic? With the aid of the previous proposition, one may in fact answer this question in a more general situation where the endmanifold is not necessarily a singleton but any spacelike submanifold:

Proposition 3.5. Let $S \subset M$ be a spacelike submanifold of a Lorentzian manifold M. Let $\gamma \in \Omega(S, q)$ be a normal null geodesic and suppose there is a focal point of S along γ before q. Then there exists a timelike curve from S to q arbitrarily close to γ.

Proof. See [O'N83], Proposition 10.48. □

Finally, previous propositions can be combined into a theorem stating sufficient and necessary conditions for there being a small deformation of a causal curve from S to q into a timelike curve:

Theorem 3.1. Let S be a spacelike submanifold of a Lorentzian manifold M. Given a causal curve $\alpha \in \Omega(S, q)$, there is a timelike curve in $\Omega(S, q)$ arbitrarily near α, except in case α is a null geodesic normal to S and there are no focal points of S along α before q.

We see that normal null geodesics define, up to their first focal point, a special subset of M that cannot be reached by following 'slower' curves starting from the submanifold. This is analogous to the Riemannian fact that a geodesic fails to be minimizing past its first conjugate point. To understand the causal structure of a Lorentzian manifold, it is thus equally important to know when a normal null geodesic forms its first focal point. This is where the focusing results of the previous section come into play. In the next chapter, this is studied in more detail.

Chapter 4

Penrose's theorem

4.1 Causal relations and causality conditions

We assume throughout this chapter that M is a time-oriented connected Lorentzian manifold. First, we define the following relations for $p, q \in M$:

1. $p \ll q$ if there exists a future-pointing timelike curve in M from p to q.
2. $p < q$ if there exists a future-pointing causal curve in M from p to q.

Also, we write $p \leq q$ if $p < q$ or $p = q$. Let $A \subset M$. The *chronological future* of A is defined as the set

$$I^+(A) = \{q \in M : \exists\, p \in A \text{ s.t. } p \ll q\} \ ,$$

and, correspondingly, the *chronological past* $I^-(A)$ is defined by turning the causal relation $p \ll q$ around. Furthermore, the *causal future* of A is defined as

$$J^+(A) = \{q \in M : \exists\, p \in A \text{ s.t. } p < q\} \ .$$

Again, the past version $J^-(A)$ is defined by requiring that there is a future-pointing causal curve from q to p. For a singleton p, we simply denote $I^\pm(p) = I^\pm(\{p\})$ and similarly for the causal past and future of the singleton. The above causal relations are transitive, which implies that $I^+(I^+(A)) = I^+(A)$ and $J^+(J^+(A)) = J^+(A)$, and similarly for the past versions.

Given an open $U \subset M$, U inherits the Lorentzian manifold structure from M in a natural way. If $A \subset U$, we write $I^+(A, U)$ for the chronological future of A in U (and similarly for other causal sets). It is clear that the causal relations holding in U also hold in M, which means that $I^+(A, U) \subset I^+(A) \cap U$. In general, the converse is not true, as is the case, *e.g.*, if U is not connected, A is contained in one of the connected components of U and $I^+(A)$ meets some other connected component.

As a corollary to Proposition 3.4, we have the following important result for the causal relations:

Corollary 4.1. Let $x, y, z \in M$. If $x \ll y$ and $y \leq z$, or $x \leq y$ and $y \ll z$, then $x \ll z$.

In terms of chronological and causal future of $A \subset M$, this gives $I^+(A) = I^+(J^+(A)) = J^+(I^+(A))$. As usual, the past versions behave in the same way.

From the definitions of causal and chronological futures we also immediately get the following corollary to Theorem 3.1:

Corollary 4.2. Let γ be a future-pointing causal curve from $A \subset M$ to $q \in J^+(A) \setminus I^+(A)$. Then γ is a null geodesic that does not meet $I^+(A)$ and has no conjugate points before q.

Proof. Let $p \in A$ be a point where γ starts from. Since $q \in J^+(A) \setminus I^+(A)$, there is, in particular, no timelike curve from p to q. Then Theorem 3.1 applied to the spacelike submanifold $\{p\}$ gives that γ is a null geodesic that has no conjugate points before q and stays outside $I^+(A)$. □

It is possible to deduce several topological properties of the sets defined by the causality relations. Connecting thereby the causal structure of a Lorentzian manifold to its topology has fundamental consequences for the global geometry of spacetime. Tracking down such features starts at the local level, from convex neighborhoods where the causal relations are basically reduced to those of Minkowski space.

Definition 4.1.1. Let $A \subset M$. We say that the relation $\leq$ is closed in A if for any $p, q \in A$ and $(p_n), (q_n) \subset A$ with $p_n \to p \implies q_n \to q$, then $p_n \leq q_n$ for all n implies that $p \leq q$.

Lemma 4.1. Let $C \subset M$ be a convex open set. Then

1. $I^+(p, C)$ is open in C.
2. $J^+(p, C)$ is the closure of $I^+(p, C)$ in C.
3. The causality relation $\leq$ is closed in C, with a stronger property that $p_n \to p, q_n \to q$, and $q_n \in J^+(p_n, C)$ $\forall n$ implies that $q \in J^+(p, C)$.

Proof. See [O'N83], Lemma 14.2. □

The first statement in the above lemma can be enhanced into a more general result:

Lemma 4.2. Let $p, q \in M$ be s.t. $p \ll q$. Then there are neighborhoods U of p and V of q s.t. $p' \ll q'$ for every $p' \in U$ and $q' \in V$.

Proof. See [O'N83], Lemma 14.3. □

Immediately, we obtain an important topological consequence:

Corollary 4.3. $I^+(A)$ is open for any $A \subset M$.

Lemma 4.3. Let $A \subset M$. Then

1. $\operatorname{int} J^+(A) = I^+(A)$,

2. $J^+(A) \subset \overline{I^+(A)}$, which are equal if and only if $J^+(A)$ is closed.

Proof. See [O'N83], Lemma 14.6. □

Typically, on physical grounds we impose conditions on the permissible causal structure of M. The *chronology condition* states that there are no closed timelike curves in M, which amounts to ruling out paradoxical time travel scenarios. Analogously, if there are no closed causal curves in M, we say that M satisfies the *causality condition*. Furthermore, we want to rule out manifolds that contain 'almost' closed causal curves, such as ones that enter any neighborhood of a point more than once. A closed causal curve could be obtained in a situation like that by an arbitrarily small perturbation of the manifold geometry. To exclude such pathologies, we may impose a further causality condition:

Definition 4.1.2. The *strong causality condition* holds at $p \in M$ if for any neighborhood U of p there exists a neighborhood $V \subset U$ of p s.t. all causal curve segments $\gamma : [a, b] \to M$ with endpoints $\gamma(a), \gamma(b) \in V$ satisfy $\gamma([a, b]) \subset U$.

Especially, compact subsets where strong causality is satisfied have various useful properties. For instance, if a sequence of future-pointing causal curve segments contained in a compact strongly causal K is such that the sequence of their starting points and the sequence of their endpoints converge to two different points $p, q \in K$, then there is a future-pointing causal broken geodesic from p to q (see [O'N83], Lemma 14.14). In particular, this implies that if K is such that for any points $p, q \in K$ with $p \leq q$ there is a causal curve α from p to q that is contained in K, then the causality relation $\leq$ is closed in K.

4.2 Global hyperbolicity and Cauchy hypersurfaces

There is yet another causality condition that is important enough to deserve a section of its own: *global hyperbolicity*. There are slightly different ways of defining global hyperbolicity but all the different but equivalent definitions capture the established notion of a causally 'well-behaved' spacetime. Here, as usual, we follow the definition adopted in [O'N83]. First, we introduce the shorthand notation $J(p, q) = J^+(p) \cap J^-(q)$. It is straightforward to see that $J(p, q)$ is the smallest set that contains every future-pointing causal curve from p to q. We call $J(p, q)$ the *causal diamond subtended by p and q*.

Definition 4.2.1 (Global hyperbolicity)**.** A set $H \subset M$ is globally hyperbolic if

1. the strong causality condition is satisfied on H, and
2. given $p, q \in H$ s.t. $p < q$, the set $J(p, q)$ is compact and $J(p, q) \subset H$.

The first nice property of globally hyperbolic sets is the following:

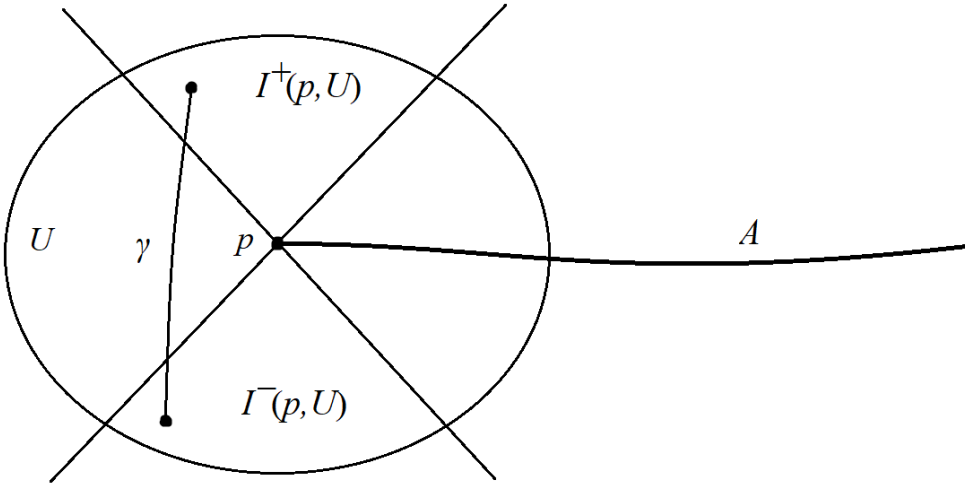

Figure 4.1: An illustration of edge(A).

Lemma 4.4. Given a globally hyperbolic open set $U \subset M$, the causality relation $\leq$ is closed in U.

Proof. See [O'N83], Lemma 14.22. □

A set $A \subset M$ is *achronal* if there is no timelike curve from p to q for any $p, q \in M$. Correspondingly, we say that A is *acausal* provided that there is no causal curve from p to q for any $p, q \in M$.

Definition 4.2.2. The *edge* of an achronal set A is defined as the set of points $p \in \overline{A}$ s.t. every neighborhood U of p contains a timelike curve $\gamma : [a, b] \to M$ s.t. $\gamma(a) \in I^-(p, U), \gamma(b) \in I^+(p, U)$, and $\gamma([a, b]) \cap A = \varnothing$.

We denote the edge of A by edge(A). For an illustration of the notion, see Fig. 4.1. For achronal topological hypersurfaces, closedness can be characterized in terms of emptyness of its edge:

Lemma 4.5. An achronal $A \subset M$ is a closed topological hypersurface if and only if edge$(A) = \varnothing$.

Proof. See [O'N83], Corollary 14.26. □

We introduce a further causal notion that will be useful later: $A \subset M$ is a *future set* if $I^+(A) \subset A$. Analogously with the previous lemma, it is possible to establish the following property of future sets:

Lemma 4.6. Let A be a future set. Then ∂A is a closed achronal topological hypersurface.

Proof. See [O'N83], Corollary 14.27. □

Next, we define a very special kind of subset, a *Cauchy hypersurface*, which turns out to be closely connected to global hyperbolicity – to the extent that sometimes global hyperbolicity is defined simply by the presence of such a hypersurface in M, see *e.g.* [Wal84].

Definition 4.2.3. Let $\Sigma \subset M$. Σ is a Cauchy hypersurface in M if every inextendible timelike curve in M intersects Σ once and only once.

The following lemma shows that not only each inextendible timelike curve meets Σ but also every inextendible causal curve does so. Moreover, since each inextendible timelike curve intersects Σ exactly once, it immediately follows that Σ is achronal. Then, as an achronal set that no timelike curve can pass without intersecting it, it has an empty edge, and Lemma 4.5 gives that it is a closed topological hypersurface:

Lemma 4.7. Any Cauchy hypersurface Σ is a closed achronal topological hypersurface that every inextendible causal curve intersects.

Proof. See [O'N83], Lemma 14.29. □

In particular, each maximal integral curve of a timelike vector field intersects a Cauchy hypersurface at a unique point. This makes it possible to define a retraction of the manifold onto the hypersurface:

Proposition 4.1. Let Σ be a Cauchy hypersurface in M and X be a timelike vector field on M. Given a point $p \in M$, a maximal integral curve of X through p intersects Σ at exactly one point $r(p)$. Furthermore, $r : M \to \Sigma$ is a continuous open surjection s.t. $r|_\Sigma = \mathrm{id}_\Sigma$, and Σ is connected.

Proof. See [O'N83], Proposition 14.31. □

Definition 4.2.4. Let $A \subset M$ be an achronal set. The *future domain of dependence* $D^+(A)$ of A is the set of points $p \in M$ s.t. each past-inextendible causal curve through p intersects A.

The *past domain of dependence* $D^-(A)$ is defined analogously. The *domain of dependence* $D(A)$ of A is the union of these: $D(A) = D^-(A) \cup D^+(A)$. For an illustration, see Fig. 4.2.

Theorem 4.1. Let $A \subset M$ be achronal. Then $\operatorname{int} D(A)$ is globally hyperbolic.

Proof. See [O'N83], Theorem 14.38. □

Corollary 4.4. If M contains a Cauchy hypersurface, then M is globally hyperbolic.

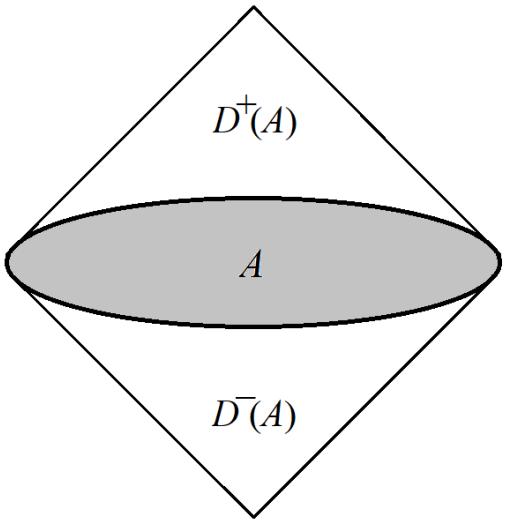

Figure 4.2: An illustration of the domain of dependence $D(A) = D^-(A) \cup D^+(A)$.

Proof. Let Σ be a Cauchy hypersurface in M. Then Σ is achronal and the domain of dependence $D(\Sigma) = M$. Thus, $\operatorname{int} D(\Sigma) = \operatorname{int} M = M$. □

4.3 Trapped surfaces and focusing of null geodesics

We are finally in a position to relate the extrinsic geometry of submanifolds to causality on a Lorentzian manifold M. First, we define:

Definition 4.3.1. A spacelike submanifold of M is *future-converging* if its mean curvature vector field H is past-pointing timelike.

Remark. The mean curvature vector field of $S \subset M$ being past-pointing timelike is equivalent to the condition that the convergence of S satisfies $\mathbf{k}(l) = \langle H, l \rangle > 0$ for every future-pointing null vector l normal to S, which in turn is equivalent to the statement that $\mathbf{k}(v) = \langle H, v \rangle > 0$ for every future-pointing causal vector v normal to S.

Then we define the notion that is the centerpiece of Penrose's theorem:

Definition 4.3.2 (Trapped surface)**.** A compact spacelike codimension 2 submanifold of M is a *trapped surface* if it is future-converging.

Given a set $A \subset M$, we denote $E^+(A) := J^+(A) \setminus I^+(A)$.

Definition 4.3.3. Let A be a closed subset of M.[1] If $E^+(A)$ is compact, we say that A is

[1] In [O'N83], A is also supposed to be achronal. This assumption, however, is never used in the proof of the proposition below so we have dropped it from the definition.

future-trapped.

A past-trapped set is defined analogously.

Next, we prove a lemma that will be needed in the main proposition of this section:

Lemma 4.8. Let S be a compact codimension 2 submanifold of the Lorentzian manifold M. Let $\|\cdot\|$ be the norm induced by a Riemannian metric on M. Then the subset $\widetilde{S} \subset NS$ of the normal bundle defined by

$$\widetilde{S} = \{(p,k) \in NS : \|k\| = 1,\ k \text{ is null}\}$$

is compact.

Proof. Let $\widetilde{S}$ be the set defined above, and let (v_i) be a sequence in $\widetilde{S}$. Mapping each element in the sequence to S by the canonical projection, we obtain a sequence $(\pi(v_i)) \subset S$. But S is compact so there is a subsequence $(\pi(v_j)) \subset (\pi(v_i))$ s.t. $\pi(v_j) \to p \in S$. Since NS is a 2-vector bundle, there is an open $U \subset S$ with $p \in U$ and a diffeomorphism $\phi : U \times \mathbb{R}^2 \to \pi^{-1}(U)$ s.t. $L_q : \vec{v} \mapsto \phi(q,\vec{v})$ is a linear isomorphism $\mathbb{R}^2 \to \pi^{-1}(q)$ for each $q \in U$. Let now $(v_k) \subset (v_j)$ s.t. $\pi(v_k) \in U$ for all k. Then we can define a diffeomorphism $\tilde{\phi} : U \times S^1 \to \pi^{-1}(U) \cap A$, where $S^1 \subset \mathbb{R}^2$ is the unit circle and $A = \{(p,k) \in NS : \|k\| = 1\}$, as follows: for each $(q,\vec{v}) \in U \times S^1$, there is a unique $v \in \pi^{-1}(q) \cap A$ s.t. $\vec{v} = L_q^{-1}(v)/|L_q^{-1}(v)|$ where $|\cdot|$ is the standard norm in $\mathbb{R}^2$, and we set $\tilde{\phi}(q,\vec{v}) = v$. Now since $\widetilde{S} \subset A$, there are vectors $\vec{v}_k \in S^1$ s.t. $\tilde{\phi}(\pi(v_k),\vec{v}_k) = v_k$ for all k. But S^1 is compact so there is a subsequence $(\vec{v}_l) \subset (\vec{v}_k)$ s.t. $\vec{v}_l \to \vec{v} \in S^1$. Then, since $(\pi(v_l),\vec{v}_l) \to (p,\vec{v})$ and each $\tilde{\phi}(\pi(v_l),\vec{v}_l) = v_l \in \widetilde{S}$, which is closed, we have $v_l \to v \in \widetilde{S}$. □

Proposition 4.2. Let M be a future null complete manifold. Suppose that $\mathrm{Ric}(k,k) \geq 0$ for all null tangent vectors k to M. Then any trapped surface in M is future-trapped.

As the proposition plays a pivotal role in the proof of Penrose's theorem, instead of referring the reader to [O'N83] we shall give a full proof of the proposition.

Proof. Let $S \subset M$ be a trapped surface. It is thus compact by definition and hence closed. We then need to show that $E^+(S)$ is compact. Let $\widetilde{S}$ be a subset of the normal bundle NS as in the previous lemma. From Proposition 3.3 it follows that for every $l \in \widetilde{S}$ there exists a focal point of S on the geodesic $\gamma_l|_{[0,1/\mathbf{k}(l)]}$ with $\mathbf{k}(l) = \langle H, l\rangle > 0$ since S is future-converging. Since $\widetilde{S}$ is compact and $\mathbf{k}$ is continuous, there is a minimum $k_{\min} \in \mathbf{k}(\widetilde{S})$ and hence for each $l \in \widetilde{S}$ there exists a focal point of S on $\gamma_l|_{[0,b)}$ where $b > 1/k_{\min}$.

Let then $q \in E^+(S)$. By Corollary 4.2, there exists a null geodesic γ from S to q. Theorem 3.1 implies that, since $q \notin I^+(S)$, γ is normal to S and does not have any

focal points of S before q. Hence, γ is obtained by reparametrizing some γ_v with $v \in \widetilde{S}$. Therefore, $E^+(S) \subset \exp(K)$ where

$$K = \{tv \in NS : v \in \widetilde{S},\, 0 \leq t \leq b\}. \tag{4.3.1}$$

Compactness of $\widetilde{S}$ implies that K is compact and thus $\exp(K)$ is also compact. Then we only need to show that $E^+(S)$ is a closed subset of the compact set $\exp(K)$. Let $p \in \overline{E^+(S)}$ and let U be a neighborhood of p. Then $U \cap \exp(K) \neq \varnothing$ and since $\exp(K)$ is closed (as a compact subset in a Haussdorff space), $p \in \exp(K)$. But $\exp(K) \subset J^+(S)$ so either $p \in I^+(S)$ or $p \in J^+(S) \setminus I^+(S) = E^+(S)$. Suppose the former is the case. But then, since $I^+(S)$ is open, p would have a neighborhood with an empty intersection with $E^+(S)$, which contradicts the fact that p belongs to the closure of $E^+(S)$. Therefore, we must have $p \in E^+(S)$. □

Remark. Proposition 4.2 supposes that M is future null complete without further qualifications. The hypothesis actually used in the proof is that the null geodesics normal to S, in particular, are extendible beyond the affine parameter value $1/k_{\min}$. Formulating the proposition in this way allows us to strengthen the conclusion of Theorem 4.2.

4.4 Penrose's theorem

In general relativity, realistic matter content of spacetime is supposed to satisfy certain energy conditions that, generally speaking, only allow for positive energy. In particular, the stress-energy tensor is expected to satisfy the *null energy condition*, which states that $T(k,k) \geq 0$ for all null vectors k. With Einstein's equation (2.3.4), this condition for matter can be translated into a geometric condition $\mathrm{Ric}(k,k) \geq 0$, which by the result of the previous section leads to focusing of null geodesics normal to a trapped surface. In conjunction with assumptions on the topology and causal structure of the Lorentzian manifold M, this entails that M is singular in the sense of null geodesic incompleteness.

Theorem 4.2 (Penrose's theorem)**.** Suppose the following conditions hold:

1. $\mathrm{Ric}(k,k) \geq 0$ for all null vectors k to M.
2. M has a non-compact Cauchy hypersurface Σ.
3. M contains a trapped surface S.

Then M is future null incomplete.

Proof. By Corollary 4.4, M is globally hyperbolic since it contains a Cauchy hypersurface. Then we want to show that $J^+(S)$ is closed in M. To this end, let $(q_n) \subset J^+(S)$ s.t. $q_n \to q \in M$. Then there is a sequence (p_n) in S with $p_n \leq q_n$ for each n, and since S is compact, there is a convergent subsequence (p_m) with $p_m \to p \in S$. Let (q_m) be the corresponding subsequence of (q_n) that converges to q. But now since $p_m \leq q_m$ for

each m, Lemma 4.4 implies that $p \leq q$. Hence $q \in J^+(p) \subset J^+(S)$ and we conclude that $J^+(S)$ is closed.

Lemma 4.3 implies that $\operatorname{int} J^+(S) = I^+(S)$ and $J^+(S) = \overline{I^+(S)}$ so $E^+(S) = \overline{J^+(S)} \setminus \operatorname{int} J^+(S) = \partial J^+(S)$, which is the boundary of a future set. By Lemma 4.6, $E^+(S)$ is hence a topological hypersurface. Furthermore, by Proposition 4.2, S is future-trapped, *i.e.* $E^+(S)$ is compact.

Next, we show that the compactness of $E^+(S)$ contradicts the non-compactness of the Cauchy hypersurface Σ. Let $r : M \to \Sigma$ be an open continuous map with $r|_\Sigma = \mathrm{id}_\Sigma$ given in Proposition 4.1 by a timelike vector field X in M. Let $\rho = r|_{E^+(S)}$ be the restriction of r to $E^+(S)$. Since $E^+(S)$ is achronal, each integral curve of X intersects it at most once, which means that ρ is injective. Above $E^+(S)$ was shown to be a topological hypersurface and, by Corollary 4.7, Σ is also a topological hypersurface. The invariance of domain then implies that $\rho(E^+(S))$ is an open subset of Σ. Continuity of ρ, on the other hand, implies that the image $\rho(E^+(S))$ of a compact set is compact and hence closed since Σ is Haussdorff. As a Cauchy hypersurface, Σ is connected, and since $E^+(S)$ is nonempty, $\rho(E^+(S))$ can be both open and closed in Σ only if $\rho(E^+(S)) = \Sigma$. But this is a contradiction since Σ was supposed to be non-compact. □

Remark. Formulated this way, the theorem seems to say nothing more than that there is some – at least one – incomplete null geodesic. By the remark above, the conclusion can be sharpened to at least identify the incomplete geodesic as one of those normal to the trapped surface S. Since $\langle H, l\rangle > 0$ for null vectors normal to S, through small deformations of S in the Cauchy hypersurface Σ new trapped surfaces can be generated. Penrose's theorem can then be applied to these surfaces to deduce the existence of other incomplete null geodesics, actually an infinite family thereof.

Remark. As a trapped surface is a key ingredient of Penrose's theorem, it is important to understand when such surfaces exist in spacetime. In exact solutions of Einstein's equation, trapped surfaces can be found, for instance, inside a Schwarzschild black hole. In general, if enough matter is concentrated in a sufficiently small region, occurrence of trapped surfaces is guaranteed even in the presence of large deviations from spherical symmetry [SY83]. A natural next question then is whether trapped surfaces can form under a time evolution from generic initial data and, in particular, initial data that is regular and contains no trapped surfaces. For a proof that this can happen in, *e.g.*, vacuum spacetimes, see [Chr08].

Remark. The singularity deduced in the theorem is understood in terms of (null) geodesic incompleteness – some light ray on the manifold comes to a sudden end. Whether there is a curvature singularity associated with (and, furthermore, responsible for) the abrupt end of the light ray, is not directly addressed by the theorem. For a recent analysis of

curvature singularities in geodesically incomplete spacetimes, see [RÓ23].

Remark. Penrose's theorem is sometimes characterized as a 'black hole singularity theorem.' Even though it is a widespread expectation that a process leading up to the formation of a trapped surface generates a black hole, a region causally separated from the rest of spacetime by an event horizon, one should notice that the theorem does not say that the singularity is necessarily veiled by an event horizon. That naked singularities should not be formed in generic, physically realistic situations, is stated in the Weak Cosmic Censorship Conjecture, which is a major open problem in mathematical relativity [Col19].

Bibliography

[Chr08] Demetrios Christodoulou. The Formation of Black Holes in General Relativity. In *12th Marcel Grossmann Meeting on General Relativity*, pages 24–34, 5 2008.

[Col19] Alan A. Coley. Mathematical General Relativity. *Gen. Rel. Grav.*, 51(6):78, 2019.

[HE23] Stephen W. Hawking and George F. R. Ellis. *The Large Scale Structure of Space-Time*. Cambridge Monographs on Mathematical Physics. Cambridge University Press, 2 2023.

[Lee97] J.M. Lee. *Riemannian Manifolds: An Introduction to Curvature*. Graduate Texts in Mathematics. Springer New York, 1997.

[Lee03] J.M. Lee. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer, 2003.

[O'N83] B. O'Neill. *Semi-Riemannian Geometry With Applications to Relativity*. ISSN. Elsevier Science, 1983.

[OS39] J. R. Oppenheimer and H. Snyder. On Continued gravitational contraction. *Phys. Rev.*, 56:455–459, 1939.

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

[Poi09] Eric Poisson. *A Relativist's Toolkit: The Mathematics of Black-Hole Mechanics*. Cambridge University Press, 12 2009.

[RÁ23] István Rácz. Spacetime singularities and curvature blow-ups. *Gen. Rel. Grav.*, 55(1):3, 2023.

[SG15] José M. M. Senovilla and David Garfinkle. The 1965 Penrose singularity theorem. *Class. Quant. Grav.*, 32(12):124008, 2015.

[SY83] Richard Schoen and S T Yau. The existence of a black hole due to condensation of matter. *Communications in Mathematical Physics*, 90:575–579, 1983.

[Wal84] Robert M. Wald. *General Relativity*. Chicago Univ. Pr., Chicago, USA, 1984.