MASTERARBEIT | MASTER'S THESIS

Titel | Title

Null Geometry and the Penrose Incompleteness Theorem

verfasst von | submitted by

Paul-Aaron Haberger B.Sc.

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 876
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Physics
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Roland Steinbauer

Abstract

In this work we prove the classical version of the Penrose Incompleteness Theorem and a somewhat different variant. Apart from the standard argument based on Differential Topology and the Variational Theory of Curves, our approach uses results on the structure of null hypersurfaces and achronal boundaries. In particular, we develop methods to rigorously study the geometry of null submanifolds. The classical part of the argument is also presented in full detail by reviewing the necessary mathematics. Moreover, the thesis is entirely self-contained and we thereby provide a complete account of all the necessary background in Lorentzian Geometry.

In dieser Arbeit beweisen wir die klassische Version des Unvollständigkeitssatzes von Penrose sowie eine seiner anderen Varianten. Neben dem klassischen Argument, das auf Methoden der Differentialtopologie und der Variationstheorie von Kurven basiert, stützt sich unser Zugang auf Resultate über die Struktur nullartiger Hyperflächen und achronaler Ränder. Insbesondere entwickeln wir Methoden zur präzisen Untersuchung der Geometrie nullartiger Untermannigfaltigkeiten. Der klassische Teil des Arguments wird vollständig und im Detail präsentiert, wobei die dafür erforderlichen mathematischen Grundlagen hergeleitet werden. Darüber hinaus ist die Arbeit in sich abgeschlossen und enthält eine umfassende Darstellung sämtlicher notwendiger Voraussetzungen aus der Lorentzschen Geometrie.

Contents

1 Introduction

The *Penrose Incompleteness Theorem*, also more commonly known as the *Penrose Singularity Theorem*, arguably poses one of the most remarkable achievements of Mathematical Relativity and Lorentzian Geometry. In his groundbreaking paper [Pen65], Sir Roger Penrose was able to prove a geometrical result, which had a profound impact on the modern understanding of Einstein's Theory of Relativity.

The statement of the Theorem concerns the property of a spacetime manifold being null geodesically incomplete. In rough terms, this means that there exists a null geodesic (possibly more), which cannot be extended to infinite parameter time. Penrose was the first to demonstrate that assuming (i) Existence of a non-compact Cauchy Surface, (ii) Null Energy Condition, and (iii) Presence of a Trapped Surface, that a spacetime must be null incomplete. This original argument hinges on results from Differential Topology and Variation of Curves in the setting of Lorentzian Manifolds.

In the set of notes by Gregory J. Galloway [Gal14], one can find outlines for a proof, where one mainly uses the structure of null hypersurfaces and achronal boundaries. This offers a very intuitive approach, setting it somewhat apart from Penrose's original argument and other proofs found in the textbook literature.

The aim of this Master Thesis is to provide a complete and mathematically rigorous account of the ideas presented in [Gal14], with the goal of proving the celebrated incompleteness theorem. In addition, another related incompleteness result is covered which lends itself to further applications. In order to achieve this, rigorous mathematics from Lorentzian geometry, Topology and Analysis will be established and applied.

To this end, Chapter 2 serves as an introduction to Lorentzian Manifolds and Causality Theory. In particular, we will display the differences between Lorentzian and Riemannian Geometry and establish several important results, as well as present the main object we want to work with: the *spacetime* (M, g). In Chapter 3, we define Semi-Riemannian submanifolds as smooth submanifolds that inherit a (non-degenerate) metric tensor and then study their induced connection. After this, in Chapter 4, we study causal curves and prove the limit curve lemma, which is a very helpful technical result in Lorentzian Geometry concerning sequences of causal curves. The latter part of the Chapter is dedicated to notions of Variational Calculus. While this is very similar to the Riemannian case, one needs to be careful in adapting standard arguments in a way such that the underlying causal structure of the Lorentzian manifold is respected. Chapter 5 deals with global topological properties in Lorentzian Geometry. In the first subsection we introduce achronal boundaries, which are formally defined as sets of the form $\partial I^{\pm}(S)$, i.e. the topological boundary of the chronological future of an arbitrary subset $S \subseteq M$. Moreover, we will establish that achronal boundaries actually carry the structure of an achronal, topological hypersurface that is generated by null geodesics. The second subsection is a continuation of our study of Causality Theory. In particular, we introduce a causal ladder, define Cauchy surfaces and Global Hyperbolicity and study how these notions relate to one another. As a central result, we verify that compact, achronal

topological hypersurfaces in globally hyperbolic spacetimes are Cauchy surfaces.
After this, we are again interested in submanifolds of spacetimes. However, unlike in Chapter 3, we will study cases where the spacetime metric does not induce a non-degenerate metric. These kind of submanifolds, called *null submanifolds*, are the main objects of interest in Chapter 6. Our goal is to understand the geometry of null hypersurfaces, which becomes important when studying smooth parts of achronal boundaries.
A proper analysis of variations of null geodesics then is subject of the penultimate Chapter 7. There, we combine the results of Chapters 3 and 4 and introduce the notion of *focal points* of submanifolds along null geodesics. In particular, will show how focal points are related to singular values of the normal exponential map and what impact they have on the causal structure.
Finally, in Chapter 8, we first introduce trapped surfaces and geodesic null congruences. Following this, we give an overview of Einstein's Theory of General Relativity and its relation to Penrose's Incompleteness Theorem. The third subsection is where we discuss the proof of the Incompleteness Theorem and a subsequent result thereof.

2 Fundamental Notions of Lorentzian Geometry

The history of Lorentzian geometry and, more generally, Semi-Riemannian geometry dates back to the early 20th century, where it founded its central significance in relation to physics. In particular, it is a Lorentzian manifold that is the framework in which *Albert Einstein* $(1879-1955)$ constructed his theory of General Relativity. Building up on *Bernhard Riemann* $(1826-1866)$ and his work in geometry, Lorentzian geometry has since become an active field of research in mathematics and mathematical physics.
The purpose of this chapter is to give a brief and sufficient introduction to fundamental notions and key results in Lorentzian geometry. The reader is expected to have some knowledge in differential geometry and Riemannian geometry, approximately on the level of e.g. [Kun22a] and [KS21]. We will however recall some basic definitions down below.

2.1 Semi-Riemannian Manifolds

Our first goal is to introduce the notion of a *Lorentzian manifold*. After that, we will briefly collect several important facts that are needed later. In this section we will follow [O'N83], [KS21] and [KS23] as our main references. Note that we will also adopt their notations and conventions, unless otherwise specified. We may now proceed with the first definition.

Definition 2.1.1. *A semi-Riemannian manifold is a pair (M, g), where M is a smooth, second countable Hausdorff manifold and g is a smooth non-degenerate $(0, 2)$-tensor field on M (i.e. $g \in \mathcal{T}^0_2(M)$) that is symmetric and has constant signature*[1].

[1] Recall that a signature is a pair (k, l), where the letters, in this order, denote the maximum dimension of a subspace W for which $g|_W$ is 1) negative definite, 2) positive definite.

The tensor field g – also called *metric tensor* or *metric* – therefore yields a map $g_p := g(p) : T_pM \times T_pM \to \mathbb{R}$ for any $p \in M$, which is symmetric, non-degenerate and has constant signature respectively. Note that we will also often use the notation $\langle \cdot, \cdot \rangle := g(\cdot, \cdot)$. One can distinguish the following cases

Definition 2.1.2. *Let* (M, g) *be a semi-Riemannian manifold with* $\dim(M) = n + 1$. *We call* (M, g) *a*

(i) ***Lorentzian manifold*** *if* g *has signature* $(1, n)$,

(ii) ***Riemannian manifold*** *if* g *has signature* $(n + 1, 0)$.

As the most simple and important example of a Lorentzian manifold, one may consider *Minkowski space* $(\mathbb{R}^{1,n}, \eta)$, where $\mathbb{R}^{1,n} := \mathbb{R}^{n+1}$ and $\eta = \operatorname{diag}(-1, 1, \dots, 1)$. In physics, Minkowski space with $n = 3$, provides the geometric framework underlying the special relativity.
It will be useful to review some of the differences and similarities of Lorentzian and Riemannian manifolds. We will see that if (M, g) is any kind of semi-Riemannian manifold, several notions from Riemannian geometry generalize to (M, g). Most importantly, M will always admit a *Levi-Civita connection* $\nabla : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$. With this, one can define the *Riemann curvature tensor* $R : \mathfrak{X}(M) \times \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$ as usual by $R(Z, X, Y) := R_{XY}Z := \nabla_{[X,Y]}Z - [\nabla_X, \nabla_Y]\, Z$. In local coordinates we will write $R^a_{bcd}\partial_a = R_{\partial_c \partial_d}\partial_b$ and define the *Ricci curvature tensor* Ric as the contraction $R^a_{ija} = R_{ij}$.
A smooth curve $\gamma : I \to M$ is a smooth map from a connected subset I of $\mathbb{R}$ that contains at least two points to M. Similarly, a *piecewise smooth curve* $\gamma : [a, b] \to M$ is a continuous map that admits a finite partition $a = t_0 < t_1 < \cdots < t_n = b$ of $[a, b]$ such that $\gamma|_{[t_i, t_{i+1}]}$ is smooth for each $i = 1, \dots, n-1$. Note that by a curve we will usually mean a piecewise smooth curve, if not otherwise specified. In order to talk about the velocities of curves, following notion is necessary

Definition 2.1.3. *Let* N, M *be smooth manifolds and* $f : N \to M$ *smooth. A vector field along* f *is a smooth map* $X : N \to TM$ *such that* $\pi \circ X = f$, *i.e.*

$$\begin{tikzcd} TM \arrow[r, "\pi"] & M \\ N \arrow[u, "X"] \arrow[ur, dashed, "f"'] & \end{tikzcd}$$

commutes, where $\pi : TM \to M$ *is the bundle projection of* TM.
The set of vector fields along f *is denoted by* $\mathfrak{X}(f)$.

Remark 2.1.4. (i) The set $\mathfrak{X}(f)$ together with point-wise addition and multiplication is a module over $\mathcal{C}^\infty(N)$.

(ii) Let N, M with $\dim(M) = m$, $\dim(N) = n$ and f be as above and consider $X \in \mathfrak{X}(f)$. Let $V \subseteq N$ open and (U, φ) be a chart in M such that $f(V) \subseteq U$. If $p \in V$, then we can write $X(p) = X^i(p)\, \partial_i|_{f(p)}$ with $X^i \in \mathcal{C}^\infty(N)$ for each $i = 1, \dots, m$.

(iii) If $\gamma : I \to M$ is a smooth curve on a smooth manifold M, then $\gamma' \in \mathfrak{X}(\gamma)$. In local coordinates $(U, \varphi = (x^1, \dots, x^n))$, we have $\gamma'(t) = \frac{d(x^i \circ \gamma)}{dt}(t)\ \partial_i|_{\gamma(t)}$. γ' is called the *velocity vector field* of γ and is often also denoted as $\dot{\gamma}$. Note in particular that if $X \in \mathfrak{X}(M)$, then $X \circ \gamma \in \mathfrak{X}(\gamma)$.

(iv) Let $\gamma : [a, b] \to M$ be a curve and $V : [a, b] \to TM$ a continuous map with $\pi \circ V = \gamma$. V is called a *piecewise smooth vector field* if there is a finite partition $a = t_0 < t_1 < \dots < t_n = b$ of $[a, b]$ such that $V|_{[t_i, t_{i+1}]}$ is smooth for all $i = 1, \dots, n-1$. In this case, we write $V \in \mathfrak{X}_{\mathrm{pw}}(\gamma)$. In particular, for any piecewise smooth curve γ one has $\gamma' \in \mathfrak{X}_{\mathrm{pw}}(\gamma)$.

We may now define the *covariant derivative* of a vector field $Z \in \mathfrak{X}(\gamma)$ along a smooth curve $\gamma : I \to M$. For this, only the Levi-Civita connection is needed and thus, one can proceed as in the case of a Riemannian manifold.

Proposition 2.1.5. *Let $\gamma : I \to M$ be a smooth curve. There exists a unique map*

$$\begin{aligned} \frac{\nabla}{dt} &: \mathfrak{X}(\gamma) \to \mathfrak{X}(\gamma) \\ Z &\mapsto \dot{Z} := Z' := \frac{\nabla Z}{dt} \end{aligned}$$

called induced covariant derivative, which satisfies

(i) $(\lambda Z_1 + \mu Z_2)' = \lambda Z_1' + \mu Z_2'$ *for all* $\lambda, \mu \in \mathbb{R}$ *and* $Z_1, Z_2 \in \mathfrak{X}(\gamma)$*,*

(ii) $(fZ)' = \frac{df}{dt} Z + f Z'$ *for all* $f \in \mathcal{C}^\infty(I)$ *and* $Z \in \mathfrak{X}(\gamma)$*,*

(iii) $(X \circ \gamma)'(t) = \nabla_{\gamma'(t)} X$ *for all* $X \in \mathfrak{X}(M)$ *and* $t \in I$*,*

(iv) $\frac{d}{dt} \langle Z_1, Z_2 \rangle = \langle Z_1', Z_2 \rangle + \langle Z_1, Z_2' \rangle$ *for all* $Z_1, Z_2 \in \mathfrak{X}(\gamma)$*.*

The proof proceeds in the exact same way as it would in Riemannian geometry (E.g. see [O'N83, Prop. 3.18]). Now, let γ and Z be as above. Using the previous remark and the properties of the induced covariant derivative, the local coordinate expression for Z' can be written as

$$Z' = \sum_{i=1}^n \frac{dZ^i}{dt} (\partial_i \circ \gamma) + \sum_{i=1}^n Z^i\, \nabla_{\gamma'} \partial_i = \sum_{i=1}^n \left(\frac{dZ^i}{dt} + \Gamma^i_{jk} \frac{d(x^j \circ \gamma)}{dt} Z^k \right) (\partial_i \circ \gamma). \tag{2.1}$$

Let us recall two important properties of the covariant derivative, which are needed later on.

Lemma 2.1.6. *Let $f : I \times J \to M$ smooth and $Z \in \mathfrak{X}(f)$. We then have* *(i) in first order*

$$\frac{\nabla}{\partial s}\frac{\partial f}{\partial s} = \frac{\nabla}{\partial t}\frac{\partial f}{\partial t}$$

(ii) and in second order

$$\left(\frac{\nabla}{\partial t}\frac{\nabla}{\partial s} - \frac{\nabla}{\partial s}\frac{\nabla}{\partial t}\right) Z = R_{f_s f_t} Z,$$

where $f_s = \frac{\partial f}{\partial s}$ *and* $f_t = \frac{\partial f}{\partial t}$.

For a proof of this result see [KS21, Lemma 2.1.20] or [O'N83, Prop. 4.44].

Let $\gamma : I \to M$ be a curve on M. We will call a vector field $Z \in \mathfrak{X}(\gamma)$ *parallel* if $Z' = 0$. Because equation (2.1) is a system of linear ordinary differential equations of first order, it follows that for initial data $z \in T_{\gamma(t_0)}M$ with $t_0 \in I$, one obtains a unique parallel vector field $Z \in \mathfrak{X}(\gamma)$ with $Z(t_0) = z$. This solution Z is called the *parallel transport of z along γ*. The *parallel transport* along γ from $\gamma(t_1)$ to $\gamma(t_2)$ then is the map $\mathcal{P}^{\gamma}_{t_1,t_2} : T_{\gamma(t_1)}M \to T_{\gamma(t_2)}$ defined by $\mathcal{P}^{\gamma}_{t_1,t_2}(z) = Z(t_2)$, where Z is the parallel transport of z along γ. One may show that $\mathcal{P}^{\gamma}_{t_1,t_2}$ is a linear isometry (see e.g. [O'N83, Lemma 3.20] or Proposition 3.2.14), which is a fact that will be used several times.

A *geodesic* c is a smooth curve $c : I \to M$ with a parallel velocity vector field. It thus solves the *geodesic equation*[2]

$$\ddot{c} = 0. \tag{2.2}$$

This equation can locally be written down as a system of ordinary differential equations of second order using equation (2.1) and the local coordinate expression of $\dot{c}$. Consequently, it always admits a unique maximal solution $c : I \to M$ for initial data $p \in M$ and $v \in T_pM$, i.e. $c(0) = p$, $\dot{c}(0) = v$ and any other geodesic $\tilde{c} : J \to M$ with same initial data satisfies $J \subseteq I$ (cf. [KS21, Chapter 2.1]).

A *pregeodesic* is a smooth curve $c : I \to M$ with $\ddot{c} = \varphi\dot{c}$ for some $\varphi \in \mathcal{C}^\infty(I, \mathbb{R})$. Any such pregeodesic admits a reparametrization as a geodesic, and while any reparametrization of a non-constant geodesic c is a pregeodesic, only for $h(t) = at + b$ with $a, b \in \mathbb{R}$ follows that $c \circ h$ is again a geodesic (cf. [KS23, Remark 3.2.2] and [O'N83, Lemma 3.26]).

The *exponential map at $p \in M$* is defined as $\exp_p : \mathcal{D}_p \subseteq T_pM \to M$ with $\exp_p(v) = c_v(1)$, where c_v is the geodesic with $c(0) = p$ and $\dot{c}(0) = v$ and $\mathcal{D}_p$ is the maximal domain of definition, which is open[3]. More generally, the *exponential map* $\exp : \mathcal{D} \subseteq TM \to M$ is given by $\exp(v) \coloneqq \exp_{\pi(v)}(v)$ with $\pi : TM \to M$ being the bundle projection of TM, and $\mathcal{D}$ such that $\mathcal{D}_p = \mathcal{D} \cap T_pM$ for each $p \in M$.

Recall[4] that for any $p \in M$ there exists a *normal neighbourhood* U of p such that $\exp_p : \tilde{U} \to U$

[2] We will primarily use the $\ddot{c}$ and $\dot{c}$ notation for geodesics instead of c'' and c'.

[3] See e.g. [O'N83, Cor. 5.4].

[4] Cf. [O'N83, Prop. 3.30 & Prop. 3.31]

is a diffeomorphism from a neighbourhood $\tilde{U}$ of $0 \in T_pM$ that is *star shaped*, i.e. if $v \in \tilde{U}$ then also $tv \in \tilde{U}$ for any $t \in [0,1]$. Note that any $q \in U$ can be connected to p via a unique geodesic $c : [0,1] \to U$, also called a *radial geodesic*, with $\dot{c}(0) = \exp_p^{-1}(q) \in \tilde{U}$.
Moreover, a normal neighbourhood may be used to define *Riemannian normal coordinates* (also called *normal coordinates*) $(x^1, \ldots, x^{n+1})$ around p on M with $\dim(M) = n+1$ (cf. [O'N83, Prop. 3.33]). In these coordinates one has $g|_p = \eta$ and $\Gamma^a_{bc}|_p = 0$, where η is the Minkowski metric and $\Gamma^a_{bc}\partial_a = \nabla_{\partial_b}\partial_b$. Additionally, if $\{e_0, e_1, \ldots, e_n\}$ is an orthonormal basis of T_pM, we may define the vector fields $E_0, E_1, \ldots, E_n$ in a normal neighbourhood U of p by parallel transporting e_i along every radial geodesic emanating from p for each $i = 0, \ldots, n$ respectively. By the properties of parallel transport, $\{E_0, E_1, \ldots, E_n\}$ must remain orthonormal at any $q \in U$. In this way, for any $p \in M$, one can always consider a local orthonormal frame around p.
An open set $C \subseteq M$ is called *convex* or *totally normal* if it is a normal neighbourhood of all of its points. The following holds true[5]

Proposition 2.1.7. *Any point $p \in M$ admits a neighbourhood basis that consists of convex sets.*

So far, we have seen that many notions and results known from Riemannian geometry easily generalize to the semi-Riemannian setting, which actually reflects the significance of the non-degeneracy of g. However, the fact that the metric is not positive definite for Lorentzian manifolds also has severe consequences. Recall that we call a Riemannian manifold (M, g) complete if all maximal geodesics are defined on all of $\mathbb{R}$. The *Hopf-Rinow* Theorem then ties this notion of completeness to completeness of the metric space (M, d_g), where d_g is the Riemannian distance function[6]. More generally, the following holds

Theorem 2.1.8 (Hopf-Rinow). *Let (M, g) be a connected Riemannian manifold. The following are equivalent*

(i) (M, d_g) *is complete, i.e. every Cauchy sequence converges.*

(ii) There exists $p \in M$ such that $\exp_p$ is defined on all of T_pM.

(iii) (M, g) *is geodesically complete.*

(iv) Every closed and bounded subset of M is compact.

For a proof and a proper discussion we refer to [O'N83, Theorem 5.21]. While the above Theorem has found a myriad of applications, it may be used in particular, to establish the following crucial fact (cf. [NO61])

Theorem 2.1.9 (Nomizu, Ozeki). *Any connected $\mathcal{C}^\infty$ manifold M admits a complete Riemannian metric.*

[5] See [KS21, Thm. 2.2.7] for a proof.
[6] Cf. [O'N83, Chapter 5]

To contrast this, any smooth, second countable Hausdorff manifold M that is compact and has non-vanishing Euler characteristic does not even admit a Lorentzian metric[7]. Furthermore, defining geodesic completeness of semi-Riemannian manifolds in the same way as for the Riemannian case, there is no result akin to the Hopf-Rinow Theorem in the Lorentzian setting. In fact, there is no canonical way to define a metric (in the metric space sense) on a Lorentzian manifold at all. Completeness is therefore a much more elusive issue in the Lorentzian setting.

2.2 Causality Theory

In this second subsection, we will introduce a very crucial aspect intrinsic to Lorentzian manifolds (M, g): *Causality Theory*. Because of the indefiniteness of a Lorentzian metric g, it is possible to distinguish so-called *causal curves* $\gamma : I \to M$, for which $g(\dot{\gamma}(t), \dot{\gamma}(t)) \leq 0$ for every $t \in I$. Points $p, q \in M$ that can be connected via such causal curves are called *causally related*. Causality theory now refers to the study of questions surrounding the issue which points in M are causally related. This is of particular interest in physics: Special relativity tells us that curves, which model the trajectory of particles or other objects, cannot surpass the speed of light. Points that can only be reached by higher velocities are thereby not causally related and cannot be influenced (in the physical sense). But apart from its significance in physics, the causal structure of a Lorentzian manifold turns out to be deeply interconnected with its geometry. In particular, causality theory provides the necessary tools to answer whether or not two points can be connected by a length maximizing geodesic. We will follow [O'N83], [Bä20] and [KS23] as the main references for this subsection.
Let (M, g) be a Lorentzian manifold. Because the metric g is not positive definite but indefinite, one may distinguish the following cases

Definition 2.2.1. *Let $p \in M$. A vector $v \in T_pM$ with $v \neq 0$ is called*

(i) ***spacelike*** *if $\langle v, v\rangle > 0$ or $v = 0$,*

(ii) ***timelike*** *if $\langle v, v\rangle < 0$,*

(iii) ***null*** *if $\langle v, v\rangle = 0$.*

If v is either timelike or null, we will call v ***causal****.*

Note that any (T_pM, g_p) is isometric to Minkowski space $(\mathbb{R}^{1,n}, \eta)$. Therefore one can always find an orthonormal basis $\{e_0, e_1, \ldots, e_n\} \subset T_pM$ such that $g_p = \eta$, so in particular $g(e_0, e_0) = -1$. We define the **future lightcone** Λ^+ and **past lightcone** Λ^- at $p \in M$ as

$$\Lambda^+ \coloneqq \{v \in T_pM \mid \langle v, v\rangle = 0,\, v^0 > 0\}, \qquad \Lambda^- \coloneqq \{v \in T_pM \mid \langle v, v\rangle = 0,\, v^0 < 0\},$$

where $v^0 \in \mathbb{R}$ is the e_0 coordinate of v. A timelike vector $w \in T_pM$ lies either inside the future or past lightcone if $w^0 > 0$ or $w^0 < 0$ respectively (See also figure 1). The union $\Lambda \coloneqq \Lambda^+ \cup \Lambda^-$

[7]In general a manifold admits a Lorentzian metric if and only if there exists a non-vanishing vector field.

is called the **double lightcone** at p. Any causal vector is called either **future directed** if it lies in or on Λ^+ or **past directed** if it lies in or on Λ^-.

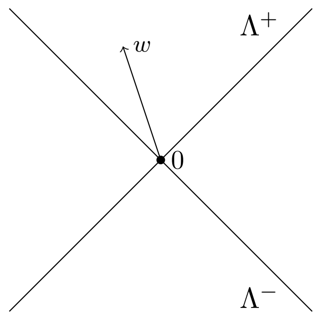


Figure 1: The future and past lightcones $\Lambda^\pm$ in T_pM. The timelike vector w lies inside the future lightcone.

Remark 2.2.2. The orthogonal group $O(1,n)$ on Minkwoski space is called the *Lorentz group*, i.e. $L \in O(1,n)$ if $L \in GL(1,n)$ with $\eta(Lx, Ly) = \eta(x,y)$ for all $x, y \in \mathbb{R}^{1,n}$. Its elements are called *Lorentz transformations*. The subgroup $SO(1,n) \subset O(1,n)$, which consists of the orientation preserving Lorentz transformations, is called the *proper* Lorentz group. Moreover, $SO(1,n)$ consists of the two connected components $SO^+(1,n) = \{L \in SO(1,n) \,|\, L_{00} > 0\}$ and $SO^-(1,n) = \{L \in SO(1,n) \,|\, L_{00} < 0\}$, where $L_{00} \in \mathbb{R}$ is the first entry of the matrix $(L_{ij})_{i,j=0,\ldots,n}$. One may show[8] that the Lorentz transformations in $SO^+(1,n)$ are *time preserving*, i.e. if $L \in SO^+(1,n)$ and $x \in \Lambda^+$, then $Lx \in \Lambda^+$. Respectively, $SO^-(1,n)$ contains all the proper time reversing Lorentz transformations. A lot more can be said about the algebraic properties of Lorentz transformations, but here we will simply refer to [Bä20, Chapter 1.1]. In particular, down below we will only need the fact that for a future directed $x \in \mathbb{R}^{1,n}$, it is possible to find a Lorentz transformation $L \in O(1,n)$ such that $x = (x^0, 0, \ldots, 0)$ (cf. [Bä20, Thm. 1.16]). Note also that all of this obviously carries over to (T_pM, g_p).

For a general Lorentzian manifold M, a continuous choice of the light cones $\Lambda^\pm$ at each $p \in M$ is what we will call a *time orientation*. In more precise terms we have

Definition 2.2.3. *Let (M,g) be a Lorentzian manifold. A smooth vector field*[9] $X \in \mathfrak{X}(M)$ *such that $X(p)$ is timelike for any $p \in M$ defines a time orientation on (M,g).*

Remark 2.2.4. A Lorentzian manifold (M,g) is called *time orientable* if there exists a time orientation. If X defines a time orientation then $-X$ also defines a time orientation, called *opposite* or *negative time orientation*, which exchanges the future and past directions. Note that this "symmetry" will often be used to reduce arguments to either the future or past direction.

For our purposes, we will mostly consider time orientable Lorentzian manifold from now on.

[8] See e.g. [Bä20, Cor. 1.21].
[9] A continuous section would be enough here. See e.g. [Bä20, Prop. 2.2].

In fact, this work will mainly revolve around the following object

Definition 2.2.5. *A spacetime is a connected Lorentzian manifold* (M, g) *that is time oriented.*

Similarly to definition 2.2.1, one may now introduce several notions of causality to the setting of curves as follows

Definition 2.2.6. *Let* $\gamma : I \to M$ *be a (piecewise smooth) curve in a spacetime* (M, g)*. We call* γ

(i) ***spacelike*** *if* $\dot{\gamma}$ *is always spacelike,*

(ii) ***timelike*** *if* $\dot{\gamma}$ *is always timelike,*

(iii) ***null*** *if* $\dot{\gamma}$ *is always null,*

(iv) ***causal*** *if* $\dot{\gamma}(t)$ *is either timelike or null for any* $t \in I$*.*

Furthermore, if γ *is causal, we call it* ***future directed*** *(resp.* ***past directed****) if* $\dot{\gamma}(t)$ *either lies in or on* Λ^+ *(resp.* Λ^-*) at any* $t \in I$*.*

The definitions for spacelike/timelike/null/causal and future/past directed vector fields follows the same logic.
From now on and for the remainder of this chapter, we will consider a spacetime (M, g). The following lemma turns out to be very useful in later calculations.

Lemma 2.2.7. *Let* $X \in \mathfrak{X}(M)$ *define a time orientation on* (M, g) *and let* $Y \in \mathfrak{X}(M)$ *be causal.* Y *is future directed if and only if* $g(X, Y) < 0$ *and it past directed if and only if* $g(X, Y) > 0$*.*

Proof. Since this is a point-wise claim, let $p \in M$. Using a Lorentz transformation, we can assume without loss of generality that $X(p) = (X^0, 0, \ldots, 0)$. Then, $g(X, Y) = \eta(X, Y) = -X^0 Y^0$. Because $X^0 > 0$ by assumption, the claim follows immediately. □

Following our previous definitions, one may now introduce the so called *causal relations* as follows

Definition 2.2.8. *Let* $p, q \in M$*. We write*

(i) $p \ll q$ *if there exists a future directed timelike curve from* p *to* q*,*

(ii) $p < q$ *if there exists a future directed causal curve from* p *to* q*,*

(iii) $p \leq q$ *if either* $p = q$ *or* $p < q$*.*

Note that these relations are all transitive. However, obviously $p \leq p$ and therefore $\leq$ is also reflexive. Moreover, recall that a partial order $\leq$ is a relation on a set M which is reflexive, transitive and antisymmetric, where antisymmetric means that $p \leq q$ and $q \leq p$ together imply $p = q$ for $p, q \in M$. Consequently, if (M, g) features no causal loops, i.e. no closed causal curves, the causal relation $\leq$ is indeed a partial order.

Definition 2.2.9. *Let $A \subset M$. The chronological future $I^+(A)$ and chronological past $I^-(A)$ of A are defined as*

$$I^+(A) = \{p \in M \,|\, \exists\, q \in A : \, q \ll p\}, \quad I^-(A) = \{p \in M \,|\, \exists\, q \in A : \, p \ll q\}.$$

Similarly, the causal future $J^+(A)$ and causal past $J^-(A)$ of A are given by

$$J^+(A) = \{p \in M \,|\, \exists\, q \in A : \, q \leq p\}, \quad J^-(A) = \{p \in M \,|\, \exists\, q \in A : \, q \leq p\}.$$

We will need to establish several properties of these subsets. In order to do so, we will need the so-called *Push-Up Principle*, which is of central importance.

Proposition 2.2.10 (Push-Up Principle)**.** *Let $p, q, r \in M$. If $p \ll q \leq r$ or $p \leq q \ll r$ then $p \ll r$.*

From a physics perspective, the Push-Up principle is very intuitive: Note that null curves travel with the speed of light. The first claim, for example, may then be understood as instead of traveling slowly from p to q and fast from q to r, we could at first slightly increase the velocity and then reduce it later, resulting in a curve which itself never reaches the speed of light. While this seems simple enough, proving this statement rigorously requires variational arguments, which we will develop later on in chapter 4.3. Therefore, for the full proof see either Proposition 4.3.5 or [O'N83, Prop. 10.46].
The transitivity of $\ll$ and $\leq$ together with the Push-Up principle may then be easily applied to show that

$$I^+(A) = I^+\left(I^+(A)\right) = I^+\left(J^+(A)\right) = J^+\left(I^+(A)\right) \subseteq J^+\left(J^+(A)\right) = J^+(A).$$

Definition 2.2.11. *Let $\Omega \subseteq M$ be open and $A \subseteq \Omega$. The relative future $I^+_\Sigma(A)$ and relative past $I^-_\Sigma(A)$ of A in Ω are defined as the sets*

$$I^+_\Omega(A) = \{p \in \Omega \,|\, \exists\, q \in A : \, q \ll p \text{ in } \Omega\}, \quad I^-_\Omega(A) = \{p \in \Omega \,|\, \exists\, q \in A : \, p \ll q \text{ in } \Omega\}.$$

Similarly, one defines

$$J^+_\Omega(A) = \{p \in \Omega \,|\, \exists\, q \in A : \, q \leq p \text{ in } \Omega\}, \quad J^-_\Omega(A) = \{p \in \Sigma \,|\, \exists\, q \in A : \, q \leq p \text{ in } \Omega\}.$$

We will also write $p \leq_\Omega q$ and $p \ll_\Omega q$ whenever $q \in J^+_\Omega(p)$ or $q \in I^+_\Omega(p)$.

While the global structure of the causal and chronological futures/pasts is not obvious a priori, their local behavior can be very well-described by considering radial geodesics. In particular

Proposition 2.2.12. *Let $p \in M$ and U be a normal neighbourhood of p with $\tilde{U} \subseteq T_pM$ such that $\exp_p : \tilde{U} \to U$ is a diffeomorphism. Then*

(i) $I^{\pm}_U(p) = \exp_p\left(I^{\pm}(0) \cap \tilde{U}\right)$,

(ii) $J^{\pm}_U(p) = \exp_p\left(J^{\pm}(0) \cap \tilde{U}\right)$.

The proof of the above result rests on properties of the exponential map and especially the fact that future directed timelike curves in the domain of $\exp_p$ are mapped into $I^+(p)$. For details see [Bä20, Lemma 2.11 & Cor. 2.13].
A basic property of the causal and relative future is that they are open sets. This is a consequence of the causal relation $\ll$ being an *open relation*. As in [Bä20, Prop. 2.14], one has

Proposition 2.2.13. *The relation $\ll$ is open, i.e. if $p \ll q$ with $p, q \in M$, then there exist neighbourhoods U of p and V of q such that $\overline{p} \ll \overline{q}$ for any $\overline{p} \in U$ and $\overline{q} \in V$.*

Proof. Let σ be a future directed timelike curve from p to q. Pick a convex neighbourhood C of q and let $q^- \in C$ be a point on σ before q. Now define $V \coloneqq I^+_C(q^-)$. By Proposition 2.2.12 one has $V = \exp_{q^-}(I^+(0) \cap \tilde{C})$. Since $\exp_{q^-} : \tilde{C} \to C$ is a diffeomorphism, U is open. Now choose U similarly, but such that $p^+ \in I^-(q^-)$. If $\overline{p} \in U$ and $\overline{q} \in V$, it then follows that $\overline{p} \ll p^+ \ll q^- \ll q$ and thus $\overline{p} \ll \overline{q}$. $\square$

Following [KS23, Cor. 3.1.15], we may immediately conclude that

Corollary 2.2.14. *Let $A \subseteq \Omega \subseteq M$. Then $I^{\pm}_{\Omega}(A)$ is open.*

Proof. Note that the previous Proposition implies that $I^{\pm}_{\Omega}(p)$ must be open. Since $I^{\pm}_{\Omega}(A) = \bigcup_{p \in A} I^{\pm}_{\Omega}(p)$, we are done. $\square$

In Minkwoski space, the topological relationship between causal and chronological future or past is quite obvious. Note for example that $I^{\pm}(A) = \operatorname{int}(J^{\pm}(A))$ and $J^{\pm}(A) = \overline{I^{\pm}(A)}$ for any $A \subseteq \mathbb{R}^{1,n}$. We will show that these properties can actually be generalized to arbitrary spacetimes. For the first claim, we follow [Bä20, Prop. 2.17] and for the second [KS23, Prop. 3.1.17].

Proposition 2.2.15. $I^{\pm}(A) = \operatorname{int}(J^{\pm}(A))$ *for any $A \subseteq M$.*

Proof. We only prove the $I^+(A)$ case since the $I^-(A)$ case is analogous. Because $I^+(A) \subset J^+(A)$ and $I^+(A)$ is open, one has $I^+(A) \subseteq \operatorname{int}(J^+(A))$. For the other direction let $p \in \operatorname{int}(J^+(A))$. Choose $q \in \operatorname{int}(J^+(A)) \cap I^-(p)$, which can be done because p lies in the interior.

It follows that there exists an $r \in A$ such that $r \leq q \ll p$ and by Push-Up $r \ll p$. Consequently $p \in I^+(A)$ and $\operatorname{int}(J^+(A)) \subseteq I^+(A)$. $\square$

Proposition 2.2.16. *$J^\pm(A) \subseteq \overline{I^+(A)}$ for any $A \subseteq M$ with equality if and only if $J^\pm(A)$ closed.*

Proof. Again, we only prove the $J^+(A)$ case. Let U be a normal neighbourhood of $p \in A$. By Proposition 2.2.12 one has $J^+_\Omega(A) = \exp_p\left(J^+(0) \cap \tilde{U}\right)$. Since $\exp_p$ is a diffeomorphism and because $J^+(0) \cap \tilde{U} = \overline{I^+(0) \cap \tilde{U}}$, one also has $\exp_p\left(J^+(0) \cap \tilde{U}\right) = \mathrm{cl}_U\left(\exp_p\left(I^+(0) \cap \tilde{U}\right)\right)$, where cl_U denotes the closure in U. Put together, $J^+_U(p) = \mathrm{cl}_U\left(I^+_U(p)\right)$.
Let $p \in \overline{I^+(p)}$ and pick a $q \in J^+(p)$. Using the above result for $q^- \in J^-_U(q)$, we find $q \in J^+_U(q^-) = \mathrm{cl}_U\left(I^+_U(q_-)\right)$. However, note that

$$I^+_U(q^-) \subseteq I^+\left(J^+(p)\right) = I^+(p)$$

(see Figure 2) and therefore $q \in \overline{I^+(p)}$. Since

$$J^+(A) = \bigcup_{p \in A} J^+(p) \subseteq \bigcup_{p \in A} \overline{I^+(p)} \subseteq \overline{\bigcup_{p \in A} I^+(A)} = \overline{I^+(A)},$$

the first part of the claim is done.

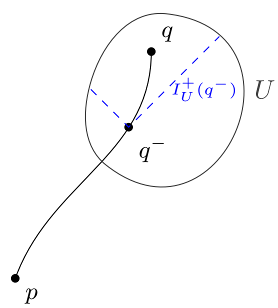


Figure 2: Situation in proof of Proposition 2.2.16.

For the final part, simply note that if $J^+(A) = \overline{I^+(A)}$ then $J^+(A)$ is obviously closed. On the other hand, if $J^+(A)$ is closed, then one has $\overline{I^+(A)} \subseteq J^+(A) \subseteq \overline{I^+(A)}$, where the first inclusion follows from $I^+(A) \subseteq J^+(A)$ and the second one from the result above. $\square$

Example 2.2.17. A small example of Proposition 2.2.16 is given by $\mathbb{R}^{1,1} \backslash \{(1,1)\}$, i.e. Minkwoski space with point $(1,1)$ removed. Here, $J^+(0)$ is not closed and in fact $J^+(0)$ is contained in $\overline{I^+(0)}$ but cannot be equal to it (see Figure 3).

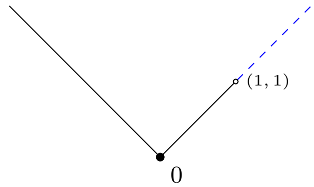


Figure 3: Minkowski space with $(1,1)$ removed. The closure of $I^+(0)$ will contain the dashed blue line that is not part of $J^+(0)$.

To summarize, one can recover certain properties of the chronological and causal sets that are analogous to Minkwoski space in a very general manner using local arguments. On a global scale however, the curvature of (M,g) could still affect the causal structure in such a way, that the analogy to Minkowski space is simply insufficient (see chapter 7 for example). To our advantage, as we have seen in Proposition 2.2.12, the local story is different. We will now see that, as one would expect, convex sets are very convenient to work with from a causality perspective. As in [O'N83, Lemma 14.2], one has

Proposition 2.2.18. *Let $C \subseteq M$ be a convex set. Then*

(i) $q \in J^+_C(p)$ *if and only if* $\exp_p^{-1}(q)$ *is future directed causal,*

(ii) $q \in I^+_C(p)$ *if and only if* $\exp_p^{-1}(q)$ *is future directed timelike,*

(iii) $\leq$ *is closed on* C*, i.e. if* $\{p_n\}_{n\in\mathbb{N}}$ *and* $\{q_n\}_{n\in\mathbb{N}}$ *are sequences in* C *with* $p_n \to p \in C$ *and* $q_n \to q \in C$ *such that* $p_n \leq q_n$ *for all* $n \in \mathbb{N}$*, then also* $p \leq q$*.*

Proof. (i) and (ii) immediately follow from Proposition 2.2.12, so we will only need to prove (iii). Let $p \neq q$ because otherwise, we are done. We can further assume $p_n \neq q_n$ for each $n \in \mathbb{N}$. Note now that the map $(p,q) \mapsto \exp_p^{-1}(q)$ is continuous. Thus, because $p_n \leq q_n$, one has

$$\langle \exp_{p_n}^{-1}(q_n), \exp_{p_n}^{-1}(q_n)\rangle \leq 0 \quad \Rightarrow \quad \langle \exp_p^{-1}(q), \exp_p^{-1}(q)\rangle \leq 0$$

and $\exp_p^{-1}(q)$ is causal. The remaining issue is whether we have $p \leq q$ or $q \leq p$. To address this, let $X \in \mathfrak{X}(M)$ be future directed timelike. By Lemma 2.2.7, one has $\langle X, \exp_{p_n}^{-1}(q_n)\rangle < 0$ and again, continuity implies that $\langle X, \exp_p^{-1}(q)\rangle \leq 0$. Assume now that $\langle X_p, \exp_p^{-1}(q)\rangle = 0$. This implies that $\exp_p^{-1}(q)$ is spacelike: As in the proof of Lemma 2.2.7, we may assume that $X_p = (X^0, 0, \ldots, 0)$ and let $v = (v^0, v^1, \ldots, v^n)$ with $v \coloneqq \exp_p^{-1}(q)$. Since $\langle X_p, v\rangle = -X^0 v^0 = 0$, we have $v^0 = 0$. But then $\langle v, v\rangle = \sum_{i=1}^n \left(v^i\right)^2 > 0$, i.e. v must be spacelike. However, this is a contradiction, and therefore $\langle X_p, v\rangle < 0$ and $p \leq q$. □

For the final part of this chapter, we pick up the discussion from the previous section. Back then we indicated the differences surrounding geodesic completeness on Riemannian and Lorentzian manifolds respectively. Having introduced the fundamental notions of Causality theory, we may now be more precise in terms of the latter. Since causality theory enables us to distinguish different classes of curves, it is possible to also distinguish types of extendibility and completeness.

Definition 2.2.19. *Let (M, g) be a spacetime and $\gamma : (a, b) \to M$ a future directed causal curve. Then γ is*

(i) future inextendible if it cannot be continuously extended to b,

(ii) past inextendible if it cannot be continuously to a.

Furthermore, γ is called inextendible if it cannot be extended either to the future or the past.

In terms of geodesics we still have completeness or incompleteness in these previous sense, i.e. a geodesic is incomplete if it does not exist on all of $\mathbb{R}$. However, one may now define the following

Definition 2.2.20. *Let (M, g) be a spacetime. Then (M, g) is*

(i) future/past null incomplete if there exists an incomplete future/past directed null geodesic,

(ii) future/past timelike incomplete if there exists an incomplete future/past directed timelike geodesic.

In the case where (M, g) is both, future and past null/timelike incomplete, one calls (M, g) null/timelike incomplete.

Note that the main result of this work, the Penrose Incompleteness Theorem, addresses (i) and gives sufficient conditions for a failure of null completeness of a spacetime (M, g). Compared to Riemannian geometry this might seem dissatisfying, but there is not much hope for expecting something stronger. Recall from before that the Hopf-Rinow Theorem connected geodesic completeness to completeness of the metric space (M, d_g) for Riemannian manifolds. We will now clarify why such a result does not hold generically for Lorentzian manifolds or spacetimes.

Definition 2.2.21. *Let $\gamma : [a, b] \to M$ be a curve. The length of γ is given by*

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

This is in complete analogy to Riemannian geometry. However, observe that in the Lorentzian case we actually have that any null curve γ satisfies $L(\gamma) = 0$. This marks the point where our analogy fails. If one would simply define the distance function as the infimum over the length functional, as in the Riemannian case, the resulting function is definitely not a metric (in the proper sense of the word). Nevertheless, one should ask if it possible to simply define a different notion of distance instead to salvage the situation. This turns out to be partially true. One can in fact define an analogue of d_g for Lorentzian manifolds as follows

Definition 2.2.22. *The time-separation function* $\tau : M \times M \to \mathbb{R}$ *is defined as*

$$\tau(p,q) = \begin{cases} \sup\{L(\gamma)\,|\,\gamma \textit{ future directed causal curve from } p \textit{ to } q\}, & \textit{if } p < q \\ 0, & \textit{if } p \not< q \end{cases}$$

Intuitively, one should think of τ as the (proper) time it takes to travel the slowest possible path from p to q. Observe that τ does not define a metric. In particular, it is not symmetric. Furthermore, it is also unbounded, i.e. there could be points $p < q$ such that $\tau(p,q) = \infty$. One may also show (see [O'N83, Lemma 14.16 & Lemma 14.17]) that τ is lower semicontinuous and satisfies a reverse triangle inequality. Nevertheless, τ still should be viewed an analogue to d_g because in normal neighbourhoods, timelike radial geodesics maximize the time-separation for Lorentzian manifolds in the same way as radial geodesics minimize the distance in the Riemannian case (cf. [O'N83, Prop. 5.34]). We will not explicitly need the time-separation function, but nevertheless emphasize its critical significance especially in settings of lower regularity (see e.g. [KS18]).

3 Semi-Riemannian Submanifolds

The Penrose Incompleteness Theorem assumes the existence of a so called *trapped surface*[10], which is a special kind of submanifold of codimension 2. In the later chapters, our goal is to study null geodesics that emanate from these trapped surfaces. It is therefore necessary that we examine the geometry of submanifolds of a Semi-Riemannian manifold (M,g) more closely. At this point, we want to deal with submanifolds where g "restricts" to a proper metric tensor. Submanifolds with this property are called *semi-Riemannian submanifolds.* This need for distinction is again a side effect of working with a Lorentzian manifold. In comparison to this, for Riemannian manifolds, any submanifold is automatically a Riemannian manifold as well. The machinery developed in this chapter turns out to be especially crucial in section 7. We furthermore emphasize that the treatment of semi-Riemannian submanifold is somewhat analogous to the Riemannian case, which can also be found in various other textbooks.

3.1 Smooth Submanifolds

Let (M,g) be a Semi-Riemannian manifold with $\dim(M) = n$. We expect the reader to be familiar with smooth submanifolds in general and otherwise refer to [Kun22b, Section 3.3]. For the rest of this section any submanifold is taken to be smooth except if explicitly stated otherwise. We will closely follow [O'N83] and [KS23].

Let $P \subseteq M$ be a submanifold of M with $\dim(P) = k$. We ask ourselves how g carries over to P. To this end, let $j : P \hookrightarrow M$ denote the inclusion map and recall that for $p \in P$, we can make the natural identification between T_pP and a subspace of T_pM via the isomorphism

[10] This will be properly introduced in chapter 8.1.

T_pj, i.e. $T_pP \subseteq T_pM$ (cf. [Kun22b, Remark 3.3.11]). Pulling back the metric, we obtain

$$\not{g}_p(v,w) \coloneqq j^*g(p)(v,w) = g(j(p))(T_pj(v), T_pj(w)) = g(p)(v,w) = \langle v,w\rangle_p$$

for $v,w \in T_pP$. So using the previously mentioned identification, $\not{g}$ is the restriction of g to P.

We now want to study submanifolds P for which $\not{g}$ is itself a semi-Riemannian metric. In this case, one has the decomposition[11] $T_pM = T_pP \oplus T_pP^\perp$ for each $p \in P$ (see [KS21, Lemma 1.1.9]) and can therefore define the linear projections $\tan : T_pM \to T_pP$ and $\mathrm{nor} : T_pM \to T_pP^\perp$. As a consequence, any $X \in T_pM$ may be written as $X = \tan(X) + \mathrm{nor}(X)$. We call a vector $X \in T_pM$ *tangential* (resp. *normal*) to P if $\mathrm{nor}(X) = 0$ (resp. if $\tan(X) = 0$).

Recall that by definition 2.1.3, the $\mathcal{C}^\infty(P)$-module $\mathfrak{X}(j)$ consists of the smooth maps $X : P \to TM$ for which $\pi \circ X = j$. We will call these kind of vector fields *M-vector fields over P* and also primarily use the notation $\overline{\mathfrak{X}}(P) \coloneqq \mathfrak{X}(j)$. Note that if $X \in \mathfrak{X}(M)$, then $X|_P \in \overline{\mathfrak{X}}(P)$ but generally $X \notin \mathfrak{X}(P)$. This reflects the fact that the vector fields in $\overline{\mathfrak{X}}(P)$ can "point out" of P and do not have to be necessarily tangent to P everywhere. Clearly $\mathfrak{X}(P) \subseteq \overline{\mathfrak{X}}(P)$.

Similarly to the case of Riemannian submanifolds, we have a smooth splitting of the tangent and normal bundle, i.e. $TM = TP \oplus TP^\perp$. Even better, there exists a local orthonormal frame that respects this splitting. For a proof we refer to [Hic63] and [KS23, Lemma 1.3.6].

Lemma 3.1.1 (Hicks)**.** *Let $P \subseteq M$ be a semi-Riemannian submanifold with $\dim(P) = k$ and $p \in P$. There exists an open neighbourhood $W \subseteq P$ of p and $E_1, \dots, E_n \in \overline{\mathfrak{X}}(W)$ such that $\{E_1, \dots, E_k\}$ is a local orthonormal frame for TP and $\{E_{k+1}, \dots, E_n\}$ for $TP^\perp$ respectively.*

Using the above lemma, we are now able to write (at least locally)

$$X = \underbrace{\sum_{i=1}^{k} \epsilon_i \langle X, E_i\rangle}_{\in \mathfrak{X}(P)} + \underbrace{\sum_{i=k+1}^{n} \epsilon_i \langle X, E_i\rangle E_i}_{\in \mathfrak{X}^\perp(P)}$$

for any $X \in \overline{\mathfrak{X}}(P)$, where $\epsilon_i = \langle E_i, E_i\rangle = \pm 1$. Therefore, the tangential and orthogonal projection operators can be generalized to the bundle level to obtain maps

$$\begin{aligned} \tan &: \overline{\mathfrak{X}}(P) \to \mathfrak{X}(P), \\ \mathrm{nor} &: \overline{\mathfrak{X}}(P) \to \mathfrak{X}^\perp(P). \end{aligned}$$

In conclusion, we have $\overline{\mathfrak{X}}(P) = \mathfrak{X}(P) \oplus \mathfrak{X}^\perp(P)$.

Above, we have remarked the fact that one may easily consider the restriction of vector fields

[11] Here, $T_pP^\perp$ denotes the orthogonal complement, i.e. $T_pP^\perp = \{v \in T_pM \mid g(v,w) = 0 \quad \forall\, w \in T_pP\}$.

on M to P. As a technical issue, it is also important to ask if we can go the other way around. More precisely, we want to know given any $X \in \mathfrak{X}(P)$, if one can construct a vector field $\tilde{X} \in \mathfrak{X}(M)$ such that the restriction to P of $\tilde{X}$ is equal to X. Following [KS23, Lemma 1.3.8], we will now show that this is feasible at least locally.

Lemma 3.1.2. *Let P, M be smooth manifolds with $\dim(P) = k$ and $\dim(M) = n$ and assume there exists an immersion $j : P \to M$. Then, for all $p \in P$, we have*

(i) $\forall f \in \mathcal{C}^\infty(P)$ there exists $\tilde{f} \in \mathcal{C}^\infty(M)$ such that $f = \tilde{f} \circ j$ near p.

(ii) $\forall X \in \mathfrak{X}(j)$ there exists $\tilde{X} \in \mathfrak{X}(M)$ such that $X = \tilde{X} \circ j$ near p.

Proof. (i) There exist charts (U, ϕ) at $p \in P$ and (V, ψ) at $j(p) \in M$ such that $\mathrm{pr} \circ \psi \circ j = \phi$, where $\mathrm{pr} : \mathbb{R}^k \to \mathbb{R}^n$ is the projection operator (see [Kun22a, Thm. 3.3.3]). Using a partition of unity, pick $\tilde{f} \in \mathcal{C}^\infty(M)$ such that $\tilde{f} = f \circ \phi^{-1} \circ \mathrm{pr} \circ \psi$ near p. Consequently

$$\tilde{f} \circ j = f \circ \phi^{-1} \circ \mathrm{pr} \circ \psi \circ j = f \circ \phi^{-1} \circ \phi = f$$

near p.
(ii) We can choose a neighbourhood $V \subseteq P$ around p such that we can write $X(q) = X^i(q)\, \partial_i\big|_{j(q)}$ with respect to some chart on $j(V)$. By (i), we can extend X^i to $\tilde{X}^i$ for each $i = 1, \dots, n$ such that $X^i = \tilde{X}^i \circ j$. Setting $\tilde{X} \coloneqq \tilde{X}^i \partial_i$ then yields the claim. $\square$

3.2 The Induced Connection

Any semi-Riemannian submanifold $P \subseteq M$ is of course a semi-Riemannian manifold in its own right. Therefore, it also possesses its own Levi-Civita connection $\not\nabla : \mathfrak{X}(P) \times \mathfrak{X}(P) \to \mathfrak{X}(P)$. It is now natural to ask how $\not\nabla$ relates to ∇. To answer this question sufficiently, we need to take a look at the mapping $\nabla : \mathfrak{X}(P) \times \overline{\mathfrak{X}}(P) \to \overline{\mathfrak{X}}(P)$ that is induced by the Levi-Civita connection on M. Suitably, this map is called *induced connection* on P and it will prove to be even more useful than simply functioning as a bridge between the two Levi-Civita connections of M and P respectively. Note that [O'N83] and [KS23] will remain to be our main sources for the rest of this chapter.
Before studying any properties of the induced connection above, we emphasize that there is an immediate issue at hand: Given $V \in \mathfrak{X}(P)$ and $X \in \overline{\mathfrak{X}}(P)$, the expression $\nabla_V X$ is not defined because $V, W \notin \mathfrak{X}(M)$. We remedy this by applying Lemma 3.1.2 to extend V, X to vector fields $\tilde{V}, \tilde{X} \in \mathfrak{X}(M)$ and then define

$$\nabla_V X \coloneqq \nabla_{\tilde{V}} \tilde{X}\Big|_P \tag{3.1}$$

However, one still needs to show that this expression is actually well-defined. We follow [O'N83, Lemma 4.1] and [KS23, Lemma 1.3.9].

Lemma 3.2.1. *The induced connection (3.1) is a well-defined smooth vector field on P.*

Proof. Clearly $\nabla_V X \in \overline{\mathfrak{X}}(P)$ as the restriction of a smooth vector field on M. Therefore, we only need to show that the definition is independent of our choice of extensions $\tilde{V}$ and $\tilde{X}$. Let $p \in P$ and choose adapted coordinates[12] $(U, \varphi = (x^1, \dots, x^n))$ for M around p. Near p, we can therefore write $V = V^i \partial_i$ and $X = X^j \partial_j$ with $V^i \in \mathcal{C}^\infty(P \cap U)$ for $i = 1, \dots, k$ and $X^j \in \mathcal{C}^\infty(P \cap U)$ for $j = 1, \dots, n$. Similarly one has $\tilde{V} = \tilde{V}^i \partial_i$ with $\tilde{V}^i \in \mathcal{C}^\infty(U)$ such that $\tilde{V}^i\big|_{U \cap P} = V^i$ and $\tilde{X} = \tilde{X}^j \partial_j$ with $\tilde{X}^i \in \mathcal{C}^\infty(U)$ such that $\tilde{X}^j\big|_{U \cap P} = X^j$.
For any $q \in U \cap P$, it then follows by the fact that we used adapted coordinates that

$$\tilde{V}(q)(\tilde{X}^j) = \tilde{V}^i(q)\, \partial_i|_q (\tilde{X}^j) = V^i(q)\, \partial_i|_q (X^j) = V(q)(X^j).$$

Additionally, since a connection is a point-wise map in the lower slot, we have

$$\nabla_{\tilde{V}} \partial_j(q) = \nabla_{\tilde{V}(q)} \partial_j = \nabla_{V(q)} \partial_j.$$

In total, we obtain

$$\nabla_{\tilde{V}} \tilde{X}(q) = \tilde{V}(q)(\tilde{X}^j)\, \partial_j|_q + \tilde{X}^j(q) \nabla_{\tilde{V}} \partial_j(q) = V(q)(X^j)\, \partial_j|_q + X^j(q) \nabla_{V(q)} \partial_j$$

and therefore $\nabla_{\tilde{V}} \tilde{X}$ only depends on V and X as desired. □

The following result displays the fact that the induced connection inherits various properties of the original Levi-Civita connection on M. We follow [O'N83, Cor. 4.2] and [KS23, Prop. 1.3.10].

Proposition 3.2.2. *Let $V, W \in \mathfrak{X}(P)$, $X, Y \in \overline{\mathfrak{X}}(P)$ and $f \in \mathcal{C}^\infty(M)$. The induced connection (3.1) has the following properties*

(i) $\nabla_V X$ *is $\mathcal{C}^\infty(M)$-linear in V.*

(ii) $\nabla_V X$ *is $\mathbb{R}$-linear in X.*

(iii) $\nabla_V (fX) = V(f)X + f \nabla_V X$.

(iv) $[V, W] = \nabla_V W - \nabla_W V$.

(v) $V\langle X, Y\rangle = \langle \nabla_V X, Y\rangle + \langle X, \nabla_V Y\rangle$.

Proof. Let $p \in P$ and extend the given vector fields and functions to vector fields and functions $\tilde{V}, \tilde{W}, \tilde{X}, \tilde{Y}, \tilde{f}$ on M. Note that the following holds near p: (a) By definition $\nabla_{\tilde{V}} \tilde{X} = \nabla_V X$. (b) $\tilde{V}(\tilde{f}) = V(f)$ from the proof of the previous lemma. (c) $\langle \tilde{X}, \tilde{Y}\rangle = \langle X, Y\rangle$. (d) $[\tilde{V}, \tilde{W}] = [V, W]$, which follows again by using adapted coordinates as in the previous lemma. The five

[12] One has $\varphi(p) = 0$ and $\varphi(q) = (x^1(q), \dots, x^k(q), 0, \dots, 0)$ for all $q \in U \cap P$. See [Kun22a, Thm. 3.3.12] for details.

properties are satisfied for the Levi-Civita connection on M and consequently the restriction will then give the results above. $\square$

As the induced connection does not necessarily have to have values tangent to P, we should study its normal and orthogonal projections in more detail. In particular, we first study its projections restricted to the case that ∇ takes only vector fields that are tangent to P. To no surprise, it turns out that the tangential projection is actually the Levi-Civita connection $\not\nabla$ on P. Just as in [O'N83, Lemma 4.3], one has

Proposition 3.2.3. *Let $V, W \in \mathfrak{X}(P)$. Then*

$$\tan(\nabla_V W) = \not\nabla_V W.$$

Proof. Let $X \in \mathfrak{X}(P)$ and locally around $p \in P$ extend all vector fields to vector fields $\tilde{X}, \tilde{V}, \tilde{W}$ on M. Recall the Koszul formula for ∇ (see e.g. [KS21, (1.3.10)])

$$\begin{aligned} 2\langle \nabla_{\tilde{V}} \tilde{W}, \tilde{X} \rangle &= \tilde{V}\langle \tilde{W}, \tilde{X} \rangle + \tilde{W}\langle \tilde{X}, \tilde{V} \rangle - \tilde{X}\langle \tilde{V}, \tilde{W} \rangle \\ &\quad - \langle \tilde{V}, [\tilde{W}, \tilde{X}] \rangle + \langle \tilde{W}, [\tilde{X}, \tilde{V}] \rangle + \langle \tilde{X}, [\tilde{V}, \tilde{W}] \rangle \\ &=: F(\tilde{V}, \tilde{W}, \tilde{X}). \end{aligned}$$

Using the properties $(a) - (d)$ of the previous proof, we find that near p

$$\begin{aligned} &\langle \nabla_{\tilde{V}} \tilde{W}, \tilde{X} \rangle \overset{(a)}{=} \langle \nabla_V W, X \rangle, \\ &\tilde{V}\langle \tilde{W}, \tilde{X} \rangle \overset{(b)}{=} V\left(\langle \tilde{W}, \tilde{X} \rangle\right) \overset{(c)}{=} V\langle W, X \rangle, \\ &\langle \tilde{V}, [\tilde{X}, \tilde{W}] \rangle \overset{(c)}{=} \langle V, [\tilde{X}, \tilde{W}] \rangle \overset{(d)}{=} \langle V, [X, W] \rangle. \end{aligned}$$

Consequently, we obtain

$$2\langle \nabla_{\tilde{V}} \tilde{W}, \tilde{X} \rangle = \langle \nabla_V W, X \rangle = F(V, W, X) = 2\langle \not\nabla_V W, X \rangle$$

locally around p, where the last equation is due to the Koszul formula for $\not\nabla$. Therefore we conclude $\langle \nabla_V W, X \rangle = \langle \not\nabla_V W, X \rangle$. Finally note that since X is tangential to P, we have

$$\langle \nabla_V W, X \rangle = \langle \tan(\nabla_V W) + \mathrm{nor}(\nabla_V W), X \rangle = \langle \tan(\nabla_V W), X \rangle.$$

Since $X \in \mathfrak{X}(P)$ was arbitrary, the claim follows. $\square$

Let us now shed some light on the normal projection of $\nabla_V W$.

Definition 3.2.4. *The second fundamental form (or shape tensor) of P in M is the map*

$\mathrm{I\!I} : \mathfrak{X}(P) \times \mathfrak{X}(P) \to \mathfrak{X}(P)^{\perp}$ *given by*

$$\mathrm{I\!I}(V, W) \coloneqq \mathrm{nor}(\nabla_V W)$$

Note that the second fundamental form is not a proper tensor field on P as it takes values in $\mathfrak{X}^{\perp}(P)$. However, it still has some very nice properties. We proceed as in [O'N83, Lemma 4.4].

Proposition 3.2.5. $\mathrm{I\!I}$ *is* $\mathcal{C}^{\infty}(P)$*-bilinear and symmetric.*

Proof. First we show symmetry. Let $V, W \in \mathfrak{X}(P)$. By linearity of the orthogonal projection, we have

$$\mathrm{I\!I}(V, W) - \mathrm{I\!I}(W, V) = \mathrm{nor}(\nabla_V W - \nabla_W V) = \mathrm{nor}([V, W]) = 0,$$

where we used Proposition 3.2.2 (iv) and the fact that $[V, W]$ is tangent to P. For the $\mathcal{C}^{\infty}(P)$-bilinearity simply note the fact that Proposition 3.2.2 implies that $\mathrm{I\!I}$ is $\mathcal{C}^{\infty}(P)$-linear in its first slot and by the symmetry above therefore also in its second. □

Note that the $\mathcal{C}^{\infty}(P)$-bilinearity still is enough to make it a point-wise object (see [O'N83, Prop. 2.2], i.e. $\mathrm{I\!I}(V, W)(p)$ depends only on $V(p), W(p)$.
Following the two preceding Propositions, one may conclude that

$$\nabla_V W = \not\nabla_V W + \mathrm{I\!I}(V, W) \tag{3.2}$$

for $V, W \in \mathfrak{X}(P)$. Observe that $\mathrm{I\!I}$ now can be understood as measuring the difference of the Levi-Civita connections ∇ and $\not\nabla$.
Our next goal is to adapt this discussion to the setting of vector fields on curves. For a smooth curve $\gamma : I \to P$, we can now consider the two separate induced covariant derivatives $\frac{\nabla}{dt}$ and $\frac{\not\nabla}{dt}$, induced by ∇ and $\not\nabla$ respectively. Their relation is summarized in the following Proposition (see [O'N83, Prop. 4.8]).

Proposition 3.2.6. *Let* $\gamma : I \to P$ *be a smooth curve on* P *and let* $Y \in \mathfrak{X}(\gamma)$ *be tangent to* P*, i.e.* $\mathrm{nor}(Y(t)) = 0$ *for all* $t \in I$*. Then*

$$\dot{Y} = Y' + \mathrm{I\!I}(\gamma', Y),$$

where $\dot{Y} = \frac{\nabla Y}{dt}$ *and* $Y' = \frac{\not\nabla Y}{dt}$.

Proof. One may prove this locally. Without loss of generality, assume that γ is contained in a single adapted coordinate chart for M. Using the local coordinate expression for $\dot{Y}$, Equation

(2.1), one finds

$$\begin{aligned}\dot{Y} &= \sum_{i=1}^{k} \frac{dY^i}{dt}(\partial_i \circ \gamma) + \sum_{i=1}^{k} Y^i \nabla_{\gamma'} \partial_i \\ &= \sum_{i=1}^{k} \frac{dY^i}{dt}(\partial_i \circ \gamma) + \sum_{i=1}^{k} Y^i \nabla\!\!\!/_{\gamma'} \partial_i + \sum_{i=1}^{k} Y^i \mathrm{I\!I}(\gamma', \partial_i) \\ &= Y' + \mathrm{I\!I}(\gamma', Y),\end{aligned}$$

where we have used (3.2), the local expression of Y' and also the $\mathcal{C}^\infty(P)$-linearity of $\mathrm{I\!I}$. □

Remark 3.2.7. The previous Proposition gives us an idea how $\mathrm{I\!I}$ encodes the geometric information of P. Suppose $p \in P$, $v \in T_pP$ and let $c : I \to P$ be a geodesic in P with initial conditions (p, v) at $t = 0$. Since $c' = \dot{c} \in \mathfrak{X}(P)$ one has $\ddot{c} = c'' + \mathrm{I\!I}(c', c')$. Therefore $\ddot{c}(0) = \mathrm{I\!I}(v, v)$ and because c is a "straight" curve with respect to P, we can see that $\mathrm{I\!I}$ tells c how much it is forced to curve with respect to the ambient manifold M to just stay in P.
The observation that $\ddot{c} = \mathrm{I\!I}(c', c')$ holds for P-geodesics naturally implies that $\ddot{c} \perp P$. However, note that the opposite holds as well: Suppose $\ddot{c}$ is normal to P. We have $c'' = \ddot{c} - \mathrm{I\!I}(c', c')$. Because the left hand side is tangent to P while the right hand side is normal to P one has $c'' = 0$. In conclusion, we have shown that a curve c on P is a P-geodesic if and only if $\ddot{c} \perp P$.
In the case that $\mathrm{I\!I} = 0$, we call P *totally geodesic*. It is easy to see that this implies that P-geodesics are then also M-geodesics and vice versa. Moreover, one can show that this is in fact an equivalence (see [O'N83, Prop. 4.13]).

The induced connection was a map $\nabla : \mathfrak{X}(P) \times \overline{\mathfrak{X}}(P) \to \overline{\mathfrak{X}}(P)$, but up until now we have put our sole attention to the case where its second argument is tangent to P. Therefore, we now want to study the orthogonal case accordingly. This motivates the following two definitions.

Definition 3.2.8. *The normal connection is the map $\nabla^\perp : \mathfrak{X}(P) \times \mathfrak{X}(P)^\perp \to \mathfrak{X}(P)^\perp$ defined by*

$$\nabla^\perp_V Z \coloneqq \mathrm{nor}(\nabla_V Z).$$

Definition 3.2.9. *The map $\tilde{\mathrm{I\!I}} : \mathfrak{X}(P) \times \mathfrak{X}(P)^\perp \to \mathfrak{X}(P)$ is given by*

$$\tilde{\mathrm{I\!I}}(V, Z) \coloneqq \tan(\nabla_V Z).$$

Thus, if $V \in \mathfrak{X}(P)$ and $Z \in \mathfrak{X}(P)^\perp$, we have

$$\nabla_V Z = \tilde{\mathrm{I\!I}}(V, Z) + \nabla^\perp_V Z. \tag{3.3}$$

We can therefore conclude that, similarly to $\mathrm{I\!I}$, the map $\tilde{\mathrm{I\!I}}$ measures the difference between ∇ and $\nabla^\perp$. Furthermore, it is also a tensorial object as it is $\mathcal{C}^\infty(P)$-bilinear: Linearity in the first argument and $\mathbb{R}$-linearity in the second follows from 3.2.2. Now let $V \in \mathfrak{X}(P), Z \in \mathfrak{X}(P)^{\perp}, f \in \mathcal{C}^{\infty}(P)$ and note that

$$\begin{aligned}\tilde{\mathrm{I\!I}}(V, fZ) &= \tan(\nabla_V(fZ)) = \tan(V(f)Z + f\nabla_V Z) = \tan(V(f)Z) + \tan(f\nabla_V Z)\\ &= f\tan(\nabla_V Z) = f\tilde{\mathrm{I\!I}}(V, Z),\end{aligned}$$

where we have used the fact that $V(f)Z \perp P$.
The following Proposition explains the similarity in notation to $\mathrm{I\!I}$. Namely, it demonstrates the fact that both, the second fundamental form and $\tilde{\mathrm{I\!I}}$, actually encode the same geometric information.

Proposition 3.2.10. *Let $V, W \in \mathfrak{X}(P)$ and $Z \in \mathfrak{X}(P)^{\perp}$. Then*

$$\langle \tilde{\mathrm{I\!I}}(V, Z), W \rangle = -\langle \mathrm{I\!I}(V, W), Z \rangle.$$

Proof. Using Proposition 3.2.2 and the fact that W is tangent to P we have

$$\begin{aligned}\langle \tilde{\mathrm{I\!I}}(V, Z), W \rangle &= \langle \tan(\nabla_V Z), W \rangle = \langle \nabla_V Z, W \rangle\\ &= V\langle Z, W \rangle - \langle Z, \nabla_V W \rangle\\ &= -\langle Z, \mathrm{nor}(\nabla_V W) \rangle\\ &= -\langle Z, \mathrm{I\!I}(V, W) \rangle.\end{aligned}$$

□

Let us now revisit the setting of vector fields along a smooth curve $\gamma : I \to P$ on P. The normal connection induces a *normal covariant derivative*, which one can then use to define the *normal transport* of a normal vector field along γ.

Definition 3.2.11. *Let $\gamma : I \to P$ be a smooth curve on P and $Y \in \mathfrak{X}(\gamma)$ with* $\tan(Y(t)) = 0$ *for all $t \in I$. The normal covariant derivative is defined as*

$$Y' := \frac{\nabla^{\perp} Y}{dt} := \mathrm{nor}\left(\frac{\nabla Y}{dt}\right).$$

By $\mathcal{C}^{\infty}$-linearity of the orthogonal projection, the usual properties of the induced covariant derivative (prop. 2.1.5) are inherited. Analogously to Proposition 3.2.6, one has

Proposition 3.2.12. *Let $\gamma : I \to P$ be a smooth curve on P and let $Y \in \mathfrak{X}(\gamma)$ which is normal to γ, i.e.* $\tan(Y(t)) = 0$ *for all $t \in I$. Then*

$$\dot{Y} = \tilde{\mathrm{I\!I}}(\gamma', Y) + Y'.$$

Proof. We use adapted coordinates just as in the proof of Proposition 3.2.6. Then, by equation (3.3)

$$\begin{aligned}\dot{Y} &= \sum_{i=k+1}^{n} \frac{dY^i}{dt}(\partial_i \circ \gamma) + \sum_{i=k+1}^{n} Y^i \, \nabla_{\gamma'} \partial_i \\ &= \sum_{i=k+1}^{n} \frac{dY^i}{dt}(\partial_i \circ \gamma) + \sum_{i=k+1}^{n} Y^i \, \nabla^{\perp}_{\gamma'} \partial_i + \sum_{i=k+1}^{n} Y^i \tilde{\mathrm{I\!I}}(\gamma', Y) \\ &= Y' + \tilde{\mathrm{I\!I}}(\gamma', Y).\end{aligned}$$

□

We will call a normal vector field $Y \in \mathfrak{X}(\gamma)$ *normal parallel* if $Y' = 0$. This is of course reminiscent of the usual definition of parallel vector fields on M or P. Much in the same way, we want to establish the notion of normal transport along a curve.

Lemma 3.2.13. *Let $\gamma : I \to P$ be a smooth curve on P and $y \in T_{\gamma(t_0)}P^{\perp}$ with $t_0 \in I$. Then there exists a unique normal parallel vector field $Y \in \mathfrak{X}(\gamma)$ such that $Y(t_0) = y$.*

Proof. Take adapted coordinates around $\gamma(t_0)$. By Equation (2.1), one has

$$Y' = \sum_{i=k+1}^{n} \left(\frac{dY^i}{dt} + \Gamma^i_{jk} \frac{d(x^j \circ \gamma)}{dt} Y^k \right) (\partial_i \circ \gamma) = 0.$$

This is a first order linear ODE and therefore[13] has a unique global (on the chart) solution with $Y(t_0) = y$. By uniqueness, we can now cover all of γ by these charts and then solve the above equation with appropriate initial conditions on each chart to obtain a unique normal parallel vector field Y with $Y(t_0) = y$. □

Analogous to the tangential case, normal parallel translation is a linear isometry. We follow the argument in [O'N83, Lemma 3.20].

Proposition 3.2.14. *Let $\gamma : I \to P$ be a smooth curve on P. Further, let $t_1, t_2 \in I$ and set $p = \gamma(t_1), q = \gamma(t_2)$. The normal parallel translation along γ from p to q defined by*

$$\begin{aligned}\mathcal{P}^{\gamma}_{t_1,t_2} : T_pP^{\perp} &\to T_qP^{\perp} \\ y &\mapsto Y(t_2),\end{aligned}$$

where Y is the unique normal parallel vector field to y, is a linear isometry.

Proof. Let $v, w \in T_pP$ and let V, W be their unique normal parallel transports. By linearity of the induced normal connection, $V + W$ is also normal parallel with $(V + W)(t_1) = v + w$.

[13] See e.g. [Arn92, Chapter 4].

Thus

$$\mathcal{P}^{\gamma}_{t_1,t_2}(v+w) = (V+W)(t_2) = V(t_2) + V(t_2) = \mathcal{P}^{\gamma}_{t_1,t_2}(v) + \mathcal{P}^{\gamma}_{t_1,t_2}(w).$$

Analogously, one finds $\mathcal{P}^{\gamma}_{t_1,t_2}(\lambda v) = \lambda \mathcal{P}^{\gamma}_{t_1,t_2}(v)$, so $\mathcal{P}^{\gamma}_{t_1,t_2}$ is indeed linear.
Now assume that $\mathcal{P}^{\gamma}_{t_1,t_2}(v) = 0$. By uniqueness of normal parallel transport, one then must have $v = 0$. Since $T_pP^{\perp}$ and $T_qP^{\perp}$ have the same dimension and $\mathcal{P}^{\gamma}_{t_1,t_2}$ is injective, it must be a linear isomorphism.
Finally, let V, W as before. Since the properties of the induced covariant derivative carried over to the induced normal covariant derivative, one still has

$$\frac{d}{dt}\langle V, W\rangle = \langle V', W\rangle + \langle V, W'\rangle = 0.$$

Thus $\langle V, W\rangle$ is constant and it follows that

$$\langle \mathcal{P}^{\gamma}_{t_1,t_2}(v), \mathcal{P}^{\gamma}_{t_1,t_2}(w)\rangle = \langle V(t_2), W(t_2)\rangle = \langle V(t_1), W(t_1)\rangle = \langle v, w\rangle.$$

□

4 Causal Curves in Spacetimes

In this Chapter we establish some results in the variational calculus of causal curves and prove the limit curve lemma, an essential and versatile tool in Lorentzian geometry, which can be applied to a very large variety of problems. As the name suggests, the lemma is concerned with sequences of causal curves on a spacetime (M, g) and provides a result for the existence of a continuous limit curve with respect to some topology. This result distinguishes itself from similar formulations in Riemannian geometry, in that it respects the additional causal structure. In particular, we will find that the limit of a sequence of causal curves will remain causal. As expected for a proof of such a statement, one needs an Ascoli Theorem, which is the subject of the first section, where we will also fix a topology to properly define what one means by limit curve. The second section is then dedicated to an introduction of continuous causal curves and the proof of the limit curve lemma. After that, in the final section, we will introduce the variational calculus of causal curves and then define a particular class of vector fields, called the Jacobi fields, which turn out to be crucial for our later analysis.

4.1 Ascoli's Theorem

An Ascoli Theorem needs some topological preparations, which we will develop here. Throughout this section, we follow [Mun14] as the primary reference.
If one wants to talk about convergence of a sequence of functions, it is necessary to define a concept of "nearness" on some function space. Note that this can then be achieved by defining a topology on Y^X, the set of functions from a space X to a space Y. Evidently, there is a

freedom of choice, as we don't want to put just any topology on our space. Recall that there already are many important and established notions of convergence in function spaces, i.e. point-wise convergence or uniform convergence. For our purposes uniform convergence is a bit too much to ask for, but as it turns out, we can work with a weaker, local version.

Definition 4.1.1. *Let X be a topological space and (Y, d) a metric space. Furthermore, let $C \subseteq X$ be compact and $\epsilon > 0$. For $f \in Y^X$, we define the ball $B_C(f, \epsilon) \subseteq Y^X$ by*

$$g \in B_C(f, \epsilon) \quad \Leftrightarrow \quad \sup_{x \in C} \{d(f(x), g(x))\} < \epsilon.$$

Proposition 4.1.2. *$\mathcal{B} = \{B_C(f, \epsilon) \mid C \subseteq X$ compact, $\epsilon > 0$, $f \in Y^X\}$ is a basis for a topology on Y^X.*

Proof. The balls clearly cover Y^X. We need to prove that any intersection contains another ball. For that, consider $B_{C_1}(f_1, \epsilon_1)$ and $B_{C_2}(f_2, \epsilon_2)$ with $B_{C_1}(f_1, \epsilon_1) \cap B_{C_2}(f_2, \epsilon_2) \neq \varnothing$.
It follows that there exists $g \in Y^X$ such that $\sup_{x \in C_i} \{d(f_i(x), g(x))\} < \epsilon_i$ for $i = 1, 2$. Now, let $\delta = \min\big(\epsilon_1 - \sup_{x \in C_1} \{d(f_1(x), g(x))\}, \epsilon_2 - \sup_{x \in C_2} \{d(f_2(x), g(x))\}\big)$. Since $C_1 \cup C_2$ is compact, we have $B_{C_1 \cup C_2}(g, \delta) \in \mathcal{B}$. We show that this ball is contained in the intersection of the other two. Let $f \in B_{C_1 \cup C_2}(g, \delta)$. We have, for $i = 1, 2$,

$$\sup_{x \in C_1 \cup C_2} \{d(f(x), g(x))\} < \delta \leq \epsilon_i - \sup_{x \in C_i} \{d(f_i(x), g(x))\}.$$

Using this and the triangle inequality we find

$$\begin{aligned}
\sup_{x \in C_i} \{d(f(x), f_i(x))\} &\leq \sup_{x \in C_i} \{d(f(x), g(x))\} + \sup_{x \in C_i} \{d(g(x), f_i(x))\} \\
&\leq \sup_{x \in C_1 \cup C_2} \{d(f(x), g(x))\} + \sup_{x \in C_i} \{d(g(x), f_i(x))\} \\
&< \epsilon_i.
\end{aligned}$$

Consequently, for $i = 1, 2$, $f \in B_{C_i}(f_i, \epsilon_i)$ and so $B_{C_1 \cup C_2}(g, \delta) \subseteq B_{C_1}(f_1, \epsilon_1) \cap B_{C_2}(f_2, \epsilon_2)$. □

The topology, induced by $\mathcal{B}$, is called **topology of compact convergence**. Note that convergence in this topology is equivalent to uniform convergence on compact subsets.

Proposition 4.1.3. *A sequence $\{f_n\}_{n \in \mathbb{N}} \subseteq Y^X$ converges with respect to the topology of compact convergence if and only if the sequence converges uniformly on compact subsets.*

Proof. Denote by f the limit of $\{f_n\}_{n \in \mathbb{N}}$ with respect to the topology of compact convergence and let $C \subseteq X$ be compact. Consider $B_C(f, \epsilon)$ for some $\epsilon > 0$. Since $\{f_n\}_{n \in \mathbb{R}}$ converges to

f, there exists $N \in \mathbb{N}$ such that $\sup_{x \in C} \{d(f(x), f_n(x))\} < \epsilon$ for all $n \geq N$. But this is equivalent to uniform convergence to f on C, so we are done. $\square$

Remark 4.1.4. At this point, one could ask if the topology of compact convergence depends on the choice of metric d. As it turns out, this is not the case on the set of continuous functions $\mathcal{C}^0(X,Y)$ (see Theorem 4.1.6 below and [Mun14, Thm. 46.8, Cor. 46.9]), which is the class of functions we are primarily concerned with.

One can in fact generalize the topology of compact convergence to the case where Y is not a metric space. For this, we define the **compact-open** topology on $\mathcal{C}^0(X,Y)$.

Definition 4.1.5. *Let X and Y be topological spaces. Let $K \subseteq X$ be compact and $U \subseteq Y$ be open and define*

$$S(K,U) = \{f \in \mathcal{C}^0(X,Y) \mid f(K) \subseteq U\}$$

The sets $S(K,U)$ then form a subbasis[14] *that generates the compact-open topology on $\mathcal{C}^0(X,Y)$.*

Returning to the metric case, one can then indeed show that the two topologies are equivalent. For a proof we refer to [Mun14, Thm. 46.8].

Theorem 4.1.6. *Let X be a topological space and (Y,d) a metric space. On $\mathcal{C}^0(X,Y)$, the topology of compact convergence and the compact-open topology are equivalent.*

Remark 4.1.7. The significance of the above result becomes more apparent in the next section, where we work with a Lorentzian manifold M which is of course metrizable. But most often, we are not at all interested in choosing a specific metric for M. Hence it can be very useful to work with the compact-open topology instead.

The next Proposition is important for later and showcases that the topology of compact convergence comes with some benefits.

Proposition 4.1.8. *Let X be a topological space and (Y,d) be a metric space. Then $C^0(X,Y)$, together with topology of compact convergence, is Haudorff.*

Proof. Let $f, g \in \mathcal{C}^0(X,Y)$ with $f \neq g$. Let $C \subseteq X$ be compact and define $M \coloneqq \sup_{x \in C} \{d(f(x), g(x))\}$. Consider $B_C(f, M/2)$ and $B_C(g, M/2)$, neighbourhoods of f and g respectively. Suppose $h \in B_C(f, M/2) \cap B_C(g, M/2)$. Then, using the triangle inequality

$$\begin{aligned}
\sup_{x \in C} \{d(f(x), g(x))\} &\leq \sup_{x \in C} \{d(f(x), h(x))\} + \sup_{x \in C} \{d(h(x), h(x))\} \\
&< M/2 + M/2 \\
&= \sup_{x \in C} \{d(f(x), g(x))\},
\end{aligned}$$

[14]Recall that a topology generated by a subbasis $\mathcal{S}$ is given by the collection of all unions of finite intersections of elements in $\mathcal{S}$.

which is a contradiction. Therefore $B_C(f, M/2) \cap B_C(g, M/2) = \varnothing$. $\square$

Having fixed a topology on Y^X, we now want to address the question of converging sequences more closely. Remember that our goal is to prove an Ascoli Theorem. This means that we want to have a result that makes sure we have continuous limits for sequences in $\mathcal{C}^0(X, Y)$. Similar to the classical version of the Theorem, we need some additional restrictions on these sequences.

Definition 4.1.9. *Let X be a topological space and (Y, d) a metric space. A subset $\mathcal{F} \subseteq Y^X$ is called equicontinuous at $x_0 \in X$ if for any $\epsilon > 0$ there exists a neighbourhood $U \subseteq X$ of x_0 such that*

$$d(f(x), f(x_0)) < \epsilon$$

for all $x \in U$ and all $f \in \mathcal{F}$.
We call $\mathcal{F}$ equicontinuous if it is equicontinuous at every point $x_0 \in X$.

The following version of Ascoli's Theorem is stated and proven in [Mun14, Thm 47.1].

Theorem 4.1.10. *Let X be a topological space and (Y, d) a metric space. Consider $\mathcal{C}^0(X, Y)$ with the topology of compact convergence. Let $\mathcal{F} \subseteq \mathcal{C}^0(X, Y)$.*

(i) If $\mathcal{F}$ is equicontinuous and the set

$$\mathcal{F}_{x_0} = \{f(x_0) \mid f \in \mathcal{F}\}$$

has compact closure for all $x_0 \in X$, then $\mathcal{F}$ is contained in a compact subspace of $\mathcal{C}^0(X, Y)$.

(ii) The converse holds if X is locally compact and Hausdorff.

We will use this result in order to prove the main statement of this subsection, which reads

Theorem 4.1.11 (Ascoli)**.** *Let X be a locally compact, second countable Hausdorff space and let (Y, d) be a complete metric space. Assume that the sequence $\{f_n\}_{n \in \mathbb{N}} \subseteq \mathcal{C}^0(X, Y)$ is equicontinuous and that for each $x_0 \in X$ the set*

$$\mathcal{F}_{x_0} = \bigcup_{n \in \mathbb{N}} \{f_n(x_0)\}$$

is bounded with respect to d. Then there exists a continuous function f and a subsequence $\{f_m\}$ which converges to f uniformly on each compact subset of X.

Before providing the proof, we need to collect a few more additional results.
First note, that the completeness assumption of the metric space comes with a beautiful

generalization of Heine-Borel's Theorem. In order to state it in a nice way, we need the following definition

Definition 4.1.12. *A metric space* (Y, d) *is totally bounded if for each* $\epsilon > 0$ *there exists a finite cover of* Y *by* ϵ*-balls.*

It turns out that this property encodes compactness for complete metric spaces as showcased below. We will just give the result, a proof can be found in [Mun14, Thm 45.1].

Theorem 4.1.13 (Heine-Borel)**.** *A metric space* (Y, d) *is compact if and only if it is complete and totally bounded.*

An important part of the proof for the main Theorem 4.1.11 is contained in the next lemma.

Lemma 4.1.14. *Let* X *be a locally compact Hausdorff space and* (Y, d) *complete. Give* $\mathcal{C}^0(X, Y)$ *the topology of compact convergence. A subspace* $\mathcal{F} \subseteq \mathcal{C}^0(X, Y)$ *has compact closure if and only if* $\mathcal{F}$ *is equicontinuous and the set*

$$\mathcal{F}_{x_0} = \{f(x_0) \,|\, f \in \mathcal{F}\}$$

is bounded for every $x_0 \in X$*.*

Proof. The if direction is just an application of 4.1.10. On the other hand, let $\mathcal{F}$ be equicontinuous under d and $\mathcal{F}_{x_0}$ bounded for all $x_0 \in X$. First, we want to use Heine-Borel (Theorem 4.1.13) on $(\overline{\mathcal{F}_{x_0}}, d)$ to show that it is compact. Completeness carries over to closed subsets, so we only need to demonstrate totally boundedness.
Let $\epsilon > 0$, we want to find finetely many balls $B_\epsilon(x_1), \dots, B_\epsilon(x_n)$ which cover the space. Since each F_{x_0} is bounded, there exists a number $M \in \mathbb{R}$ such that $\operatorname{diam}(\mathcal{F}_{x_0}) < M$. Consequently $n > M/\epsilon$ many balls are sufficient. By Heine-Borel it now follows that $\overline{\mathcal{F}_{x_0}}$ is compact.
Applying Theorem 4.1.10 again, we find that $\mathcal{F}$ is contained in a compact subset of $\mathcal{C}^0(X, Y)$. Since $\mathcal{C}^0(X, Y)$ is Hausdorff by Proposition 4.1.3, compactness for $\overline{\mathcal{F}}$ follows.

□

We close this subsection with the proof of the main statement.

Proof of Theorem 4.1.11. Give $\mathcal{C}^0(X, Y)$ the topology of compact convergence. We want to show that $\mathcal{C}^0(X, Y)$ is first countable. Because in first countable Hausdorff spaces compactness implies sequential compactness[15], the Theorem follows after applying Lemma 4.1.14 to $\mathcal{F} = \bigcup_{n\in\mathbb{N}} f_n$.
Let $f \in \mathcal{C}^0(X, Y)$, we need to show that there exists a countable neighbourhood-basis. Recall[16], that for second countable, locally compact Hausdorff spaces X, there exists a (strong)

[15] See [Lee11, Lemma 4.42f].
[16] See [Lee11, Prop. 4.76].

exhaustion by compact sets, i.e. a sequence $\{K_i\}_{i\in\mathbb{N}}$ of compact subsets of X such that $X = \bigcup_{i\in\mathbb{N}} K_i$ and $K_i \subseteq \operatorname{int} K_{i+1}$. Given such a sequence, we define $\mathcal{B} = \{B_{K_i}(f, 1/n) \,|\, i, n \in \mathbb{N}\}$, where $B_{K_i}(f, 1/n)$ are the balls from definition 4.1.1. Clearly $\mathcal{B}$ is countable and contains neighbourhoods of f. Let $U \subseteq C^0(X, Y)$ be any neighbourhood of f. We will show that there exists $B \in \mathcal{B}$ such that $B \subseteq U$.

$\{B_C(g, \epsilon) \,|\, C \subseteq X \text{ compact}, \epsilon > 0, g \in \mathcal{C}^0(X, Y)\}$ is a basis for our topology, so there is some element such that $f \in B_C(g, \epsilon) \subseteq U$. By compactness, C can be covered with finitely many sets $\operatorname{int} K_i$. Let j be the maximum index of this finite cover. By construction, $C \subseteq K_j$. Define $M := \sup_{x\in C} \{d(f(x), g(x))\} < \epsilon$ and take $B_{K_j}(f, 1/n) \in \mathcal{B}$ for some $n \in \mathbb{N}$ with $1/n < \epsilon - M$. Then $h \in B_{K_j}(f, 1/n)$ and the triangle inequality implies

$$\begin{aligned}
\sup_{x\in C} \{d(h(x), g(x))\} &\leq \sup_{x\in C} \{d(h(x), f(x))\} + \sup_{x\in C} \{d(f(x), g(x))\} \\
&\leq \sup_{x\in K_j} \{d(h(x), f(x))\} + M \\
&< 1/n + M \\
&< \epsilon.
\end{aligned}$$

Therefore $B_{K_j}(f, 1/n) \subseteq B_C(g, \epsilon)$ and we conclude that $\mathcal{B}$ indeed is the desired neighbourhood basis, finishing the proof.

□

4.2 Causal $\mathcal{C}^0$-Curves and the Limit Curve Lemma

After the discussion on convergence in the space of continuous functions, we want return to the setting of Semi-Riemannian manifolds. In particular, we want to prove that by applying the Ascoli Theorem from the previous section, one obtains a limit curve result which respects the additional causal structure on a Lorentzian manifold. Our primary references are given by [BEE96] and [Kri99].

Let us consider a spacetime (M, g) and endow it with an additional complete Riemannian background metric h, which we know to exist by Theorem 2.1.9. The need for this auxiliary metric h may be apparent from Ascoli's Theorem, where a complete metric space is required. Now, if we look at a sequence of curves on M, it is not expected a priori that we will have convergence[17] in the class of piecewise smooth functions but rather that a limit curve is merely continuous. Certainly, we still want some sensible notion of causality to be retained for this continuous limit.

Definition 4.2.1. *A continuous ($\mathcal{C}^0$) curve $\gamma : I \to M$ is called causal (resp. timelike) if for any $t_0 \in I$ there exists a convex set $C \subseteq M$ around $\gamma(t_0)$ and an open interval $J \subseteq I$ around t_0 such that for any pair $t_1, t_2 \in J$ with $t_1 < t_2$ there exists a smooth causal (resp. timelike)*

[17] With respect to the topology of compact convergence.

curve contained in C which connects $\gamma(t_1)$ and $\gamma(t_2)$.
Furthermore, we call γ future directed (resp. past directed) if the connecting curves are future directed (resp. past directed).

Remark 4.2.2. Any (piecewise) smooth causal curve is automatically causal in the above sense, so these definitions are nicely compatible. See Figure 4 for an example of the non-smooth case.
While it is sufficient to find a single convex neighbourhood of each $\gamma(t_0)$ such that the property in the definition is satisfied, we could have been more restrictive and have it required to hold for any convex neighbourhood instead. This is not necessary however, since both of these definitions are actually equivalent. This can be seen by the fact that $\gamma(t_1) \leq \gamma(t_2)$ for $t_1, t_2 \in I$ with $t_1 < t_2$ implies that any radial geodesic connecting two points on γ are causal by Proposition 2.2.18.

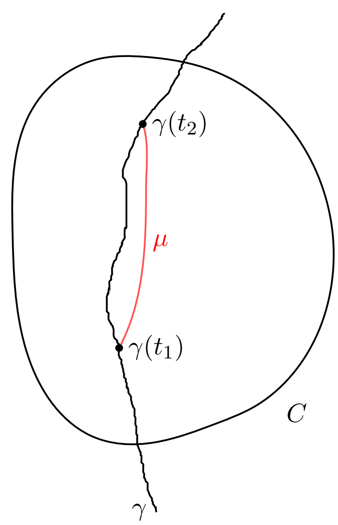


Figure 4: Definition 4.2.1. μ is a smooth causal curve connecting $\gamma(t_1)$ and $\gamma(t_2)$.

As shown below, $\mathcal{C}^0$ causal curves satisfy a local Lipschitz condition. We follow [Kri99, Lemma 8.2.1.]

Proposition 4.2.3. *For all $p \in M$ there exists a chart $(U, (x^0, \ldots, x^n))$ around p with U convex and a constant $K > 0$ such that all causal curves $\gamma : I \to M$ can be parametrized by x^0 and*

$$\|\gamma(t_1) - \gamma(t_2)\|_2 \leq K|t_1 - t_2|, \tag{4.1}$$

where $\|\gamma(t)\|_2$ denotes the usual Euclidean norm of $\gamma(t)$ in the new coordinates.

Proof. Let $p \in M$. We first show that there is a chart $(U, (x^0, \ldots, x^n))$ around p such that U is convex and x^0 has a timelike gradient[18].

[18] Cf. [KS21, Definition 3.2.6].

For this, take normal coordinates $(V, (x^0, \ldots, x^n))$ at p. Recall that in local coordinates, the gradient is given by $\operatorname{grad} x^0 = g^{ab} \frac{\partial x^0}{\partial x^a} \frac{\partial}{\partial x^b} = g^{0a} \frac{\partial}{\partial x^a}$. Since $g|_p = \eta$, it follows that $\operatorname{grad} x^0|_p = -\frac{\partial}{\partial x^0}\big|_p$. Thus, one obtains

$$g(\operatorname{grad} x^0, \operatorname{grad} x^0)|_p = \eta\left(\left.\frac{\partial}{\partial x^0}\right|_p, \left.\frac{\partial}{\partial x^0}\right|_p\right) = \eta_{00} = -1,$$

meaning that $\operatorname{grad} x^0$ is timelike at p.
Continuity then implies that the gradient is still timelike on some open subset $\tilde{U} \subseteq V$ of p. Furthermore, by Theorem 2.1.7, there exists a convex neighbourhood U of p with $\overline{U} \subseteq \tilde{U}$ and $\overline{U}$ compact. Note that $(U, (x^0, \ldots, x^n))$ is the chart we are looking for.
Without loss of generality let $\operatorname{grad} x^0$ have negative time orientation. We now show that x^0 increases (resp. decreases) along any future (resp. past) directed $\mathcal{C}^0$ causal curve γ, i.e. $\gamma^0(t_1) < \gamma^0(t_2)$ (resp. $>$) for $t_1 < t_2$.
Since γ is causal, there exists a smooth, future (resp. past) directed causal curve $\mu : [a, b] \to M$ with $\mu(a) = \gamma(t_1)$ and $\mu(b) = \gamma(t_2)$. Recall that by Lemma 2.2.7 $g(\operatorname{grad} x^0, \dot{\mu})(t) > 0$ (resp. < 0). On the other hand, by the definition of the gradient

$$g(\operatorname{grad} x^0, \dot{\mu}(t)) = \dot{\mu}(t)(x^0) = \frac{\mathrm{d}(x^0 \circ \mu)}{\mathrm{d}\, t}(t),$$

which implies that $\mu^0(a) < \mu^0(b)$ and in turn $\gamma^0(t_1) < \gamma^0(t_2)$ (resp. > 0).
We conclude that x^0 is monotonous along any causal curve γ and may thereby be used as a parametrization, proving the first part of the Proposition.
By continuity, for U small enough and due to compactness of $\overline{U}$, there is a constant $k_0 > 0$ such that any causal vector is also causal with respect to the flat metric

$$\tilde{g} = -k_0\, dx^0 \otimes dx^0 + \sum_{i=1}^{n} dx^i \otimes dx^i.$$

Take γ and μ like before and parametrize with respect to x^0. We have $\gamma(t_1) = \mu(t_1)$, $\gamma(t_2) = \mu(t_2)$ and $\mu^a(t)e_a = (t, x^1(t), \ldots, x^n(t))$, where $\{e_0, \ldots, e_n\}$ denotes the standard basis of $\mathbb{R}^{1,n}$. Note, that $\dot{\mu}^0(t) = 1$ and $\tilde{g}(\dot{\mu}(t), \dot{\mu}(t)) < 0$ imply

$$k_0 \geq \sum_{i=1}^{n} \left(\dot{\mu}^i(t)\right)^2$$

for all t. Therefore, if $t_1 < t_2$

$$\begin{aligned}\|\gamma(t_1) - \gamma(t_2)\|_2 &= \|\mu(t_1) - \mu(t_2)\|_2 = \|\mu^a(t_1)e_a - \mu^a(t_2)e_a\|_2 \\ &= \left\|\int_{t_2}^{t_1} \dot{\mu}^a(t)e_a\, \mathrm{d}t\right\|_2 \leq \int_{t_1}^{t_2} \|\dot{\mu}^a(t)e_a\|_2\, \mathrm{d}t\end{aligned}$$

$$
\begin{aligned}
&= \int_{t_1}^{t_2} \left[\left(\dot{\mu}^0(t)\right)^2 + \sum_{i=1}^{n} \left(\dot{\mu}^i(t)\right)^2 \right]^{1/2} \mathrm{d}t \\
&\le \int_{t_1}^{t_2} \sqrt{1+k_0}\, \mathrm{d}t = \sqrt{1+k_0}\, |t_1 - t_2|,
\end{aligned}
$$

which finishes the proof. □

The fact that γ is locally Lipschitz in the above coordinates has a surprising consequence following from the subsequent result in measure theory due to Rademacher. For a proof we refer to [EG15, Thm 3.2].

Theorem 4.2.4 (Rademacher)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be locally Lipschitz. Then f is differentiable almost-everywhere*[19]*.*

The previous Proposition and the Rademacher Theorem then immediately imply the Corollary below, which as a result is interesting on itself but which for our purposes is primarily of technical significance.

Corollary 4.2.5. *Let $\gamma : I \to M$ be a causal and $(U, (x^0, \ldots, x^n))$ like before. Then $\dot{\gamma}^i(t)$ exists almost-everywhere and there is a $K > 0$ such that $|\dot{\gamma}^i(t)| < K$ for all $i = 0, \ldots, n$, where γ^i are the chart components of γ.*

Proof. The first part follows by Theorem 4.2.4. For the second part we have by Equation (4.1) that

$$
|\gamma^i(t+h) - \gamma^i(t)| \le \|\gamma(t+h) - \gamma(t)\|_2 \le K|h|.
$$

Dividing and taking the appropriate limit on both sides then proves the claim. □

The final preparation for the proof of the Limit Curve Lemma is related to the complete Riemannian metric h on M and is again of technical nature. We follow [BEE96, p. 74-76].

Lemma 4.2.6. *Let $\gamma : I \to M$ be a causal with I compact. Then $L_h(\gamma)$ is finite.*

Proof. First note that $L_h(\gamma)$ is well-defined for γ causal by 4.2.5. Let $t_1, t_2 \in I$ with $t_1 < t_2$ and such that $\gamma|_{[t_1,t_2]}$ is contained in a chart $(U, (x^0, \ldots, x^n))$ as in the proof of Proposition 4.2.3.

Denote by $\overline{h}$ be the supremum of the metric components h_{ab} on the compact set $\overline{U}$. Then

$$
L_h(\gamma|_{[t_1,t_2]}) = \int_{t_1}^{t_2} \sqrt{h(\dot{\gamma}, \dot{\gamma})(t)}\, \mathrm{d}t = \int_{t_1}^{t_2} \left[\sum_{a,b} h_{ab}|_{\gamma(t)} \dot{\gamma}^a(t) \dot{\gamma}^b(t) \right]^{1/2} \mathrm{d}t
$$

[19] With respect to the Lebesgue-Measure on $\mathbb{R}^n$.

$$\leq \int_{t_1}^{t_2} \sqrt{K^2(n+1)^2\overline{h}}\,\mathrm{d}t = K(n+1)\overline{h}^{1/2}|t_2 - t_1|,$$

where we have used Corollary 4.2.5. Hence, L_h is locally bounded.
This, together with compactness of I and the fact that we can cover M by a locally finite collection of the chart neighbourhoods U, we obtain the result. $\square$

We may now proceed to show the main result of this section, which concerns limits of sequences of future inextendible, piecewise smooth causal curves. The proof itself proceeds in three steps. One starts with an application of Ascoli's Theorem to show that a limit exists and then continues with two separate arguments for future inextendibility and causality. We closely follow [BEE96, Prop. 3.31].

Lemma 4.2.7 (Limit Curve Lemma)**.** *Let (M, g) be a spacetime and h a complete Riemannian metric on M. Further, let $\gamma_n : [0, \infty) \to M$ be a sequence of future inextendible causal curves parametrized by h-arclength and $p \in M$ an accumulation point of $\gamma_n(0)$. Then there exists a subsequence $\{\gamma_k\}_{k\in\mathbb{N}}$ that converges to a future inextendible continuous causal curve $\gamma : [0, \infty) \to M$ in the topology of compact convergence with $\gamma(0) = p$.*

Proof. The first step is to get to a point where we are able to apply Ascoli's Theorem. Note that, since the γ_n are parametrized with respect to h-arclength, we have

$$d_h(\gamma_n(t_1), \gamma_n(t_2)) \leq |t_1 - t_2| \quad \forall n \in \mathbb{N}, \forall t_1, t_2 \in [0, \infty) \tag{4.2}$$

which implies that $\{\gamma_n\}_{n\in\mathbb{N}}$ is equicontinuous.
Furthermore, as $\gamma_n(0) \to p$ for $n \to \infty$, it follows that there exists an $N \in \mathbb{N}$ such that $d_h(\gamma_n(0), p) < 1$ for all $n \geq N$. Consequently we obtain, using the triangle inequality and (4.2), that for any fixed $t_0 \in [0, \infty)$ the set $\gamma_n|_{[0,t_0]}$ is contained in the closed, compact ball $\{q \in M \mid d_h(p, q) \leq 1 + t_0\}$ for all $n \geq N$. Therefore the union $\bigcup_{n\in\mathbb{N}}\{\gamma_n(t_0)\}$ is bounded for any t_0.
Now we are able to use Theorem 4.1.11 to obtain a subsequence $\{\gamma_k\}_{k\in\mathbb{N}}$ which converges to a continuous function $\gamma : [0, \infty) \to M$ uniformly with respect to h on any compact subset of $[0, \infty)$. Naturally we have $\gamma_k(0) \to p = \gamma(0)$, so we are done with the first part.
All that is left to show now is that γ is causal and future inextendible. We begin with the prior: Let $t_1 \in [0, \infty)$ and let $U \subseteq M$ be a convex set around $\gamma(t_1)$. Take $\delta > 0$ such that $\{q \in M \mid d_h(\gamma(t_1), q) < \delta\} \subseteq U$. We show that for $t_1 < t_2 < t_1 + \delta$ we have $\gamma|_{[t_1,t_2]} \subseteq U$. First, there is an $\epsilon > 0$ such that $t_2 = t_1 + \delta - \epsilon$. By uniform convergence it follows that for k sufficiently large $d_h(\gamma_k(t_1), \gamma(t_1)) < \epsilon$. This, together with the triangle inequality and (4.2) indeed gives

$$d_h(\gamma(t_1), \gamma_k(t_2)) \leq d_h(\gamma(t_1), \gamma_k(t_1)) + d_h(\gamma_k(t_1), \gamma_k(t_2)) < \epsilon + |t_1 - (t_1 + \delta - \epsilon)| = \delta.$$

Similarly, if $t_1 \neq 0$ and δ small enough, then for $t_1 - \delta < t_2 < t_1$ one obtains $\gamma|_{[t_2,t_1]} \subseteq U$.
Let $s, t \in (t_1 - \delta, t_1 + \delta)$ (or $[0, \delta)$ resp.) with $s < t$. Again, uniform convergence implies $\gamma_k(s) \to \gamma(s)$ and $\gamma_k(t) \to \gamma(t)$. Note that the γ_k are future directed, i.e. for large k we obtain $\gamma_k(s) \leq_U \gamma_k(t)$. Now recall that $\leq$ is closed on convex sets (Proposition 2.2.18). Thus $\gamma(s) \leq_U \gamma(t)$ and γ is causal.
At last, we show future inextendibility. Assume, to the contrary, that γ is not inextendible. Then there exists a $q \in M$ with $\gamma(t) \to q$ as $t \to \infty$. Take a chart $(U, (x^0, \ldots, x^n))$ around q with $\operatorname{grad} x^0$ timelike as in Proposition 4.2.3. Let V be a convex neighbourhood of q and relatively compact in U. Note that there is some t_0 for which $\gamma|_{[t_0,\infty)} \subseteq V$. Lemma 4.2.6 then tells us that the length of any causal curve in V between the level sets $(x^0)^{-1}(\gamma^0(t_0))$ and $(x^0)^{-1}(x^0(q))$ is bounded since we can always parametrize these curves by x^0. One may therefore conclude that the length is bounded by some $\delta > 0$.
However, taking into account that the γ_k are parametrized with respect to h-arclength, one finds $\gamma_k|_{[t_0+1,t_0+\delta+2]} \subseteq (x^0)^{-1}([\gamma^0(t_0), x^0(q)])$ for large enough k. But this yields a contradiction since $L_h(\gamma_k|_{[t_0+1,t_0+\delta+2]}) = \delta + 1 > \delta$. $\square$

Remark 4.2.8. Recalling Remark 4.1.4, we again emphasize that the Riemannian metric h only serves as an auxiliary tool to facilitate our arguments. After all, we are only really interested in the structure of the spacetime (M, g) itself.
There are numerous results in similar spirit, found in the literature, which are also called *Limit Curve Lemma*. For our purposes however, the above statement is enough. Still, we refer the avid reader to [Min08], in which one can find a proper review on the history of such results and an elegant generalization thereof.

4.3 Variation of Causal Curves

Now that we have established the Limit Curve Lemma, we want to expand our existing toolkit by some elementary notions from the calculus of variation on differentiable manifolds. We will introduce the concept of varying piecewise smooth curves just as one would do for Riemannian manifolds. However, taking into account that we want to work on a spacetime (M, g), we want to ensure that the underlying causal structure is respected as well. The key result of this section is to demonstrate that causal curves with fixed end points, which are not null pregeodesics, can be deformed into timelike curves with the same endpoints and, in fact, such that they lie arbitrarily close to the original curve with respect to the topology of compact convergence[20]. After that, we discuss variations of geodesics and their connection to Jacobi fields, which will be especially important for Chapter 7. Our main references are [O'N83], [KS23] and [dC92].
Let (M, g) be any Semi-Riemannian (or Riemannian) manifold. We make the following definition

[20]See Remark 4.1.7.

Definition 4.3.1. *Let $\gamma : [a, b] \to M$ be a piecewise smooth curve. A variation of γ is a continuous map $\alpha : [a, b] \times (-\epsilon, \epsilon) \to M$ with $\epsilon > 0$ such that*

(i) $\alpha(t, 0) = \gamma(t)$ *for all* $t \in [a, b]$,

(ii) there exist break points $t_0 \coloneqq a < t_1 < \cdots < t_{k+1} \coloneqq b$, *such that the restriction of* α *to* $[t_{i-1}, t_i] \times (-\epsilon, \epsilon)$ *is smooth for all* $i \in \{1, \ldots, k\}$.

The variation α *of* γ *is called proper, if* $\alpha(a, s) = \gamma(a)$ *and* $\alpha(b, s) = \gamma(b)$ *for each* $s \in (-\epsilon, \epsilon)$.

Remark 4.3.2. For a variation $\alpha : [a, b] \times (-\epsilon, \epsilon) \to M$ of a piecewise smooth curve γ, we will mainly use the shorthand notation $\gamma_s(t) \coloneqq \alpha(t, s)$. Note that $t \mapsto \gamma_s(t)$ is a piecewise smooth curve for each $s \in (-\epsilon, \epsilon)$ respectively. Therefore, $s \mapsto \gamma_s$ can be understood as the *deformation* of γ into the neighbouring curves γ_s. The parameter $t \in [a, b]$ also defines a one-parameter family of smooth curves $s \mapsto \gamma_s(t)$, which are called *transversal curves.* These transversal curves then define a well-defined, piecewise smooth vector field $V \coloneqq \left.\frac{\partial \gamma_s}{\partial s}\right|_{s=0}$ along γ, i.e. $V \in \mathfrak{X}_{\mathrm{pw}}(\gamma)$, the *variation vector field* of γ_s.

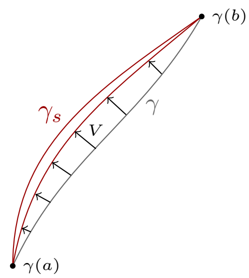


Figure 5: A proper variation γ_s of a smooth curve $\gamma : [a, b] \to M$ with variation vector field V.

See Figure 5 for a visualization of Definition 4.3.1 and Remark 4.3.2. An important aspect of the construction of variations is stated in the next Lemma, for which we follow [dC92, Ch. 9, Prop. 2.2].

Lemma 4.3.3. *Let* $\gamma : [a, b] \to M$ *be a piecewise smooth curve and* $V \in \mathfrak{X}_{pw}(\gamma)$. *Then there exists a variation* γ_s *of* γ *such that* V *is the variational vector field of* γ_s. *In addition, if* $V(a) = V(b) = 0$, *then* γ_s *is proper.*

Proof. Because $\gamma([a, b]) \subseteq M$ is compact we can cover it by finitely many convex sets $U_1, \ldots, U_n$. For each $i = 1, \ldots, n$ choose $\delta_i > 0$ such that $\exp_{\gamma(t)}(B_{\delta_i}(0)) \subseteq U_i$ for all $t \in [a, b]$ with $\gamma(t) \subseteq U_i$ and where $B_{\delta_i}(0) \subset T_{\gamma(t)}M$ is the Euclidean ball with radius δ_i around $0 \in T_{\gamma(t)}M$. Now let $\delta \coloneqq \min\{\delta_1, \ldots, \delta_n\}$ and we have shown that for each $t \in [a, b]$ the map $\exp_{\gamma(t)}$ is well-defined for all $v \in T_{\gamma(t)}M$ with $\|v\|_2 < \delta$, where $\|\cdot\|_2$ denotes the Euclidean norm on $T_{\gamma(t)}M$.

By choosing $N := \max_{t\in[a,b]} \|V(t)\|_2$ and $\epsilon < \frac{\delta}{N}$, the map

$$\begin{aligned} \alpha : &[a,b] \times (-\epsilon, \epsilon) \to M \\ &(t,s) \mapsto \exp_{\gamma(t)}(sV(t)) \end{aligned}$$

is well-defined. Furthermore, given that $\exp_{\gamma(t)}(sV(t))$ is the radial geodesic $s \mapsto \sigma(s)$ starting from $\gamma(t)$ with initial speed $V(t)$ and the fact that the solutions to the geodesic equation depend smoothly on the initial conditions, one can conclude that α is piecewise smooth. Then we also have

$$\alpha(t,0) = \exp_{\gamma(t)}(0) = \gamma(t)$$

and therefore α is a variation of γ. Its variation vector field is given by

$$\begin{aligned} \left.\frac{\partial \alpha}{\partial s}\right|_{s=0}(t) &= T_{(t,0)}[\exp_{\gamma(t)}(sV(t))](\partial_s) = T_0 \exp_{\gamma(t)} \circ T_{(t,0)}(sV(t))(\partial_s) \\ &= T_0 \exp_{\gamma(t)}(V(t)) = V(t), \end{aligned}$$

where we have used that $T_0 \exp_{\gamma(t)}(v) = v$ for any $v \in T_0(T_{\gamma(t)}M) \cong T_{\gamma(t)}M$ (see e.g. [KS21, Eq. (2.1.14)]) and thus α is the desired variation.
Finally, if $V(a) = V(b) = 0$, it immediately follows that α is proper because

$$\alpha(a,s) = \exp_{\gamma(a)}(0) = \mathrm{id}_s = \exp_{\gamma(b)}(0) = \alpha(b,s),$$

where id_s is the identity map with respect to s on $(-\epsilon, \epsilon)$. □

Of course, we are interested in the case where (M, g) is a spacetime, which we will assume to be the case for the rest of the section. One then has the following important Lemma [O'N83, Ch. 10, Lemma 45].

Lemma 4.3.4. *Let $\gamma : [a,b] \to M$ be a smooth causal curve and let γ_s be a variation of γ with variation vector field $V \in \mathfrak{X}(\gamma)$. If $\langle V', \dot\gamma\rangle(t) < 0$ for all $t \in [a,b]$, then there exists $\epsilon > 0$ such that neighbouring curves γ_s are timelike for each $s \in (0, \epsilon)$.*

Proof. By assumption, one has

$$\left\langle \frac{\partial \gamma_s}{\partial t}, \frac{\partial \gamma_s}{\partial t} \right\rangle (t,0) = \langle \dot\gamma(t), \dot\gamma(t)\rangle \leq 0.$$

Note also that

$$\begin{aligned} \frac{\partial}{\partial s}\left\langle \frac{\partial \gamma_s}{\partial t}, \frac{\partial \gamma_s}{\partial t} \right\rangle (t,0) &= 2\left\langle \frac{\nabla}{\partial_s}\frac{\partial \gamma_s}{\partial t}, \frac{\partial \gamma_s}{\partial t} \right\rangle (t,0) = 2\left\langle \frac{\nabla}{\partial_t}\frac{\partial \gamma_s}{\partial s}, \frac{\partial \gamma_s}{\partial t} \right\rangle (t,0) \\ &= 2\left\langle \frac{\nabla}{\partial_t}\left[\left.\frac{\partial \gamma_s}{\partial s}\right|_{s=0}\right](t), \frac{\partial \gamma_s}{\partial t}(t,0) \right\rangle = 2\left\langle V'(t), \dot\gamma(t)\right\rangle \end{aligned}$$

$$< 0,$$

where we have used the fact that the s-derivative and t-derivate commute [Lemma 2.1.6 (i)]. For fixed t, the Taylor expansion of $\left\langle \frac{\partial \gamma_s}{\partial t}, \frac{\partial \gamma_s}{\partial t} \right\rangle (t,s)$ is given by

$$\left\langle \frac{\partial \gamma_s}{\partial t}, \frac{\partial \gamma_s}{\partial t} \right\rangle (t,s) = \left\langle \frac{\partial \gamma_s}{\partial t}, \frac{\partial \gamma_s}{\partial t} \right\rangle (t,0) + \frac{\partial}{\partial s} \left\langle \frac{\partial \gamma_s}{\partial t}, \frac{\partial \gamma_s}{\partial t} \right\rangle (t,0)\, s + \mathcal{O}(s^2)$$

and we can conclude by the previous equations, that there exists an $\epsilon > 0$ small enough, such that the above expression stays negative for $s \in (0, \epsilon)$. By compactness this is true uniformly in t. $\square$

Naturally, the same statement holds for general causal curves (piecewise smooth) if one applies the above Lemma to each smooth segment.
Our goal now is to show that there is a proper and timelike variation for general causal curves between two points $p, q \in M$, i.e. we want to be able to approximate causal curves between p and q by timelike curves to arbitrary precision. However, this cannot be true in general, as showcased in Figure 6.

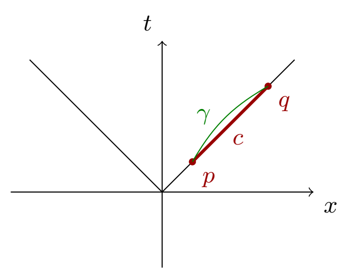


Figure 6: A null geodesic c between two points p, q in Minkowski space $\mathbb{R}^{1,1}$. Any other curve between p and q, e.g. γ, cannot be timelike, because then there need to be points on γ, where it has spacelike tangent.

This example is reminiscent of the push-up principle (Proposition 2.2.10), in that we can think of the situation very physically as follows: As one travels a null geodesic, the velocity is always equal to the speed of light. Slowing down then should imply that we have to make up for the lost time at other parts, i.e. one has to exceed the speed of light somewhere else, which is physically impossible. While this reasoning seems to hold true on a level of physical intuition, it actually turns out to be wrong. In fact, one can show that if any nonspacelike curve γ between $p, q \in M$ has a conjugate point, that then there exists a variation γ_s, from which one obtains a timelike curve between p and q (cf. [HE73, Prop. 4.5.12]). Roughly speaking[21], conjugate points are points where nearby geodesics almost meet. By this, one should intuitively think of the existence of a one-parameter family of geodesics that meets the

[21]We will discuss conjugate points and *focal points* in Chapter 7.

conjugate point in first order. An important result concerning conjugate points is that after such points a causal geodesic fails to be maximizing. Note that the existence of such points is closely related to the underlying geometric structure of M as we will see later. As a final remark, one might add that our intuition holds true at least not only in Figure 6 but also for any two dimensional spacetime since null geodesics do not exhibit conjugate points in this case (see [BEE96, Ch. 10.3, Lemma 10.45]).
Without putting any attention, for this moment, to the particular case of null geodesics (or in fact null pregeodesics) with conjugate points, we may now show the following ([O'N83, Ch. 10, Prop. 46] and [KS23, Lemma 3.2.4])

Proposition 4.3.5. *Let γ be a causal curve between $p \in M$ and $q \in M$. If γ is not a null pregeodesic, then there exists a timelike curve between p and q which is arbitrarily close to γ with respect to the topology of compact convergence.*

Proof. We will show that there exists a proper timelike variation of γ. By considering the compact-open topology and using Theorem 4.1.6, this then implies the above result.
Without loss of generality, assume that $\gamma : [0,1] \to M$ with $\gamma(0) = p$, $\gamma(1) = q$ (otherwise reparametrize). We first discuss the following two cases.
(i) *$\dot\gamma(t_0)$ is timelike for some $t_0 \in [0,1]$*
Assume first, that $t_0 = 1$ and define $W \in \mathfrak{X}_{\mathrm{pw}}(\gamma)$ by the parallel transport of $\dot\gamma(1)$ along γ. Thus $\langle W, W\rangle = \text{const.} < 0$ and $W(t)$ is timelike for all $t \in [0,1]$. Since $\dot\gamma(t)$ and $W(t)$ lie in the same light cone, it must follow that $\langle W, \dot\gamma\rangle < 0$. By continuity, one can find a $\delta > 0$ such that $\langle \dot\gamma(t), \dot\gamma(t)\rangle < 0$ on $[1-\delta, 1]$. Let $f : [0,1] \to \mathbb{R}$ be smooth with $f(0) = 0 = f(1)$ and $f' > 0$ on $[0, 1-\delta]$. Define $V = fW$ and consider the proper variation γ_s of γ, defined by the variation vector field V, which exists by Lemma 4.3.3. We then have

$$\langle V', \dot\gamma\rangle = \langle f'W + fW', \dot\gamma\rangle = f'\langle W, \dot\gamma\rangle,$$

which is negative on $[0, 1-\delta]$. Using Lemma 4.3.4, we conclude that $t \mapsto \gamma_s(t)$ is timelike for small s on $[0, 1-\delta]$. But since also, by continuity of $s \mapsto \gamma_s(t)$, γ_s stays timelike on $[1-\delta, 1]$ for s small, we have found a proper timelike deformation[22]. The $t_0 = 0$ case follows analogously. Finally, if $t_0 \in (0,1)$ instead, we can just apply the above argument to the restrictions $\gamma|_{[0,t_0]}$ and $\gamma|_{[t_0,1]}$ respectively and therefore are done.
(ii) *γ is smooth and everywhere null.*
We have $\langle \dot\gamma, \dot\gamma\rangle = 0$ and therefore

$$0 = \frac{d}{dt}\langle \dot\gamma, \dot\gamma\rangle = 2\langle \dot\gamma, \ddot\gamma\rangle,$$

[22] Note that this argument also serves as a proof of the push-up principle (Prop. 2.2.10).

which means that $\ddot{\gamma} \perp \dot{\gamma}$ on $[0,1]$. Since $\dot{\gamma}$ is null, we can write

$$\ddot{\gamma}(t) = \phi(t)\dot{\gamma}(t) + \xi(t)E(t)$$

for some spacelike $E \in \mathfrak{X}(\gamma) \perp \dot{\gamma}$ and $\phi, \xi \in \mathcal{C}^\infty([0,1])$. Note that $\xi \not\equiv 0$, because otherwise $\ddot{\gamma} = \phi\dot{\gamma}$, i.e. γ is a null pregeodesic, contrary to our assumption. Consequently, there exists $t_0 \in [0,1]$, such that $\xi(t_0) \neq 0$ and one has

$$\langle \ddot{\gamma}, \ddot{\gamma} \rangle(t_0) = [\phi(t_0)]^2 \langle \dot{\gamma}, \dot{\gamma} \rangle(t_0) + [\xi(t_0)]^2 \langle E, E \rangle(t_0) = [\xi(t_0)]^2 \langle E, E \rangle(t_0) > 0$$

since E is spacelike. By definition and the above, it follows that

$$\langle \ddot{\gamma}, \ddot{\gamma} \rangle \geq 0, \qquad \langle \ddot{\gamma}, \ddot{\gamma} \rangle \not\equiv 0. \tag{4.3}$$

Now let Y be the parallel transport of a timelike vector $Y_0 \in T_{\gamma(0)}M$ with $\langle Y_0, \dot{\gamma}(0) \rangle < 0$ along γ. Then $Y(t)$ is timelike for all $t \in [0,1]$ and lies in the same causal cone as $\dot{\gamma}(t)$ since parallel transport is an isometry, i.e.

$$\langle Y, \dot{\gamma} \rangle < 0. \tag{4.4}$$

We now want to show that there exists $X = \alpha Y + \beta\ddot{\gamma}$ with $\alpha, \beta \in \mathcal{C}^\infty([0,1])$ such that α and β vanish at the endpoints and $\langle X', \dot{\gamma} \rangle < 0$.
Note that because $\ddot{\gamma} \perp \dot{\gamma}$ and $Y' = 0$, one has

$$\langle X', \dot{\gamma} \rangle = \langle \dot{\alpha} Y + \dot{\beta}\ddot{\gamma} + \beta \,\dddot{\gamma}, \dot{\gamma} \rangle = \dot{\alpha}\langle Y, \dot{\gamma} \rangle + \langle \,\dddot{\gamma}, \dot{\gamma} \rangle = \dot{\alpha}\langle Y, \dot{\gamma} \rangle - \beta\langle \ddot{\gamma}, \ddot{\gamma} \rangle, \tag{4.5}$$

where the last equality stems from the fact that

$$0 = \frac{d}{dt}\langle \ddot{\gamma}, \dot{\gamma} \rangle = \langle \,\dddot{\gamma}, \dot{\gamma} \rangle + \langle \ddot{\gamma}\, \ddot{\gamma} \rangle.$$

We want to construct α and β such that (4.5) is negative. For this, let $\chi := \frac{\langle \ddot{\gamma}, \ddot{\gamma} \rangle}{\langle Y, \dot{\gamma} \rangle}$, which is defined since the denominator is non-vanishing by (4.4). Equations (4.3) and (4.4) together then imply that

$$\chi \leq 0, \qquad \chi \not\equiv 0.$$

We can therefore define $\beta \in \mathcal{C}^\infty([0,1])$ such that

$$\int_0^1 \beta(t)\chi(t)\, dt = -1$$

and $\beta(0) = \beta(1) = 0$. Furthermore, let $\alpha(t) := \int_0^t (\beta(s)\chi(s) + 1) ds$, which is clearly smooth and satisfies $\alpha(0) = \alpha(1) = 0$. From (4.5), one may then conclude that

$$\langle X', \dot{\gamma} \rangle = [\beta\chi + 1]\langle Y, \dot{\gamma} \rangle - \beta\langle \ddot{\gamma}, \ddot{\gamma} \rangle = [\beta\chi + 1]\langle Y, \dot{\gamma} \rangle - \beta\chi\langle Y, \dot{\gamma} \rangle = \langle Y, \dot{\gamma} \rangle < 0$$

by (4.4). Finally, just note that applying Lemma 4.3.4 gives the claim.

Suppose γ is now a piecewise causal curve, which is everywhere null and contains at least one segment which is not a null pregeodesic. By (ii) we can deform this segment to a timelike curve and then simply argue as in (i). Consequently, we only have one remaining case left to check.

(iii) *γ is piecewise smooth and every segment is a null pregeodesic.*

Assume that there exists only one breakpoint $t_0 \in (0,1)$. Let $\Delta\dot{\gamma} = \dot{\gamma}(t_+) - \dot{\gamma}(t_-)$, where $\dot{\gamma}(t_+)$ is the tangent vector of γ at t_0 with respect to the limit from above and $\dot{\gamma}(t_-)$ with respect to the limit from below (see Figure 7).

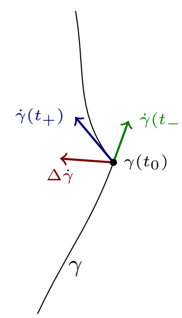


Figure 7: Setup for case (iii).

Define $W^{\pm}$ to be the parallel transport of $\dot{\gamma}(t_{\pm})$ along γ respectively and let $W := W^+ - W^-$, i.e. W is the parallel transport of $\Delta\dot{\gamma}$. For $t \in [0, t_0]$ one has that $\langle W^-, \dot{\gamma}\rangle = 0$ and $\langle W^+, \dot{\gamma}\rangle < 0$ as W^+ and $\dot{\gamma}$ lie in the same causal cone. Therefore

$$\langle W, \dot{\gamma}\rangle(t) = \langle W^+, \dot{\gamma}\rangle(t) - \langle W^-, \dot{\gamma}\rangle(t) < 0$$

for $t \in [0, t_0]$. Similarly, $\langle W^+, \dot{\gamma}\rangle = 0$ and $\langle W^-, \dot{\gamma}\rangle < 0$ on $[t_0, 1]$ and thus

$$\langle W, \dot{\gamma}\rangle(t) > 0$$

for $t \in [t_0, 1]$. Let f be a piecewise smooth function with vanishing endpoints and such that $f' > 0$ on $[0, t_0]$ and $f' < 0$ on $[t_0, 1]$ respectively. Now, simply define $V = fW$, for which then follows

$$\langle V', \dot{\gamma}\rangle = f'\langle W, \dot{\gamma}\rangle + f\langle W, \dot{\gamma}\rangle = f'\langle W, \dot{\gamma}\rangle < 0$$

for all $t \in [0,1]$, as $W' = 0$. Like before, Lemma 4.3.4 then implies that there exists a timelike variation. Finally note, that the case for multiple break points now simply follows from above and (i). $\square$

One can in fact generalize the above result to the case of continuous causal curves. This is also needed later, as we will have to work with a limit curve obtained by Lemma 4.2.7, which we only know to be continuous.

Corollary 4.3.6. *Let γ be a continuous causal curve between two points $p, q \in M$. If γ is not a null pregeodesic, then there exists a timelike curve between p and q which is arbitrarily close to γ with respect to the topology of compact convergence.*

Proof. Note that if γ is a continuous causal curve, we can always find a piecewise smooth causal curve arbitrarily close with respect to the compact-open topology. To achieve this, let U be a neighbourhood of γ. One can choose a collection of convex sets which cover γ and are contained in U. Now just consider a concatenation of radial geodesics. The claim then follows by applying the previous Proposition. $\square$

After the discussion of some results on the variation of causal curves, we finally want to look at a very particular type of variation. Namely variations $c_s : I \times (-\epsilon, \epsilon) \to M$ of a geodesic $c : I \to M$ for which each $t \mapsto c_s(t)$ is itself a geodesic. Such variations are called *geodesic variations*. We will show that these exist and that they are closely related to a very special class of vector field, the *Jacobi fields*.

Definition 4.3.7. *Let c be a geodesic. $J \in \mathfrak{X}(c)$ is a Jacobi field if it satisfies the following Jacobi equation (or Geodesic Deviation equation)*

$$\frac{\nabla^2}{\partial t^2} J = R_{J\dot{c}}\dot{c}. \tag{4.6}$$

Note that equation (4.6) is a linear system of second order ordinary differential equations. Consequently, by application of standard ODE theory, one has (see e.g. [Lee18, Prop. 10.2]).

Proposition 4.3.8. *Let $c : I \to M$ be a geodesic. Assume one has initial data $v, w \in T_{c(t_0)}M$ for some $t_0 \in I$. Then there exists a unique Jacobi field $J \in \mathfrak{X}(c)$ such that $J(t_0) = v$ and $\frac{\nabla J}{dt}(t_0) = w$.*

The connection of Jacobi fields and geodesic variations is evident from the next result. We follow [KS23, Lemma 3.2.9].

Proposition 4.3.9. *Let $c : I \to M$ be a geodesic and $J \in \mathfrak{X}(c)$. Then J is a Jacobi field if and only if there exists a geodesic variation $c_s : I \times (-\epsilon, \epsilon) \to M$ of c, with variation vector field J.*

Proof. We first show the if direction. Assume that J is the variation vector field of a geodesic variation c_s. Therefore, one has $\frac{\nabla}{\partial t}\frac{\partial c_s}{\partial t} = 0$ and we obtain

$$\begin{aligned}\frac{\nabla^2}{\partial t^2} J &= \frac{\nabla}{\partial t}\frac{\nabla}{\partial t}\left.\frac{\partial c_s}{\partial s}\right|_{s=0} = \frac{\nabla}{\partial t}\frac{\nabla}{\partial s}\left.\frac{\partial c_s}{\partial t}\right|_{s=0} \\ &= \frac{\nabla}{\partial s}\frac{\nabla}{\partial t}\left.\frac{\partial c_s}{\partial t}\right|_{s=0} + R_{\frac{\partial c_s}{\partial s}\frac{\partial c_s}{\partial t}}\left.\frac{\partial c_s}{\partial t}\right|_{s=0} = R_{J\dot{c}}\dot{c},\end{aligned}$$

where we have used Lemma 2.1.6.

For the only if direction, suppose that J is a Jacobi field. Assume $0 \in I$ (otherwise reparametrize c) and let σ be any smooth curve such that $\sigma(0) = c(0)$ and $J(0) = \sigma'(0)$. Let $A, B \in \mathfrak{X}(\sigma)$ be defined by the parallel transport of $\dot{c}(0)$ and $J'(0)$ along σ respectively. Define $X \in \mathfrak{X}(\sigma)$ by $X(s) = A(s) + sB(s)$. Note that

$$\begin{aligned} X(0) &= A(0) = \dot{c}(0) \\ \frac{\nabla X}{ds}(0) &= A'(0) + B(0) + sB'(0) = B(0) = J'(0), \end{aligned}$$

where we have used the fact that A and B are parallel. Now, similar to the proof of Lemma 4.3.3, set $c_s(t) \coloneqq \exp_{\sigma(s)}(tX(s))$. We will show that c_s is the required variation. For this, first note that $(t, s) \mapsto (\sigma(s), tX(s))$ is continuous and therefore maps the compact set $I \times \{0\}$ into a compact subset of TM. Furthermore, this subset lies in the domain of $\exp(p, v) = (p, \exp_p(v))$, which is open[23]. Consequently there exists an $\epsilon > 0$ such that $c_s : I \times (-\epsilon, \epsilon) \to M$ is well-defined.

Next note that

$$c_0(t) = \exp_{\sigma(0)}(tX(0)) = \exp_{c(0)}(t\dot{c}(0)) = c(t).$$

Since for each $s \in (-\epsilon, \epsilon)$, c_s is a (radial) geodesic, $(t, s) \mapsto c_s$ is indeed a geodesic variation of c. All that is left is to show that the variation vector field is given by J. Let $\tilde{J} \coloneqq \frac{\partial c_s}{\partial s}\big|_{s=0}$. Following from the proof of the if direction, $\tilde{J}$ is a Jacobi field. We will show that $J(0) = \tilde{J}(0)$ and $J'(0) = \tilde{J}'(0)$ which then implies $J = \tilde{J}$ by Proposition 4.3.8. Indeed, one has

$$\tilde{J}(0) = \frac{d}{ds} \exp_{c(s)}(0)\bigg|_{s=0} = \frac{d}{ds}\sigma(s)\bigg|_{s=0} = \sigma'(0) = J(0)$$

and

$$\begin{aligned} \tilde{J}'(0) &= \frac{\nabla \tilde{J}}{dt}(0) = \frac{\nabla}{\partial t}\frac{\partial c_s}{\partial s}\bigg|_{s=t=0} = \frac{\nabla}{\partial s}\frac{\partial c_s}{\partial t}\bigg|_{s=t=0} \\ &= \frac{\nabla}{\partial s}\frac{\partial}{\partial t}\exp_{\sigma(s)}(tX(s))\bigg|_{s=t=0} = \frac{\nabla}{\partial s}X(s)\bigg|_{s=0} \\ &= X'(0) = J'(0), \end{aligned}$$

where we have used Lemma 2.1.6 (i) for the third equation and the proof of Lemma 4.3.3 for the derivative of the exponential map. □

5 Topology on Spacetimes and more on Causality

One of the key differences between Lorentzian and Riemannian manifolds is, of course, the existence of a causal structure. Having established the most basic notions and results regarding

[23]See [KS21, Prop. 2.2.3].

causality in Chapter 2, we may now continue by introducing several other important concepts such as global hyperbolicity, which can generally be viewed as an analogue to completeness in Riemannian Geometry. Before that however, there is a certain class of subsets in a Lorentzian manifold that will take the center stage in our later arguments and are therefore essential to discuss. These kind of subsets, called *achronal boundaries*, are of the form $\partial I^{\pm}(S)$ for some arbitrary $S \subseteq M$. We are going to discuss achronal boundaries in great detail in the first part of this chapter and also show what kind of structure they carry. The second part then features general global topological properties of Lorentzian manifolds.

5.1 Achronal Boundaries

Achronal boundaries are sets of the form $\partial I^{+}(S)$ (resp. $\partial I^{-}(S)$) where S can be any subset of M. Looking at Minkowski space it is immediately clear that these sets are naturally interesting when null geodesics are considered (see Figure 8).

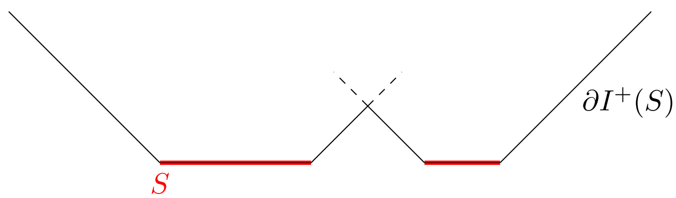


Figure 8: Minkowski space $\mathbb{R}^{1,1}$ and subset S with achronal boundary $\partial I^{+}(S)$. Each point on the boundary and not in S lies on a null geodesic.

As we will make explicit, this fact does not hold true only in Minkwoski space but for more general Lorentzian manifolds. It should be emphasized that studying achronal boundaries is not solely useful for the purposes of our later arguments but also because they naturally appear in relation to the study of null geodesics and in fact carry interesting structural properties, i.e. they turn out to be closed toplogical hypersurfaces. We will discuss these issues in this section, for which we will follow [Gal14], [O'N83] and [KS23] as the primary references.
Let us first start by elaborating on the definition of achronal sets. To this end, we define the notion of future and past sets.

Definition 5.1.1. *A subset F of M is called a future set (resp. past) if there exists some set $S \subset M$ such that $F = I^{+}(S)$ (resp. $F = I^{-}(S)$).*

Note that this definition immediately implies that F is a future set if and only if $F = I^{+}(F)$ (resp. $P = I^{-}(P)$ for past sets). As previously introduced, we define

Definition 5.1.2. *Let $F \subseteq M$ be a future or past set. Then ∂F is called achronal boundary.*

One may now appropriately conclude that every achronal boundary is indeed a set of the form $\partial I^{\pm}(S)$. Note that we will be mainly interested in the future case.
The fact that we deal with the set of boundary points of a future set has the following nice implication (see also Figure 9)

Lemma 5.1.3. *Let $S \subseteq M$ and $p \in \partial I^+(S)$. Then $I^+(p) \subseteq I^+(S)$ and $I^-(p) \subseteq M \setminus \overline{I^+(S)}$.*

Proof. We start with the first part. Let $q \in I^+(p)$. Since $p \in \partial I^+(S)$ and $p \in I^-(q)$ we must have $I^-(q) \cap I^+(S) \neq \varnothing$. Therefore $q \in I^+(S)$.
For the second part let $q \in I^-(p)$. It follows $p \in I^+(q)$ and similarly to the above that

$$I^+(q) \cap M \setminus \overline{I^+(S)} \neq \varnothing. \tag{5.1}$$

Now suppose $q \in \overline{I^+(S)}$. We distinguish two cases: If $q \in I^+(S)$ then $p \in I^+(S)$, which contradicts the fact that p is on the boundary. On the other hand if $q \in \partial I^+(S)$ we obtain by the first part $I^+(q) \subseteq I^+(S)$ which contradicts (5.1). □

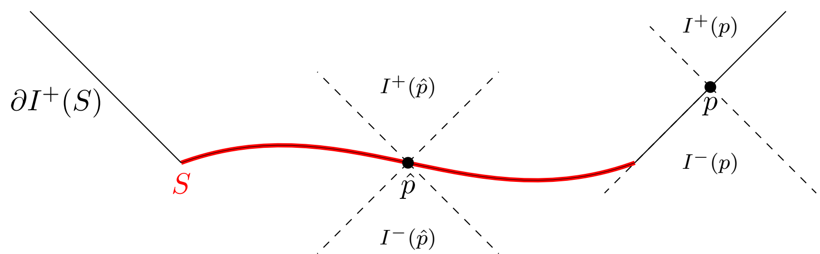


Figure 9: Visualization of Lemma 5.1.3 for two different boundary points $p, \hat{p} \in \partial I^+(S)$.

Returning to the original definition of achronal boundaries, the attentive reader will have noticed that while the term “boundary” is certainly justified, we still have to make sense of what it means to be achronal. For this we now properly define what achronal sets actually are.

Definition 5.1.4. *A subset $S \subseteq M$ is called achronal if no two points in S can be connected by a timelike curve, i.e. $I^+(S) \cap S = \varnothing$.*

It is not clear at all at this point why achronal boundaries are achronal in the above sense of the word. However, using Lemma 5.1.3, this becomes quite easy to prove.

Proposition 5.1.5. *Let $S \subseteq M$. Then $\partial I^+(S)$ is indeed achronal.*

Proof. Assume to the contrary that there exists $p, q \in \partial I^+(S)$ where $q \in I^+(p)$. Because of Lemma 5.1.3 we conclude that $q \in I^+(S)$ and because $I^+(S)$ is open, $I^+(S) \cap \partial I^+(S) = \varnothing$, which proves the claim. □

We want to investigate the structure of $\partial I^+(S)$ more closely, but first recall the following definition

Definition 5.1.6. *A second countable Hausdorff space M is an $n-$dimensional topological manifold if each point has a neighbourhood which is homeomorphic to an open set in $\mathbb{R}^n$.*
A subspace $S \subseteq M$ is a topological hypersurface if for each $p \in S$ there exists a neighbourhood $U \subseteq M$ of p together with an open set $V \in \mathbb{R}^n$ and a homeomorphism $\phi : U \to V$ such that $\phi(U \cap S) = \phi(U) \cap (\{0\} \times \mathbb{R}^{n-1})$.

In particular, note that any smooth manifold is also automatically a topological manifold. Speaking of topology, further down below we will need a highly non-trivial result, known as *Invariance of Domain*. The statement is related to subspaces of $\mathbb{R}^n$ and shows that "openness" is a topological invariant. A proof, using methods of algebraic topology, can be found in [Hat02, Thm. 2B.3].

Theorem 5.1.7 (Invariance of Domain)**.** *Let U be an open subset of $\mathbb{R}^n$ and let $\varphi : U \to \mathbb{R}^n$ be continuous and injective (i.e. a topological embedding). Then $\varphi(U)$ is open in $\mathbb{R}^n$ and $\varphi : U \to \varphi(U)$ is a homeomorphism.*

We now introduce yet another notion, related to achronal sets (see also Figure 10).

Definition 5.1.8. *Let $A \subseteq M$ be an achronal subset. The edge of A is the set of points $p \in \overline{A}$ for which for any neighbourhood U of p there exists a timelike curve γ in U from $I^-_U(p)$ to $I^+_U(p)$ with $\gamma \cap A = \varnothing$. We will denote this set by* edge A.

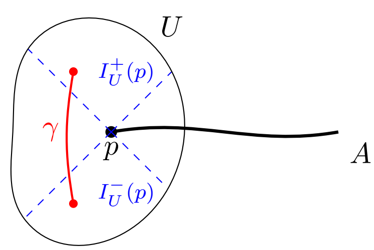


Figure 10: A point p which lies in edge(A).

As stated before, our goal is to show that $\partial I^+(S)$ carries some nice structure. To be specific, it is a closed achronal topological hypersurface of M. The next Proposition will be of central importance for this. We follow [O'N83, Ch. 14, Prop. 25] and in particular [KS23, Prop. 3.5.11].

Proposition 5.1.9. *Let $A \subseteq M$ be achronal. The following is equivalent*

(i) $A \cap \text{edge}(A) = \varnothing$

(ii) *A is a topological hypersurface*

Proof. ***(i)* $\Rightarrow$ *(ii)*** Let $p \in A$. By assumption $p \notin$ edge A and therefore, by definition of edge(A), there exists a neighbourhood U of p such that any timelike future directed curve in U from $I^-_U(p)$ to $I^+_U(p)$ intersects A in U.

Withouth loss of generality, let U be a chart domain for $\varphi : U \to \varphi(U) \subseteq R^{n+1}$ with[24] $\varphi = (x^0, \ldots, x^n)$ and such that $\frac{\partial}{\partial x^0}$ is future directed timelike on U.
By shrinking U, we obtain a neighbourhood $V \subseteq U$ of p such that for $\delta > 0$

(I) $\varphi(V) = (a - \delta, b + \delta) \times N \subseteq \mathbb{R} \times \mathbb{R}^n$,

(II) $\{x \in V \mid x^0 = a\} \subseteq I_U^-(p)$ and $\{x \in V \mid x^0 = b\} \subseteq I_U^+(p)$,

where $a, b \in \mathbb{R}$ with $a < b$ and $N \subseteq \mathbb{R}^n$ open.
Now, let $y \in N$ and define a timelike curve $\alpha : [a, b] \to V$ from $I_U^-(p)$ to $I_U^+(p)$ by $s \mapsto \varphi^{-1}(s, y)$. Clearly, α intersects A and by achronality of A this happens only once. Let $h(y) \in (a, b)$ such that $\varphi^{-1}(h(y), y) \in A$ is the point of intersection. Note that naturally there exists a unique $h(y)$ for each $y \in N$ (see Figure 11 for a depiction of the situation).

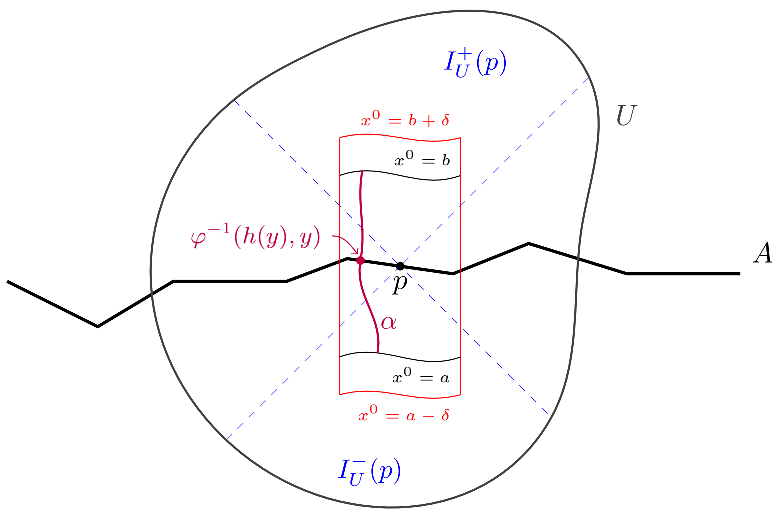


Figure 11: The situation in proof of Proposition 5.1.9, $(ii) \Leftarrow (i)$.

We want to show that $h : N \to (a, b)$ is continuous. To this end, let $\{y_m\}_{m \in \mathbb{N}}$ be a sequence in N that converges to some $y \in N$. Suppose to the contrary that $h(y_m) \not\to h(y)$ for $m \to \infty$. Since $h(y_m)$ is contained in the compact set $[a, b]$ for any $m \in \mathbb{N}$, there exists a subsequence $\{h(y_k)\}_{k \in \mathbb{N}}$ and an $r \in [a, b]$ with $h(y_k) \to r \neq h(y)$.
Let $q = \varphi^{-1}(h(y), y) \in A$. Then both, q and $\varphi^{-1}(r, y) \neq q$ are contained in the image of α. Because α is timelike, it follows that $\varphi^{-1}(r, y)$ is contained in the open set $I_V^-(q) \cup I_V^+(q)$. Using the fact that $\varphi^{-1}(h(y_m), y_m) \to \varphi^{-1}(r, y)$, there exists $m_0 \in \mathbb{N}$ with

$$\varphi^{-1}(h(y_{m_0}), y_{m_0}) \in I_V^-(q) \cup I_V^+(q).$$

But the left hand side is in A so one obtains a contradiction to achronality. Thus, h is indeed continuous.

[24] For example $\varphi = exp_p^{-1}$.

Next, note that

$$V \cap A = \varphi^{-1}(\{(h(y), y) \mid y \in N\}).$$

This means that A is the graph of h in terms of φ. We will now show that A must therefore be a topological hypersurface.
Write $\varphi = (\varphi^0, \varphi')$ and define $\psi : V \to \mathbb{R}^{n+1}$ by

$$\psi(p) = (\varphi^0(p) - h(\varphi'(p)), \varphi'(p)). \tag{5.2}$$

Note that ψ is continuous with inverse

$$\psi^{-1}(x^0, x') = \varphi^{-1}(x^0 + h(x'), x'). \tag{5.3}$$

Therefore, ψ is bijective.
In fact, $\psi \circ \varphi^{-1} : \varphi(V) \to \psi(V)$ is continuous and injective. Since $\varphi(V)$ is open in $\mathbb{R}^{n+1}$, we can apply Theorem 5.1.7 and find that $\psi \circ \varphi^{-1}(\varphi(V)) = \psi(V)$ is open and ψ is a homeomorphism. All things considered, we obtain

$$\begin{aligned}\psi(V \cap A) = \psi \circ \varphi^{-1}(\{(h(y), y) \mid y \in N\}) &\overset{(5.2)}{=} \{(0, y) \mid y \in N\} = \{0\} \times N \\ &\overset{(5.3)}{=} \psi \circ \varphi^{-1}((a - \delta, b + \delta) \times N) \cap (\{0\} \times N) \\ &= \psi(V) \cap (\{0\} \times \mathbb{R}^n)\end{aligned}$$

and therefore A is a topological hypersurface.
(i)* $\Leftarrow$ *(ii) Let $p \in A$. Since (i) is a local condition, using Proposition 2.2.12, we can assume without loss of generality that $M = \mathbb{R}^{1,n}$. Let (φ, U) be as in definition 5.1.6 with U connected, i.e. $\varphi : U \to V$ is a homeomorphism with $\varphi(U \cap A) = V \cap (\{0\} \times \mathbb{R}^n)$. Set $V_1 := \varphi(U \cap A)$ and define the restriction $\varphi_1 := \varphi|_{U \cap A} : U \cap A \to V_1$. Further, denote by π the projection map $\pi : \mathbb{R}^{1,n} \to \mathbb{R}^n$ with $(x^0, x') \mapsto x'$ (see Figure 12 for a depiction of the situation).

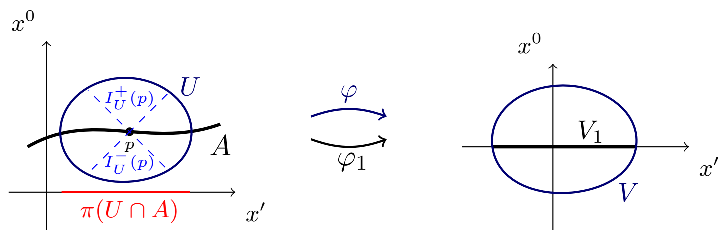


Figure 12: The situation in proof of Proposition 5.1.9, $(i) \Leftarrow (ii)$.

The vertical lines $t \mapsto (t, x')$ are timelike for any x' and intersects A at most once by achronality. Therefore, $\pi|_{U \cap A}$ is injective. Since the restriction map φ_1 is a homeomorphism one concludes that the composition $\pi \circ \varphi_1^{-1} : V_1 \to \pi(U \cap A)$ is continuous and bijective. Then, by invariance of domain (Theorem 5.1.7), we know that $\pi \circ \varphi_1^{-1}$ is a homeomorphism and $\pi(U \cap A)$ is open. Define $f : \pi(U \cap A) \to \mathbb{R}$ by $f(x') = \mathrm{pr}_0 \circ \pi^{-1}(x')$ with pr_0 being the map $x \mapsto x^0$. f is clearly continuous and furthermore

$$U \cap A = \mathrm{graph}(f) = \{(f(x'), x') \,|\, x' \in \pi(U \cap A)\}.$$

Using f, we can decompose $U \setminus A$ into two connected components

$$\begin{aligned} U^+ &= \{(x^0, x') \in U \,|\, x^0 > f(x')\}, \\ U^- &= \{(x^0, x') \in U \,|\, x^0 < f(x')\}. \end{aligned}$$

Once again using Proposition 2.2.12, we have that $I_U^\pm(p)$ are connected and because they are also open and A is achronal, both sets must lie in $U \setminus A$. Consequently, $I_U^\pm(p) \subseteq U^+ \cup U^-$. The vertical line through p meets $I_U^-(p)$ and $I_U^+(p)$, and U^- and U^+ in the right order. From this follows $I_U^\pm(p) \subseteq U^\pm$ and thus any future direct timelike curve γ from $I_U^-(p)$ to $I_U^+(p)$ is a continuous curve from U^- to U^+. Finally, by connectedness of $\mathrm{im}(\gamma)$, γ meets $\partial U^- = \partial U^+ = U \cap A \subseteq A$ and one has $p \notin \mathrm{edge}(A)$. □

Using our newly established results, we may now prove the following.

Corollary 5.1.10. *Let $S \subseteq M$. If $\partial I^+(S)$ is nonempty, it is a closed, achronal topological hypersurface in M.*

Proof. $\partial I^+(S)$ is closed by definition and achronal by Proposition 5.1.5. Consequently, what is left is to show that it is a topological hypersurface.
From Lemma 5.1.3 follows that any timelike curve from $I^-(p)$ to $I^+(p)$ will meet $\partial I^+(S)$ for any $p \in \partial I^+(S)$. Thus $\mathrm{edge}(\partial I^+(S)) = \varnothing$ and the previous result gives the claim. □

Remark 5.1.11. The above Corollary has the natural consequence that there generally is not too much hope for better regularity in Proposition 5.1.9. For example, consider the null cone C_p at a point p in Minkowski space $\mathbb{R}^{1,d}$. Note that here we have exactly $C_p = \partial I^+(p)$. Clearly, the vertex at p makes it impossible for C_p to admit a differentiable structure. However, the light cone is a Lipschitz manifold[25], so in general one could at least have differentiability almost everywhere.

Finally, we want to familiarize ourselves a bit more with the nature of achronal boundaries and in particular study their relation with null geodesics. Recall that in the introduction we remarked that in Minkowski space achronal boundaries are governed by the behaviour of null geodesics. We are now going to demonstrate that this also holds true in the general case. In particular, we want to show that any point on an achronal boundary lies on some null

[25]For a Lipschitz manifold, the transition functions are Lipschitz continuous.

geodesic. Note that the proof of the precise statement down below will require an application of the Limit Curve Lemma. We follow the argument in [Gal14, Prop. 3.4].

Proposition 5.1.12. *Let $S \subseteq M$ be closed. Then every $p \in \partial I^+(S) \setminus S$ lies on a null geodesic which is contained in $\partial I^+(S)$ and either is past inextendible or has a past end point in S.*

Proof. Endow (M, g) with a complete Riemannian metric h, which exists by Theorem 2.1.9, and let $p \in \partial I^+(S) \setminus S$. Since p is a boundary point, there exists a sequence $\{p_n\}_{n\in\mathbb{N}} \subseteq I^+(S)$ with $p_n \to p$ for $n \to \infty$. For each $n \in \mathbb{N}$ we now define a past directed timelike curve $\gamma_n : [0, a_n] \to M$ from p_n to some $q_n \in S$ such that γ_n is parametrized with respect to h-arclength.
We can now simply extend each γ_n to gain past inextendible curves $\overline{\gamma}_n : [0, \infty) \to M$, which we still parametrize with respect to h-arclength. Since p is an accumulation point of $\overline{\gamma}_n(0)$, we are now able to apply (the time reversed version) of the Limit Curve Lemma 4.2.7. Consequently, we obtain a past inextendible continuous causal curve $\gamma : [0, \infty) \to M$ with $\gamma(0) = p$ and to which the subsequence $\{\overline{\gamma}_m\}_{m\in\mathbb{N}}$ converges uniformly on compact sets (see Figure 13).

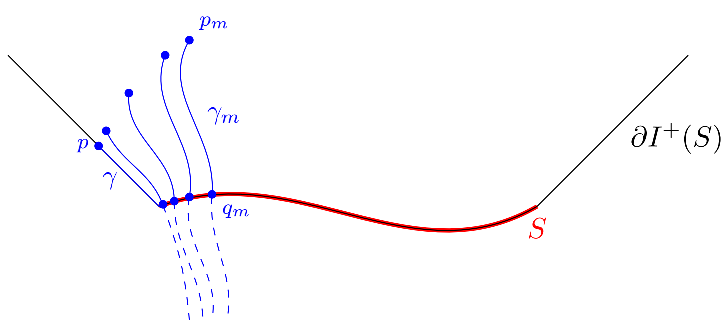


Figure 13: Setup for Limit Curve Lemma in proof of 5.1.12.

Choosing a further subsequence (if necessary), we have $a_m \nearrow a$ where $a \in (0, \infty) \cup \{\infty\}$. We want to show that either $\gamma|_{[0,a]}$ or $\gamma|_{[0,\infty)}$ is a null geodesic that satisfies the claim.
Let $t \in (0, a)$. By our assumption $a_m > t$ for large enough m. Then $\overline{\gamma}_m(t) = \gamma_m(t) \in I^+(S)$ (Note that this is also anyway satisfied for a subsequence with $a_m \searrow a$). Since $\gamma(t) = \lim_{m\to\infty} \gamma_m(t)$, we can conclude that $\gamma(t) \in \overline{I^+(S)}$.
If $\gamma(t)$ were in $I^+(S)$, then there exists $x \in S$ such that $x \ll \gamma(t) < p$. By push-up, it follows that $x \ll p$, i.e. $p \in I^+(S)$, which contradicts the fact that p is on the boundary. Therefore $\gamma(t) \in \partial I^+(S)$ for any $t \in (0, a)$ and $\gamma|_{[0,a)} \subset \partial I^+(S)$. By achronality there exists no timelike curve between two points in $\partial I^+(S)$. Thus, by Corollary 4.3.6, $\gamma|_{[0,a)}$ must be a null pregeodesic.
Finally, we consider the cases $0 < a < \infty$ and $a = \infty$. By uniform convergence and continuity we have for the first case that $\gamma(a) = \lim_{m\to\infty} \gamma_m(a_m) = \lim_{m\to\infty} q_m \in S$ as S is closed. Therefore, under an suitable parametrization, p lies on null geodesic contained in $\partial I^+(S)$ with end point on S while for $a = \infty$ the null geodesic is contained in $\partial I^+(S)$ and is past

inextendible. □

5.2 Global Hyperbolicity, Cauchy Surfaces and Domain of Dependence

After the discussion of achronal boundaries we turn to introduce new notions related to the causal structure on a spacetime. In fact, our goal is to introduce several definitions that enable us to distinguish different classes of spacetimes and to get a better view of their global structure. For this, we keep to [Gal14], [Wal84], [KS23], [O'N83] and [Ger70] as the primary references.
It is useful to impose additional constraints on the class of spacetimes (M, g) that we want to deal with. This is primarily motivated by physical considerations. For instance, any time oriented, compact Lorentzian manifold admits closed timelike curves (see e.g. [Gal14, Prop. 4.1] or [BEE96, Prop. 3.10]) and therefore, one can reasonably rule out such cases on physical grounds. In conclusion, we want to focus on studying spacetimes with reasonable causality restrictions. We make the following definitions

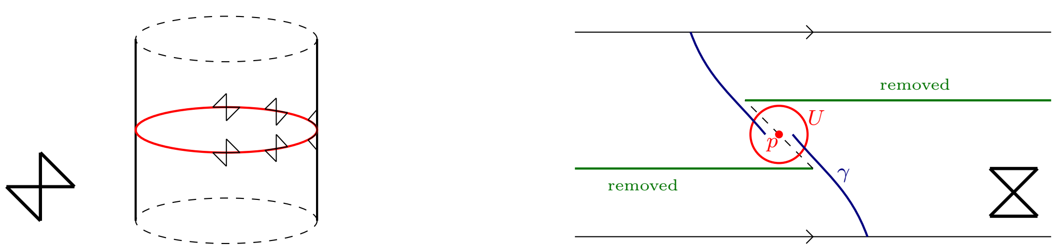
Definition 5.2.1. *A spacetime (M, g) is called*

(i) chronological, if there exist no closed timelike curves,

(ii) causal, if there exist no closed causal curves,

(iii) strongly causal, if for any $p \in M$ and for all neighbourhoods U of p there exists a neigbourhood $V \subseteq U$ of p such that any causal curve starting and ending in V is completely contained in U.

Remark 5.2.2. It is clear that we have $(iii) \Rightarrow (ii) \Rightarrow (i)$. However, observe that the opposite implications do not hold, as one may for example see in Figure 14(a). There, all great circles (red line) are closed null curves, but these are also the only causal ones, i.e. $(i) \not\Rightarrow (ii)$.
As seen in Figure 14(b), the strong causality condition prohibits not only closed causal curves but also ones that are almost closed. There, for any neighbourhood V around p one can find γ such that the definition is violated.



(a) Cylinder with tilted light cones is chronological but not causal.

(b) Violation of strong causality at $p \in M$. The lines with arrows are to be identified.

Figure 14: Examples of Remark 5.2.2.

In the literature one also encounters the following definition of strong causality: (M, g) is called strongly causal at $p \in M$ if for any neighbourhood U of p there exists another neighbourhood $V \subseteq U$ of p whose intersection with any timelike curve is connected or equivalently $x, y \in V$ with $x \ll z \ll y$ implies $z \in V$. (M, g) is strongly causal if it is strongly causal at each $p \in M$. Note however, that following from [MS08, Lemma 3.21], this notion of strong causality is in fact equivalent to the prior definition 5.2.1 (iii).

Studying strongly causal spacetimes is not only useful to avoid certain pathological situations but it also comes with some additional benefits. For instance, in the next Lemma we will show that inextendible causal curves cannot be imprisoned in strongly causal compact subsets. We follow [O'N83, Lemma 14.13].

Lemma 5.2.3. *Let K be a compact subset of M such that strong causality holds on K. If $\gamma : [0, b) \to M$ is a future inextendible causal curve starting in K, then there exists $t_0 \in [0, b)$ such that $\gamma([t_0, b)) \subseteq M \setminus K$, i.e. γ leaves K and does not return.*

Proof. Assume to the contrary that γ remains in K or keeps returning to it. Let $\{t_i\}_{i\in\mathbb{N}}$ be a sequence in $[0, b)$ with $t_i \to b$ such that $\{\gamma(t_i)\}_{i\in\mathbb{N}}$ is contained in K. By compactness there exists a subsequence $\{\gamma(t_j)\}_{j\in\mathbb{N}}$ which converges to a point $p \in K$. However, since γ is inextendible, there must be another sequence $\{\tilde{t}_j\}_{j\in\mathbb{N}}$ converging to b such that $\gamma(\tilde{t}_j) \not\to p$. Taking a further subsequence if necessary, there is a neighbourhood U of p which contains no $\gamma(\tilde{t}_k)$. Without loss of generality let $t_1 < \tilde{t}_1 < \cdots < t_k < \tilde{t}_k < \ldots$. It now follows that $\gamma_{[t_k,t_{k+1}]}$ contradicts strong causality at p for some k: For any neighbourhood V of p there is a k such that $\gamma(t_k), \gamma(t_{k+1}) \in V$ while at the same time $\gamma_{[t_k,t_{k+1}]}$ leaves U. $\square$

Considering a general Riemannian manifold, one of the most benign properties one could have is completeness. While we mentioned at the beginning that the situation surrounding geodesic completeness for Lorentzian manifolds is more delicate, there still is a Lorentzian analogue to complete Riemannian manifolds. This property is called global hyperbolicity and is defined as follows

Definition 5.2.4. *$U \subseteq M$ is globally hyperbolic if the following two conditions hold:*

(i) U is causal.

(ii) The sets $J^-(q) \cap J^+(p)$ are compact for all $p, q \in U$.

Usually the first condition is replaced by strong causality. However, this actually turns out to be equivalent once we have (ii). More precisely, Bernal and Sanchez were able to prove the following[26], [BS07].

[26] In fact, much more recently (cf. [HM19, Theorem 2.8]), it was shown that for non-compact spacetimes with dimension ≥ 3, the first assumption can be dropped completely.

Proposition 5.2.5. *If condition (ii) in definition 5.2.4 holds, then $U \subseteq M$ is causal if and only if it is strongly causal.*

We emphasize that it is not obvious at all from the definition that global hyperbolicity is an analogue of geodesic completeness for Riemannian manifolds. But as a matter of fact, it is possible to show the following (see e.g. [KS23, Thm. 3.6.7])

Theorem 5.2.6 (Avez-Seifert)**.** *Let M be globally hyperbolic and let $p, q \in M$ with $p < q$. Then there exists a causal geodesic between p and q that maximizes the Lorentzian length.*

One should emphasize here the striking difference between Lorentzian and Riemannian manifolds, where we have minimizing geodesics instead.
Additionally, one may also show that the Lorentzian time-separation function $\tau : M \times M \to \mathbb{R}$ is continuous for globally hyperbolic spacetimes instead of being only lower-semi continuous in the general case (see e.g. [O'N83, Ch. 14, Lemma 22]), but this will not be of much importance to us.
However, there is an essential property of globally hyperbolic spacetimes that we do in fact need. Recall that the causal relation $\leq$ is closed on convex subsets (cf. 2.2.18 (*iii*)). This holds true more generally.

Proposition 5.2.7. *Let M be globally hyperbolic. Then $\leq$ is closed on M.*

Proof. Let $p \in M$. We first show that $J^+(p)$ is closed. For this, let $q \in \overline{J^+(p)} \setminus J^+(p)$ and consider a sequence $\{q_n\}_{n\in\mathbb{N}}$ in $J^+(p)$ with $q_n \to q$. Now choose $r \in I^+(q)$. By global hyperbolicity follows that the causal diamond $J^+(p) \cap J^-(r)$ is compact and thus closed. Since $I^-(r)$ is an open neighbourhood of q, one further has $q_n \in J^-(r)$ for large n. Consequently

$$q = \lim_{k\to\infty} q_k \in J^+(p) \cap J^-(r) \subseteq J^+(p)$$

and we obtain a contradiction.
Finally, we need to demonstrate that $\leq$ is closed. To this end, define sequences $\{x_n\}_{n\in\mathbb{N}}$ and $\{y_n\}_{n\in\mathbb{N}}$ in M with $x_n \leq y_n$ and $x_n \to x$ and $y_n \to y$. Pick $a \in I^+(y)$. Since $I^-(a)$ is a neighbourhood of y, we have that $y_n \in J^-(a)$ for large n and therefore also $x_n \in J^-(a)$ for large n. Because $J^-(a)$ is closed, this means that also $x \in J^-(a)$ and in fact $a \in J^+(x)$. Defining yet another sequence $\{a_n\}_{n\in\mathbb{N}}$ in $I^+(y)$ with $a_n \to q$, we have that $a_n \in J^+(x)$ for each $n \in \mathbb{N}$. It follows that $y \in J^+(x)$ and we conclude $p \leq q$ as claimed. □

Using the prior result, one may now establish two very important topological properties of causal future and past sets that turn out to be crucial in later arguments. We follow [KS23, Lemma 3.10.8 & Lemma 3.11.3].

Proposition 5.2.8. *Let M be globally hyperbolic. Then*

(i) if $K \subseteq M$ is compact, $J^\pm(K)$ is closed,

(ii) if $K_1, K_2 \subseteq M$ are compact, $J^+(K_1) \cap J^-(K_2)$ is compact.

Proof. (i) Let $\{p_n\}_{n\in\mathbb{N}}$ be a sequence in $J^+(K)$ with $p_n \to p \in M$. By assumption, for each $n \in \mathbb{N}$ there exist $q_n \in K$ such that $q_n \leq p_n$. Since K is compact, there is a subsequence $\{q_k\}_{k\in\mathbb{N}}$ that converges to $q \in K$. Proposition 5.2.7 then infers that $p \in J^+(q) \subseteq J^+(K)$ and thus, $p \in J^+(K)$. The case for $J^-(K)$ follows analogously.
(ii) Let $\{p_n\}_{n\in\mathbb{N}}$ be a sequence in $J^+(K_1) \cap J^-(K_2)$. For each $n \in \mathbb{N}$ there exist $q_n \in K_1$ and $r_n \in K_2$ such that $q_n \leq p_n \leq r_n$. By compactness of K_1 and K_2, we may assume without loss of generality that $q_n \to q \in K_1$ and $r_n \to r \in K_2$. Now let $x, y \in M$ such that $x \ll q$ and $r \ll y$ (see also Figure 15).

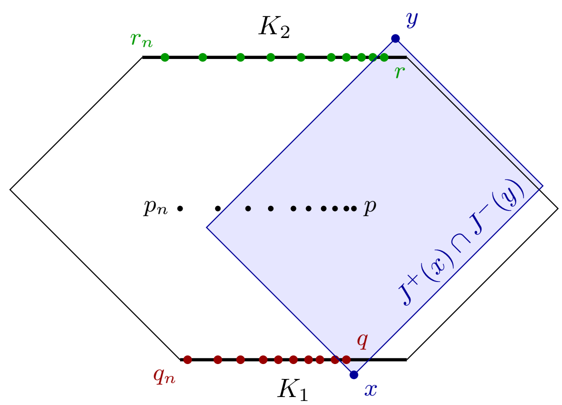


Figure 15: The setup for the proof of Proposition 5.2.8 (ii).

For large n one therefore has $x \ll p_n \ll y$ and thus $p_n \in J^+(x) \cap J^-(y)$. By compactness there exists a subsequence $\{p_k\}_{k\in\mathbb{N}}$ that converges to p. Again, using Proposition 5.2.7, one has $q \leq p \leq r$ and the claim follows. $\square$

The notion of global hyperbolicity is closely related to the existence of so called Cauchy surfaces, which are defined as follows

Definition 5.2.9. *An achronal subset S of M is called a Cauchy surface if it is intersected by any inextendible causal curve.*

From a physical perspective Cauchy surfaces arise very naturally which we will discuss later on in Chapter 8.2. Note also that the definition immediately implies that $M = I^-(S) \cup S \cup I^+(S)$. Now let $p \in \partial I^+(S)$, since $S \subseteq \partial I^+(S)$, if $p \notin S$ then $p \in I^-(S)$. Because $I^-(S)$ is open there exists $q \in I^-(S) \cup I^+(S)$, which yields a contradiction. Thus $\partial I^+(S) = S$ and similarly $\partial I^-(S) = S$. In conclusion, after applying Corollary 5.1.10, we have proved the following

Proposition 5.2.10. *If S is a Cauchy surface of M, then S is a closed, achronal topological hypersurface.*

The connection to global hyperbolicity follows from the next result due to Geroch. For a proof, we refer to [Ger70].

Theorem 5.2.11 (Geroch)**.** *M is globally hyperbolic if and only if it admits a Cauchy surface S.*

One needs to emphasize here that Cauchy surfaces do not only carry an interesting structure and yield an elegant way to define the notion of global hyperbolicity but that most remarkably, they encode the nontrivial part of the global topology as seen in the subsequent Theorem. We follow [O'N83, Prop. 14.31 & Cor. 14.32] and [KS23, Thm. 3.5.24 & Prop. 3.11.2].

Theorem 5.2.12. *Let M be globally hyperbolic. Then*

(i) if S_1 and S_2 are two Cauchy surfaces of M, then they are homeomorphic,

(ii) if S is a Cauchy surface of M, then S is connected and M is homeomorphic to $\mathbb{R} \times S$.

Proof. Consider a Cauchy surface S and let $T \in \mathfrak{X}(M)$ be timelike. We will first show that there exists a well-defined projection map $\rho^T : M \to S$ such that ρ^T is a continuous open map with $\rho^T = \mathrm{id}_S$. For this define $\rho(p)$ to be the unique point of intersection between S and the maximal integral curve of T on which lies p (cf. Figure 16). Because maximal integral curves are inextendible[27] this map is well-defined.

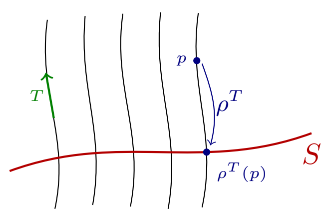


Figure 16: The flow lines of T in black. $\rho^T(p)$ is defined as the unique intersection of the Cauchy surface S and the maximal integral curve of T that runs through p.

Let $\mathrm{Fl}^T : U \subseteq \mathbb{R} \times M \to M$ be the flow of T, where U is the maximal domain. Following from 5.2.10, $\mathbb{R} \times S$ is a topological hypersurface in $\mathbb{R} \times M$. Define $U_S := (\mathbb{R} \times S) \cap U$. Note that since U is open[28], U_S is itself a topological hypersurface in $\mathbb{R} \times M$. We want to show that the restriction $\mathrm{Fl}^T\big|_{U_S}$ is continuous and bijective. Continuity is immediate, so let $p \in M$. By definition of S, the curve $t \mapsto \mathrm{Fl}^T_t(p)$ intersects S and therefore there exists $t_0 \in \mathbb{R}$ such that $\mathrm{Fl}^T_{t_0}(p) = q \in S$. By the properties of the flow map, one obtains

$$p = \left(\mathrm{Fl}^T_{t_0}\right)^{-1}(q) = \mathrm{Fl}^T_{-t_0}(q) = \mathrm{Fl}^T(-t_0, q) = \mathrm{Fl}^T\big|_{U_S}(-t_0, q),$$

[27] See [Kun22a, Thm. 2.3.3] (*i*).
[28] See [Kun22a, Thm. 2.3.3] (*iii*).

which proves surjectivity. For injectivity assume $\mathrm{Fl}^T\big|_{U_S}(t_1, q_1) = \mathrm{Fl}^T\big|_{U_S}(t_2, q_2)$ with (t_1, q_1), $(t_2, q_2) \in U_S$. It follows that $\mathrm{Fl}^T_{t_1}(q_1) = \mathrm{Fl}^T_{t_2}(q_2)$ and, by again using the properties of the flow map, we find that

$$q_2 = \left(\mathrm{Fl}^T_{t_2}\right)^{-1} \circ \mathrm{Fl}^T_{t_1}(q_1) = \mathrm{Fl}^T_{-t_2} \circ \mathrm{Fl}^T_{t_1}(q_1) = \mathrm{Fl}^T_{t_1 - t_2}(q_1).$$

Thus, the same integral curve meets S twice in q_1 and q_2. Because S is achronal, this can only be true if $q_1 = q_2$. But then $\mathrm{Fl}^T_{t_1}(q_1) = \mathrm{Fl}^T_{t_2}(q_1)$ and uniqueness of integral curves implies $t_1 = t_2$ and injectivity follows.
Note that because M itself is also a topological manifold, we can evoke the invariance of domain (Thm. 5.1.7) to conclude that $\mathrm{Fl}^T\big|_{U_S}$ is a homeomorphism and therefore open. If $\pi : \mathbb{R} \times S \to S$ is the natural projection operator, i.e. $\pi(t, p) = p$, one may now define $\rho^T := \pi \circ \left(\mathrm{Fl}^T\big|_{U_S}\right)^{-1}$. Since π is continuous and open, ρ^T is continuous and open as well. Further note that for $p \in S$ we can write $p = \mathrm{Fl}^T_0(p)$ and thus

$$\rho^T(p) = \pi(0, p) = p,$$

which shows $\rho^T\big|_S = \mathrm{id}_S$.
We may now prove statements (i) and (ii) as follows:

(i) Let $T \in \mathfrak{X}(M)$ be timelike. Note that we can define projection operators $\rho^T_{S_1}, \rho^T_{S_2}$ for S_1 and S_2 respectively. The claim follows from the fact that $\rho^T_{S_1}\big|_{S_2}$ is the inverse of $\rho^T_{S_2}\big|_{S_1}$ and vice versa.

(ii) Connectedness easily follows by picking any projection operator ρ and realizing $\rho(M) = S$. For the other part, let h be a complete Riemannian metric on M and $T_1 \in \mathfrak{X}(M)$ timelike. Normalizing T_1 with respect to h, we obtain another timelike vector field $T = \frac{T_1}{\|T_1\|_h}$. Let $\gamma : (t_-, t_+) \to M$ be an integral curve of T, where $(t_-, t_+) \subseteq \mathbb{R}$ is the maximal domain of γ. Suppose $t_+ < \infty$. It follows that

$$L_h(\gamma|_{[0,t_+)}) = \int_0^{t_+} \|\dot{\gamma}(t)\|_h \, dt = t_+ < \infty.$$

But because h is complete, the above implies that $\gamma|_{[0,t_+)}$ remains in a compact set [29] which yields a contradiction to the flow of T (cf. [Kun22a, Thm. 2.3.3]). Therefore T is complete, i.e. $\mathrm{Fl}^T : \mathbb{R} \times M \to M$. Following from the previous discussion, the map $(t, p) \in \mathbb{R} \times S \mapsto \mathrm{Fl}^T_t(p)$ is a homeomorphism onto M.

□

In regards to the splitting, Bernal and Sanchez were able to prove an even stronger statement (see [BS05] or [KS23, Thm. 3.11.20]), namely that $(M, g) \simeq (\mathbb{R} \times S, -\beta \mathrm{d}t^2 + g_t)$ where

[29] See Hopf-Rinow Theorem 2.1.8.

$\beta : \mathbb{R} \times S \to (0, \infty)$ is smooth and g_t is a smooth family of Riemannian metric on S. Intuitively, we may therefore think of globally hyperbolic spacetimes as Lorentzian manifolds that admit a global flow of time.

Up next we want to define a very fundamental notion called the domain of dependence.

Definition 5.2.13. *Let S be an achronal subset of M. The future domain of dependence of S, is given defined as*

$$D^+(S) = \{p \in M \,|\, \text{Any past inextendible causal curve from } p \text{ intersects } S\}.$$

Similarly, we define the past domain of dependence of S by

$$D^-(S) = \{p \in M \,|\, \text{Any future inextendible causal curve from } p \text{ intersects } S\}.$$

The domain of dependence of S then is

$$D(S) = D^+(S) \cup D^-(S).$$

See Figure 17 for examples that illustrate the above definition. From a physical perspective, $D^+(S)$ is precisely the set of points that are causally determined by S. Thereby making it a concept of central importance in mathematical relativity[30].

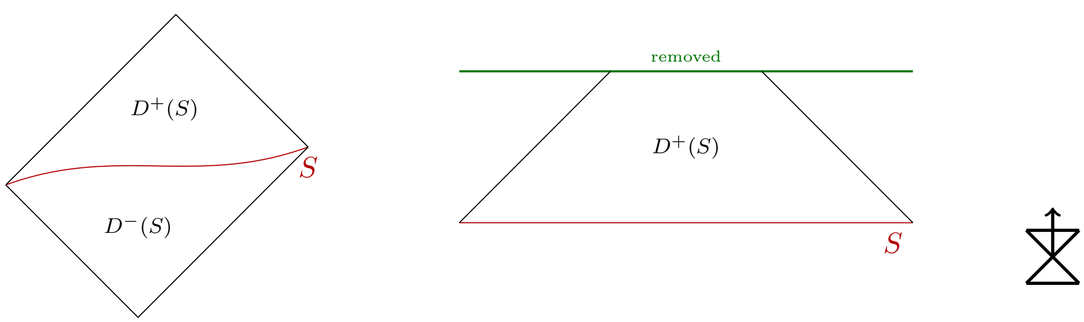


Figure 17: Examples for definition 5.2.13

Let us now collect some properties that somehow justify the above picture and which are also useful for later. We follow [KS23, Rem. 3.8.2].

Proposition 5.2.14. *Let S be an achronal subset of M. Then*

(i) $S \subseteq D^\pm(S) \subseteq I^\pm(S) \cup S \subseteq J^\pm(S)$,

(ii) $D^\pm(S) \cap I^\mp(S) = \varnothing$,

(iii) $D^+(S) \cap D^-(S) = S$,

[30] More on this in section 8.2.

(iv) $D(S) \cap I^{\pm}(S) = D^{\pm}(S) \setminus S$.

Proof. (i) is clear. For (ii) let $p \in D^+(S) \cap I^-(S)$. Then there exists a past directed timelike curve γ from S to p. Also, because $p \in D^+(S)$, any inextendible, past directed timelike curve $\tilde{\gamma}$ meets S. But then $\gamma \cup \tilde{\gamma}$ is a timelike curve which meets S twice and therefore contradicts achronality. The case with $p \in D^-(S) \cap I^+(S)$ follows in the same way. To prove (iii) just note by (i) and (ii) that

$$S \subseteq D^+(S) \cap D^-(S) \subseteq D^+(S) \cap \big(S \cup I^-(S)\big) = D^+(S) \cap S = S.$$

Finally, note that again by (i) and (ii) it follows $D(S) \cap I^{\pm}(S) = D^{\pm}(S) \cap I^{\pm}(S) = D^{\pm}(S) \setminus S$, which shows (iv). □

The domain of dependence fits quite naturally into our previously defined concepts of global hyperbolicity and Cauchy surfaces. Assume for example that one has an achronal subset $S \subseteq M$. If S is a Cauchy surface one has $M = I^+(S) \cup S \cup I^-(S)$ and because S is achronal, it immediately follows that $D^{\pm}(S) = I^{\pm}(S) \cup S$. Consequently $D(S) = D^+(S) \cup D^-(S) = M$. On the other hand, let us now suppose $D(S) = M$. Take $p \in M$ and let γ be any inextendible causal curve through p. Without loss of generality $p \in D^+(S)$ ($p \in D^-(S)$ is analogous). Since γ is in particular past inextendible, it has to meet S at some point. As γ and p were arbitrary, one may conclude that S is a Cauchy surface of M. So without much effort, we were able to prove the following fact

Proposition 5.2.15. *Let S be an achronal subset of M. Then S is a Cauchy surface of M if and only if $D(S) = M$.*

A nice consequence of this and Theorem 5.2.11 is that if $D(S)$ is open, it must also be globally hyperbolic.

As we will see in Chapter 8.2, the domain of dependence is also closely related to what is called the maximal Cauchy development for the Einstein equations. From an hyperbolic PDE viewpoint, one is therefore certainly curious about the future (or past) boundary of this set. This is simply the case because questions about extendibility are naturally posed and answered there. But even geometrically speaking it seems reasonable to ponder on what happens there. Indeed, it turns out that this boundary shares some similarities with achronal boundaries, which were introduced in the previous section. For now, let us make a precise definition of what we mean here by *boundary.*

Definition 5.2.16. *Let S be an achronal subset of M. The future Cauchy horizon of S is the set*

$$H^+(S) = \{p \in \overline{D^+(S)} \,|\, I^+(p) \cap D^+(S) = \varnothing\} = \overline{D^+(S)} \setminus I^-(D^+(S)).$$

Similarly, the past Cauchy horizon is

$$H^-(S) = \{p \in \overline{D^-(S)} \mid I^-(p) \cap D^-(S) = \varnothing\} = \overline{D^-(S)} \setminus I^+(D^-(S)).$$

The total Cauchy horizon then is

$$H(S) = H^+(S) \cup H^-(S).$$

See Figure 18 for a visualization of the above definition.

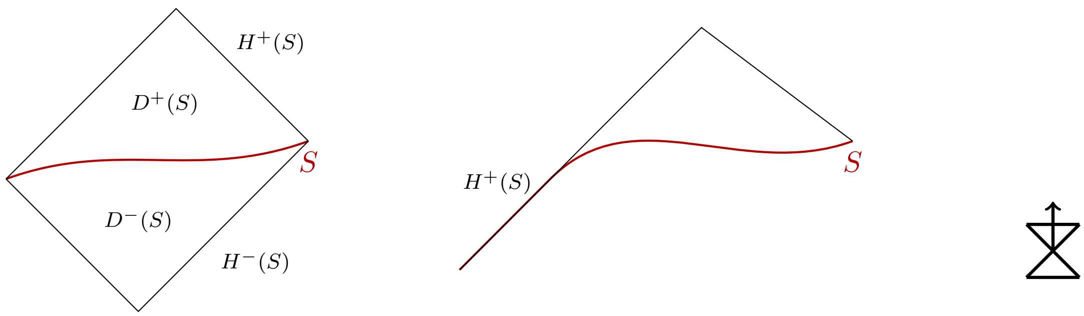


Figure 18: Examples for definition 5.2.16

We may now be able to make the following addition to the previous Proposition.

Proposition 5.2.17. *Let S be an achronal subset of M. Then S is a Cauchy surface of M if and only if $H(S) = \varnothing$.*

Proof. We want to show $\partial D(S) = H(S)$. Then if S is a Cauchy surface, by 5.2.15 $D(S) = M$ and naturally $\partial D(S) = H(S) = \varnothing$. Conversely, if $H(S) = \varnothing$ then $\partial D(S) = \varnothing$. Consequently $D(S)$ is closed and open, and because M is connected we have $D(S) = M$.
Let $p \in \partial D(S)$ and assume that $p \notin H(S)$. Therefore, there must exist $q_\pm \in I^\pm(p) \cap D^\pm(S)$. Then $I^-(q_+)$ and $I^+(q_-)$ are two nonempty open neighbourhoods of p and so is $I^-(q_+) \cap I^+(q_-)$. Let $r \in I^-(q_+) \cap I^+(q_-)$ and let $\gamma_\pm$ be two timelike curves from r to q_+ and q_- to r respectively. Because $q_\pm \in D^\pm(S) \subseteq I^\pm(S) \cup S$, and by achronality of S, the curve $\gamma_- \cup \gamma_+$ has to meet S exactly once and thus also either $q_+ \notin S$ or $q_- \notin S$ (or both). Suppose $q_+ \notin S$. We have three cases (see Figure 19): (i) solely the segment γ_+ meets S, (ii) solely the segment γ_- meets S or (iii) both segments meet S at r.

For (i) we have $r \in I^-(S)$ and therefore $r \in I^-(S) \cap I^+(q_-)$. Hence $r \in D^-(S) \subseteq D(S)$. For (ii) we can conclude that $r \in D^+(S) \subseteq D(S)$ since any past inextendible causal curve has to meet S as we can see by concetenating with γ_+ to obtain a causal curve from $q_+ \in D^+(S)$. (iii) simply implies $r \in S \subseteq D(S)$. Note that the case where $q_+ \in S$ but $q_- \notin S$ is analogous. Altogether, we always have $r \in D(S)$. Therefore, we have shown that any point in the open neighbourhood $I^-(q_+) \cap I^+(q_-)$ of $p \in \partial D(S)$ is contained in $D(S)$, which yields a

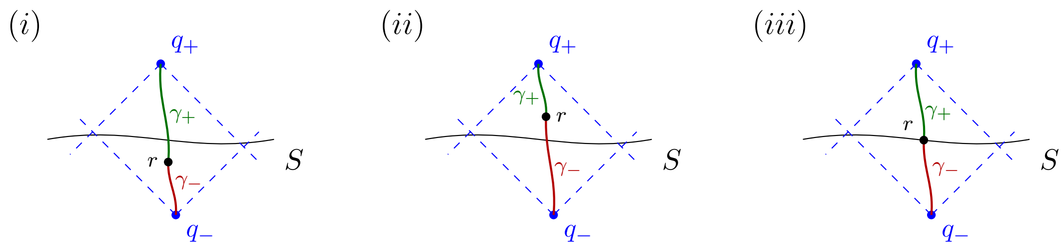


Figure 19: The three cases $(i), (ii), (iii)$ from left to right for $q_- \notin S$.

contradiction. Consequently, one has $p \in H(S)$.
Now let $p \in H(S)$. We need to show $p \in \partial D(S)$. For this, suppose $p \in H^+(S)$ and let $U \subset M$ be any open neighbourhood of p. By definition, it follows that $p \in \overline{D^+(S)}$. Note that we then must have $D(S) \cap U \neq \varnothing$. On the other hand, let $\gamma : [0, a) \to M$ be any future directed timelike curve starting from $\gamma(0) = p$. By continuity, the set $J := \gamma^{-1}(U)$ is open and in particular non-empty since $0 \in J$. Therefore, there exist $t \in J \subseteq [0, a)$ such that $\gamma(t) \in U$ and we can conclude $U \cap I^+(p) \neq \varnothing$. Let $q \in U \cap I^+(p)$. Because $p \in H^+(S)$ it follows that $q \notin D^+(S)$. We show next that $q \notin D^-(S)$ as well. By Proposition 5.2.14 (i) and (iii) it remains to show that $q \notin I^-(S)$. But this follows simply because $q \in I^+(S)$ and S is achronal. Thus $q \notin D^\pm(S)$ and $q \in M \setminus D(S)$. We have shown that each open neighbourhood U of $p \in H^+(S) \subseteq \overline{D(S)}$ contains points in $D(S)$ and $M \setminus D(S)$ and therefore $p \in \partial D(S)$. The case for $p \in H^-(S)$ follows analogously. □

We now want to collect useful properties of $H^\pm(S)$. However, before that, we will have to prove a very convenient result. We follow [O'N83, Ch. 14, Lemma 30] and [KS23, Lemma 3.5.21 (i)].

Lemma 5.2.18. *Let $C \subseteq M$ be closed and γ a past directed and inextendible causal curve emanating from p such that $\gamma \cap C = \varnothing$. Then for any $q \in I^+_{M \setminus C}(p)$ there exists a past directed timelike curve starting in q which is past inextendible and does not meet C.*

Proof. Without loss of generality let $\gamma : [0, \infty) \to M \setminus C$ with $\gamma(0) = p$. Note that $\{\gamma(n)\}_{n \in \mathbb{N}}$ does not converge. Further, let h be an auxiliary, complete Riemannian metric h on M.
Set $p_0 := q \gg p$. We have $\gamma(1) \leq \gamma(0) \ll p_0$ and therefore by push-up $\gamma(1) \ll p_0$ and consequently there exists a timelike curve γ_1 from p_0 to $\gamma(1)$.
Next, we pick a point p_1 on γ_1 with $0 < d_h(p_1, \gamma(1)) < 1$ and note that $\gamma(2) \ll p_1$. Again, there exists a past directed timelike curve γ_1 from p_1 to $\gamma(2)$ (see Figure 20 for a depiction of the setup). Iterating this procedure yields $\gamma(k) \ll p_k \ll p_{k-1}$ and $d(\gamma(k), p_k) < \frac{1}{k}$ for any $k \in \mathbb{N}$.

In this way, we obtain a past directed timelike curve $\tilde{\gamma}$, which starts in p_0 and such that all p_k lie on $\tilde{\gamma}$. We want to show that $\tilde{\gamma}$ is past inextendible.
Assume for the sake of contradiction that $\tilde{\gamma}$ can be extended continuously to some endpoint

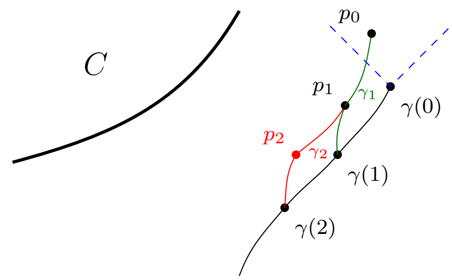


Figure 20: Setup in proof for Lemma 5.2.18.

x. Then

$$d(\gamma(k), x) \leq d(\gamma(k), p_k) + d(p_k, x) \to 0$$

since the second term on the right hand side goes to zero as $p_k \to x$, while the first is bounded by $\frac{1}{k}$ and therefore also vanishes in the limit. Consequently, $\gamma(k) \to x$ as $k \to \infty$, which contradicts the past inextendibility of γ. $\square$

Drawing from [KS23, Lemma 3.8.14], one may now establish several properties of Cauchy horizons.

Proposition 5.2.19. *Let S be an achronal subset of M. We have*

(i) *$H^{\pm}(S)$ is closed and achronal.*

(ii) *If S is closed, then $p \in \overline{D^+(S)}$ if and only if any past directed inextendible timelike curve through p meets S.*

(iii) *If S is closed, then*

$$\partial D^{\pm}(S) = S \cup H^{\pm}(S).$$

Proof. (i) By Definition 5.2.16, the Cauchy horizons are defined as the complement of an open set in a closed set and must thereby be closed.
Next we show achronality for $H^+(S)$. Note first that, also by definition, $I^+(H^+(S)) \cap D^+(S) = \varnothing$. Openness of $I^+(H^+(S))$ then implies $I^+(H^+(S)) \cap \overline{D^+(S)} = \varnothing$. Since $H^+(S)$ is contained in $\overline{D^+(S)}$ it follows that $I^+(H^+(S)) \cap H^+(S) = \varnothing$ and we obtain the result. The case for $H^-(S)$ is analogous.
(ii) " $\Rightarrow$ " Assume to the contrary that there exists a $p \in \overline{D^+(S)}$ and a past directed inextendible timelike curve $\gamma : [0, b) \to M$ with $\gamma(0) = p$ that does not meet S. By closedness of S there exists a convex neighbourhood U of p with $S \cap U = \varnothing$. Let $\epsilon > 0$ such that $q \coloneqq \gamma(\epsilon)$ with $p \in I^+_U(q)$. Choose an $r \in I^+_U(q) \cap D^+(S)$, which exists as $I^+_U(q)$ is an open neighbourhood of p. Let c be the radial geodesic from r to q in U (see Figure 21). Using Proposition 2.2.12 one concludes that c is timelike and past directed.

One has $c \cap S = \varnothing$, but at the same time, the concatenation $c \cup \gamma|_{[\epsilon,b)}$ is a past directed

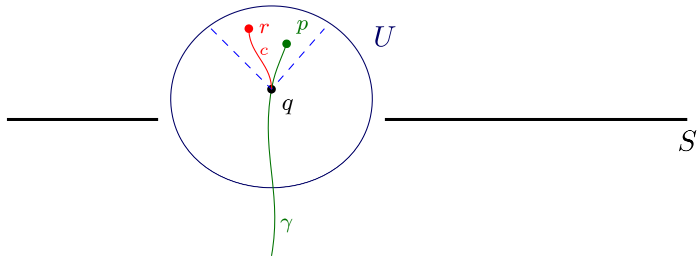


Figure 21: (ii) in proof of Proposition 5.2.19.

inextendible timelike curve starting at $r \in D^+(S)$. Therefore it has to meet S and one obtains a contradiction.
(ii) " $\Leftarrow$ " Let $p \notin \overline{D^+(S)}$ and $q \in I^-_{M \setminus \overline{D^+(S)}}(p)$. Note that $q \notin D^+(S)$, which implies that there exists a past directed inextendible causal curve γ from q which does not intersect S. By the previous Lemma 5.2.18 there is a past directed inextendible timelike curve $\tilde{\gamma}$ from p with $\tilde{\gamma} \cap S = \varnothing$ and thus we are done.
(iii) We show the inclusion $S \subseteq \partial D^+(S)$. Note that we have $S \subseteq D^+(S) \subseteq \overline{D^+(S)}$ and let $p \in S \cap \operatorname{int} D^+(S)$. Choose $q \in \operatorname{int} D^+(S) \cap I^-(p)$ and let γ be a past directed inextendible timelike curve starting at q, i.e. $\gamma \cap S \neq \varnothing$ (see Figure 22). For any $r \in \gamma \cap S$ one has $r \ll p$, which contradicts achronality of S.

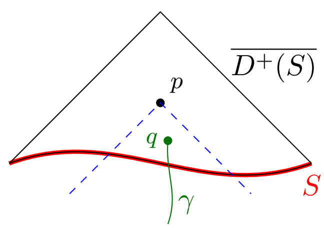


Figure 22: (iii) in proof of Proposition 5.2.19.

Next, we prove $H^+(S) \subseteq \partial D^+(S)$. By definition $H^+(S) \subseteq \overline{D^+(S)}$ so suppose $p \in H^+(S) \cap \operatorname{int} D^+(S)$. It follows that $I^+(p) \cap D^+(S) \neq \varnothing$, which contradicts that $p \in H^+(S)$.
Finally, one needs to demonstrate that indeed $\partial D^+(S) \subseteq S \cup H^+(S)$. Again, assume for the sake of contradiction that there exists $p \in \partial D^+(S) \setminus (S \cup H^+(S))$. By (ii) we have $p \in I^+(S)$. Also, since $p \in \overline{D^+(S)} \setminus H^+(S)$, there exists a $q \in I^+(p) \cap D^+(S)$. We therefore conclude that p is contained in the open set $I^+(S) \cap I^-(q)$.
Let $r \in I^+(S) \cap I^-(q)$ and suppose γ is past directed inextendible causal starting at r. Further, let $\tilde{\gamma}$ be a past directed timelike curve from q to r. Given $r \in I^+(S)$, we must have $\tilde{\gamma} \subseteq I^+(S)$. By achronality, $S \cap I^+(S) = \varnothing$ and thus $\tilde{\gamma} \cap S = \varnothing$. Since $q \in D^+(S)$ the curve $\tilde{\gamma} \cup \gamma$ meets S, so γ must meet S and we have $r \in D^+(S)$. Since r was arbitrary, we have shown that $I^+(S) \cap I^-(q) \subseteq D^+(S)$. But this is impossible since p can then only be in the interior and not on the boundary of $D^+(S)$. $\square$

Property (i) together with Proposition 5.1.9 yields the following useful Corollary

Corollary 5.2.20. *Let S be an achronal subset of M. Then $H^+(S) \setminus \operatorname{edge} H^+(S)$ is a closed, achronal topological hypersurface of M if it is nonempty.*

We may now show a result reminiscent of Proposition 5.1.12. The proof is also of similar nature, i.e. we construct the null geodesics directly using the Limit Curve Lemma. For this, we adapt arguments in [Wal84, Thm. 8.3.5] and [Kri99, Prop. 8.3.1].

Proposition 5.2.21. *Let S be a closed, achronal subset of M and $p \in H^+(S) \setminus \operatorname{edge}(S)$. Then p is the future endpoint of a null geodesic contained in $H^+(S)$ which is either past inextendible in M or meets* $\operatorname{edge}(S)$.

Proof. Note that by definition $H^+(S) \subseteq \overline{D^+(S)} \subseteq S \cup I^+(S)$. We consider two cases $p \in I^+(S)$ or $p \in S$ separately.
For the first, let $p \in (H^+(S) \setminus \operatorname{edge} S) \cap I^+(S)$ and take a sequence $\{p_n\}_{n\in\mathbb{N}}$ with $p_n \in I^+(p)$ and $p_n \to p$ for each $n \in \mathbb{N}$. Note that for all $n \in \mathbb{N}$ there exist past directed inextendible timelike curves γ_n with $\gamma_n(0) = p_n$ and such that $\gamma_n \cap S = \varnothing$ (see Figure 23). One may now apply the Limit Curve Lemma 4.2.7 to obtain a subsequence $\{\gamma_m\}_{m\in\mathbb{N}}$ which converges uniformly on compact sets to a continuous past directed inextendible causal curve γ with $\gamma(0) = p$.

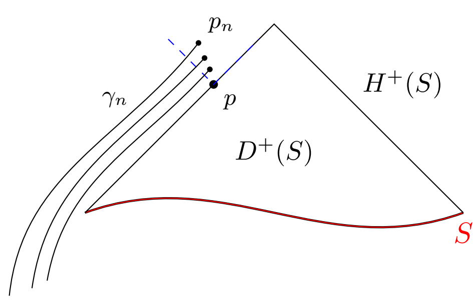


Figure 23: Setup for first case in proof for Proposition 5.2.21.

By construction $\gamma_m(t) \notin \overline{D^+(S)}$. We have $\gamma_m(t) \to \gamma(t)$ and therefore $\gamma(t) \in M \setminus \operatorname{int} D^+(S)$. By continuity, we have that $\gamma(t) \in I^+(S)$ for small t. Suppose now that $\gamma(t) \notin \partial D^+(S) \subseteq \overline{D^+(S)}$. Consequently, there exists a past directed inextendible timelike curve c starting at $\gamma(t)$ that does not meet S, i.e. $c \cap S = \varnothing$. Note that for any point $c(s)$ one has $c(s) \ll \gamma(t)$. Using the fact that $\gamma(t) \leq p$ it follows that $c(s) \ll p$ by push-up. Therefore, we must conclude $c(s) \in I^-(S)$ as c does not meet S. Observe that by continuity of c then $\gamma(t) \in \overline{I^-(S)}$. But $I^+(S)$ is an open neighbourhood of $\gamma(t)$ and thus $I^+(S) \cap I^-(S) \neq \varnothing$, which contradicts the fact that S is achronal. Hence $\gamma(t) \in \partial D^+(S)$ and Proposition 5.2.19 (iii) implies $\gamma(t) \in H^+(S)$. That $\gamma|_{\gamma^{-1}(I^+(S))}$ is the null geodesic we are looking for then follows in the exact same way as in the proof of 5.1.12 together with the fact that $H^+(S)$ is achronal.

For the second case let $p \in H^+(S) \setminus \mathrm{edge}(S) \cap S$ and obtain subsequences $\{p_m\}_{m\in\mathbb{N}}$ and $\{\gamma_m\}_{m\in\mathbb{N}}$ together with a limit curve γ, in the same way as above (see Figure 24).

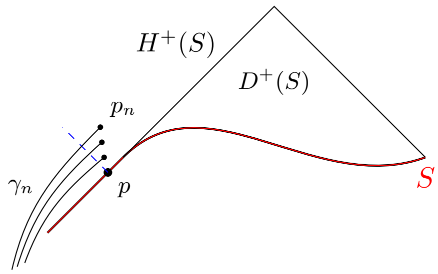


Figure 24: Setup for second case in proof for Proposition 5.2.21.

Since $p \notin \mathrm{edge}(S)$, there exists a neighbourhood $U \subset M$ of p such that any timelike curve from $I_U^-(p)$ to $I_U^+(p)$ in U intersects S. Note that by construction $\gamma_m \cap I_U^-(p) = \varnothing$ and therefore $\gamma \cap I_U^-(p) = \varnothing$ because $I_U^-(p)$ is open. Denote by $\tilde{\gamma}$ the segment of γ inside U, i.e. $\tilde{\gamma} = \gamma \cap U$. We want to show that $\tilde{\gamma}(t) \in S$ for any t, so suppose to the contrary that there exists such $t \neq 0$ with $\tilde{\gamma}(t) \notin S$. Choose a convex neighbourhood $C \subseteq U$ of $\tilde{\gamma}(t)$ such that $C \cap S = \varnothing$ (possible since S is closed). One has that $I_C^-(p) \neq \varnothing$ since $\tilde{\gamma}(t) \in J^-(p) \subseteq \overline{I^-(p)}$. Because C also contains infinitely many γ_m and $I^-(p) \subseteq I^-(\gamma_m)$, there exists a timelike curve α in C from $q \in I_C^-(p)$ to $\gamma_m(t) \in I_C^+(p)$ for large enough m. But note that α is naturally contained in U, i.e. α is a timelike curve from $I_U^-(p)$ to $I_U^+(p)$ and because it does not meet S, we get a contradiction with the fact that $p \notin \mathrm{edge}(S)$. Consequently, $\tilde{\gamma}(t) \in S \subseteq \overline{D^+(S)}$.
We now want to show that $\tilde{\gamma}(t) \in H^+(S)$. Again, for the sake of contradiction, assume that there exists $q \in D^+(S) \cap I^+(\tilde{\gamma}(t))$. Choose a convex neighbourhood C of $\tilde{\gamma}(t)$ such that $C \subseteq I_U^-(q)$. By push-up follows $I_C^-(\tilde{\gamma}(t)) \subseteq I_C^-(p)$ and thus $I_C^+(I_C^-(p))$ is an open neighbourhood of $\tilde{\gamma}(t)$. This means, together with the fact that C needs to intersect γ_m for large m, we can pick a point $r \in \gamma_m \cap C \cap I_C^+(I_C^-(p))$. Now, let β be a timelike curve from r to q and let λ be timelike from $I_C^-(p)$ to r in C (see Figure 25).

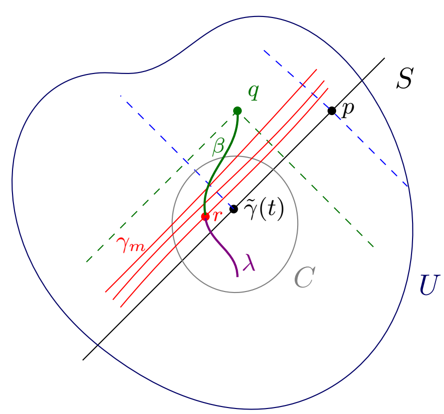


Figure 25: Setup for showing $\tilde{\gamma}(t) \in H^+(S)$ in proof of Proposition 5.2.21.

The concatenation of λ with the future part of γ_m is a timelike curve from $I^-_U(p)$ to $I^+_U(p)$ since $C \subseteq U$ and therefore intersects S. By construction, none of the γ_m meet S, so we can conclude that $\lambda \cap S \neq \varnothing$. By achronality of S then follows, that $\beta \cap S = \varnothing$. Finally, consider $\beta \cup \gamma_m|_{[\gamma_m^{-1}(r),\infty)}$. This is a past directed, inextendible timelike curve that starts at q and does not meet S. But this yields a contradiction because $q \in D^+(S)$. Therefore, $D^+(S) \cap I^+(\tilde{\gamma}(t)) = \varnothing$ and equivalently $\tilde{\gamma}(t) \in H^+(S)$.
Note that we have shown that $\tilde{\gamma}(t) \in S$, so by achronality of S it follows that γ is a null geodesic by the exact same reasons as before.
In summary, we have shown that any $p \in H^+(S) \setminus \mathrm{edge}(S)$ is contained on a null geodesic. One only needs to prove that this null geodesic γ is either inextendible or has a past endpoint in $\mathrm{edge}(S)$. For this, suppose that γ leaves $H^+(S)$. Since $H^+(S)$ is closed there exists a past endpoint $q \in H^+(S)$, after which γ leaves $H^+(S)$. Denote by γ the geodesic segment from p to q. Assume that $q \notin \mathrm{edge}(S)$. By the above there exists a null geodesic segment γ' that starts in q and remains in $H^+(S)$. If γ' is not a continuation of the geodesic γ, then by Proposition 4.3.5 there exist timelike curves from points on γ' to γ, which contradicts achronality of $H^+(S)$. Therefore $q \in \mathrm{edge}(S)$ and we are done. □

Remark 5.2.22. By time symmetry one may easily show that $H^-(S)$ is similarly ruled by null geodesics which are either future inextendible or terminate in $\mathrm{edge}(S)$.
Taking another look at Figure 18, this previous result is not really surprising and rather intuitive or expected. Nevertheless, it should be highlighted that the study of null curves or geodesics seems to be of fundamental significance in understanding the geometry of several significant subsets in M.

We close this section with a result that will be extremely important in proving the Penrose Incompleteness Theorem later on, and which also connects a lot of the results we have collected so far in an elegant manner. We follow [Gal14, Prop. 4.8].

Proposition 5.2.23. *Let M be globally hyperbolic and S be a compact, achronal topological hypersurface. Then S is a Cauchy surface.*

Proof. We show that $H(S) = \varnothing$ and the result follows by Proposition 5.2.17. Let $p \in H^+(S)$. Since S is compact, it is also closed and by Propositions 5.1.9 and 5.2.21 we have that $\mathrm{edge}(S) = \varnothing$ and that p is the future endpoint of a past inextendible null geodesic c with $c \subset H^+(S)$. Clearly $c \subset J^-(p)$ and by Proposition 5.2.14 (i) one has $c \subset H^+(S) \subseteq \overline{D^+(S)} \subseteq J^+(S)$ since $J^+(S)$ is closed following from Proposition 5.2.8 (i). Therefore, again by Proposition 5.2.8, c is contained in the compact set $J^+(S) \cap J^-(p)$. But this yields a contradiction because by the past version of Lemma 5.2.3, c has to leave any compact set. Consequently $H^+(S) = \varnothing$. The case $p \in H^-(S)$ then follows in the same exact way by time symmetry. Thus $H(S) = H^+(S) \cup H^-(S) = \varnothing$ and we are done. □

6 Null Geometry

The Penrose Incompleteness Theorem is a result that concerns the geodesic inextendibility of null geodesic on a spacetime. In order to prove such a statement it is not at all surprising that one should investigate objects that are closely related to null geodesics, e.g. achronal boundaries. The goal of this chapter is therefore to discuss and introduce the notion of *null submanifolds*. Recall that we already introduced Semi-Riemannian Submanifolds back in Chapter 3. Summarized, these objects were defined as subsets P of a spacetime (M, g) that carried the structure of a smooth submanifold such that the metric g carried over to a metric tensor on P. One question now remains: if P is a smooth submanifold, what happens when g does not restrict to a metric tensor $\not{g}$ on P? The answer to this is that $\not{g}$ may be degenerate. It is this particular case, where one calls P a *null submanifold*. As one will see, P features a well-defined global null vector field and, similarly to achronal boundaries or Cauchy horizons, is ruled by null geodesics. We emphasize that null submanifolds are a crucial part of the machinery that enables us to prove the Incompleteness Theorem later on, thereby making them objects of central importance.

6.1 Null Submanifolds and Hypersurfaces

The first section is dedicated to the introduction of null submanifolds and, in particular, null hypersurfaces. Our primary goal being to define geometric notions that are analogous to the Semi-Riemannian case. To this end, we have a big issue to face, namely the degeneracy of $\not{g}$. Note that this has severe implications on our analysis and ones needs to address several technical issues to enable us to develop the right tools, which one needs in the later sections. As for the main references, we will follow [Gal14] and [Are18].

As mentioned in the introduction, at this point we already have encountered two interesting types of subsets of a spacetime (M, g), namely $\partial I^+(S)$ and $H^+(S)$ for some achronal $S \subset M$. Note that both of these sets have in common that they exhibit the property of being ruled by null geodesics. By this we mean that any point in $\partial I^+(S)$ or $H^+(S)$ is contained on a null geodesic that is itself contained within these sets (cf. Proposition 5.1.12 and Proposition 5.2.21). Furthermore, we have also seen that they carry the structure of a topological hypersurface with the caveat that there is no hope in expecting these subsets to be smooth manifolds in general. However, note that this may be rectified by getting rid of the "problematic" part of the set, e.g. by considering something like $\partial I^+(p) \setminus \{p\}$ instead of $\partial I^+(p)$.

Turning to smooth submanifolds of M, we will see that null submanifolds of M also exhibit the property of being ruled by null geodesics. In order to properly define this class of submanifold, recall from Chapter 3 that for a smooth submanifold P in M we denoted by $\not{g}$ the pullback of the metric g under the inclusion map $j : P \to M$. Note now, since g has Lorentzian signature, Sylvester's law of inertia (e.g. [Gre75, Chapter 9.2]) then implies that we have only the following three cases to consider

Definition 6.1.1. *Let $N \subseteq M$ be a smooth submanifold and let $\not{g}$ be the restricted metric on*

N. If for all $p \in N$ we have that $\not{g}_p$

*(i) has positive signature, then N is called **spacelike**,*

*(ii) has Lorentzian signature, then N is called **timelike**,*

*(iii) is degenerate, then N is called **null**.*

For cases (i) and (ii) we immediately see that N is either a Riemannian or a Lorentzian manifold in its own right respectively. In case (iii) however, the restriction $\not{g}$ is degenerate and therefore does not define metric tensor on N. Alas, this makes our live more difficult since, at first glance, the usual way of studying the geometry of submanifolds does not work out. Nevertheless, we will show how one may solve this problem and that all the curvature related quantities can in fact be introduced much in the same spirit as done for non-degenerate submanifolds. In what follows, we assume that any submanifold is smooth if not specifically stated otherwise. Furthermore, we want to concern ourselves primarily with null hypersurfaces.

Definition 6.1.2. *A null hypersurface H in M is a null submanifold with* $\mathrm{codim}(S) = 1$.

Example 6.1.3. Null hypersurfaces (or null submanifolds respectively) are of particular interest in several applications found in physics. Especially in general relativity, one encounters them in the form of various horizons. Let us consider a few examples.

(i) $M = \mathbb{R}^{1,d}$ and let $x \in \mathbb{R}^{1,d}$. Then the null cone minus the vertex $\partial I^+(p) \setminus \{p\}$ is a null hypersurface.

(ii) Again $M = \mathbb{R}^{1,d}$. Pick any null vector $X \in T_p\mathbb{R}^{1,d}$ and consider the hyperplane $\{q \in \mathbb{R}^{1,d} \,|\, \eta(v, X) = 0 \text{ with } v = p - q\}$, which is a null hypersurface (see Figure 26).

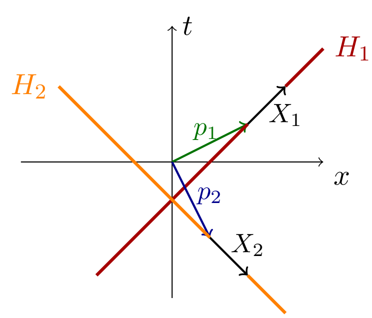


Figure 26: Two null hypersurfaces H_1 and H_2 in Minkowski space $\mathbb{R}^{1,1}$ given by the above, for foot points p_1, p_2 and null vectors X_1, X_2 respectively.

(iii) Consider the so called Eddington-Finkelstein extension of the Schwarzschild solution to Einstein's vacuum field equations, i.e. $M = \mathbb{R} \times (0, \infty) \times S^2$ with coordinates (v, r, θ, ϕ)

and the metric (see e.g. [Rea24, Chapter 2])

$$g = -\left(1 - \frac{2m}{r}\right) dv \otimes dv + 2\, dv \otimes dr + r^2 d\theta \otimes d\theta + r^2 \sin^2\theta\, d\phi \otimes d\phi$$

for some $m > 0$. Note that r is to be interpreted as a radial coordinate. This spacetime models the vaccuum region outside of a spherically symmetric stellar object, e.g. a non-rotating star. In that case m has the interpretation of mass. Let $r_s := 2m$, which is known as the *Schwarzschild radius*. The *event horizon* is defined as the set $H = r^{-1}(\{r_S\}) \subseteq M$, which forms the boundary of the so called *Schwarzschild black hole region* (see Figure 27). One may show that $dr|_{r=r_S} \neq 0$ and that $\frac{\partial}{\partial v}$ is a degenerate vector field for H. From this follows then that H is a null hypersurface.

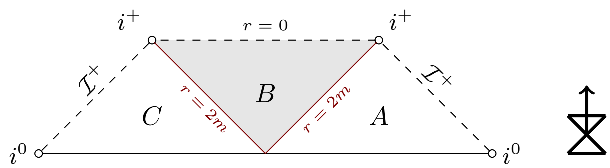


Figure 27: The fully future extended Schwarzschild solution as a Penrose diagram with regions A and B given by the Eddington-Finkelstein extension. The event horizon is drawn in red and the black hole region is the shaded part of the picture. Region B is called black hole region: No future directed causal geodesic in B is able to escape this region and must in fact travel towards $r = 0$. For a detailed discussion of this diagram, including the special subsets i^0, i^+ and $\mathcal{I}^+$, we refer to [Rea24, Chapters 2 & 5], [Wal84, Ch. 6,11 and 12] and [DR05, Appendix C].

Suppose $(H, \not{g})$ is a null submanifold of a spacetime (M, g). The degeneracy of $\not{g}$ implies that for any $p \in H$, there exists $L_p \in T_pH$ with $\not{g}_p(L_p, X) = \langle L_p, X\rangle = 0$ for all $X \in T_pH$. Therefore we have that $L_p \in T_pH^\perp$ and because $\dim(T_pH^\perp) = 1$, it follows that $T_pH^\perp = \operatorname{span}(\{L_p\})$ and similarly $T_pH = \operatorname{span}(\{L_p\}^\perp)$. Note that in particular, $\langle L_p, L_p\rangle = 0$, which means that L_p is null. In summary, at any point in H there exists a null vector L_p which is simultaneously tangent and normal to H.
The next Lemma contains a very useful fact concerning the tangent vectors of null hypersurfaces H. We follow [Are18, Lemma 2.2.1, p.37f].

Lemma 6.1.4. *Let H be a null hypersurface in M and $p \in H$. Let $L_p \in T_pH$ be the null normal at p and let $X \in T_pH$ with $X \neq 0$. Then, X is either spacelike or $X = \lambda L_p$ for some $\lambda \in \mathbb{R}$.*

Proof. Since this is a local problem, we can work with an orthonormal basis at T_pM. Hence we can write $X = (x^0, \overline{x})$ and $L = (l^0, \overline{l})$, with $\overline{x} = (x^1, \dots, x^n)$ and $\overline{l} = (l^1, \dots, l^n)$. Note that by assumption we have $\langle L_p, X\rangle = 0$ and $\langle L_p, L_p\rangle = 0$, i.e. $0 = -x^0 l^0 + \overline{l}\cdot\overline{x}$ and $0 = -(l^0)^2 + \|\overline{l}\|^2$, where $\|\cdot\|$ is the Euclidean norm. Using this and applying the Cauchy-Schwarz inequality yields

$$\left(x^0\right)^2 \left(l^0\right)^2 = (\overline{x} \cdot \bar{l})^2 \leq \|\bar{l}\|^2 \, \|\overline{x}\|^2.$$

Since $(l^0)^2 = \|\bar{l}\|^2$ we then get $(x^0)^2 \leq \|\overline{x}\|^2$. Therefore $\langle X, X\rangle \geq 0$ with equality if and only if $\overline{x} = \lambda \, \bar{l}$, which implies $x^0 l^0 = \bar{l} \cdot \overline{x} = \lambda \|\bar{l}\|^2 = \lambda \, (l^0)^2$ and thus also $x^0 = \lambda \, l^0$. In summary, either X is spacelike or $(x^0, \overline{x}) = \lambda \, (l^0, \bar{l})$. □

Taking into account that there exists a null normal vector at each point of a null hypersurface H that is also tangent to H, one may use the above result to infer the following.

Proposition 6.1.5. *Any null hypersurface H in M exhibits a smooth, future directed null vectorfield $L \in \mathfrak{X}(H) \cap \mathfrak{X}^{\perp}(H)$, which is unique, up to some positive smooth scalar function.*

Proof. We want to define a smooth map $L : H \to TH$ with $p \mapsto L_p$ where $L_p \in T_pH \cap T_pH^{\perp}$, i.e. L_p is a null normal vector of H at p. Of course, this can always be done locally in H using a local orthonormal frame. Therefore, let $U_i \subseteq H$ with $i \in \mathbb{N}$ such that $\bigcup_{i \in \mathbb{N}} U_i = H$ and that we can define a smooth null normal vector field L_i on U_i. Let T be a global future directed timelike vector field that gives the time orientation on M. For each i, define $\tilde{L}_i = f_i L_i$ with $f_i = -\langle L_i, T\rangle \neq 0$. Thus we have normalized the L_i in a uniform way such that $\langle \tilde{L}_i, T\rangle = -1$, i.e. the $\tilde{L}_i$ are future directed. Let $i \neq j$ be such that $U_i \cap U_j \neq \varnothing$. Note that on this overlap, there must exist a nonzero smooth function h such that $L_i = hL_j$ since they point in the same direction. It follows that

$$\tilde{L}_i = \frac{L_i}{-\langle L_i, T\rangle} = \frac{hL_j}{-\langle hL_j, T\rangle} = \frac{hL_j}{-h\langle L_j, T\rangle} = \frac{L_j}{-\langle L_j, T\rangle} = \tilde{L}_j$$

and $\tilde{L}_i$ agrees with $\tilde{L}_j$ on overlaps. Now simply define $L : H \to TH$ by $L(p) = L_i(p)$ if $p \in U_i$, which is clearly smooth. For uniqueness, observe that we can scale L by any positive smooth function we desire. □

Remark 6.1.6. The existence of a null normal vector field L as above is quite convenient. In particular, following [Kup87, Cor. 5], one may easily show that if M is orientable then H is also orientable: Let $X \in \mathfrak{X}(M)$ be future directed and timelike and let $\omega \in \Omega^{n+1}(M)$ be the volume form on M. Note that by Lemma 6.1.4 X is transversal to H and therefore $i_X\omega \in \Omega^n(M)$ —the inner product of X and ω— restricts to a nowhere vanishing n-form on H.

In general, dealing with null hypersurfaces (resp. null submanifolds) is different and more delicate as compared to non-degenerate hypersurfaces (resp. submanifolds). The existence of a smooth null vector field which is tangent but also normal to H reflects that we do **not** have the usual splitting of the tangent bundle $TM|_H = TH^{\perp} \oplus TH$ because $TH^{\perp} \cap TH \neq \varnothing$ and in fact $T_pM \neq T_pH^{\perp} + T_pH$ for $p \in H$. Thus, there are no canonical projections of TM into a tangential and orthogonal part. This has the direct consequence that the Levi-Civita connection on M generally does not induce a linear connection on H in a generic way. In particular, any linear connection defined by projecting into TH will depend on the choice of the null vector field L. We are therefore unable to define the usual notions that were used to examine the intrinsic curvature (e.g. second fundamental form, shape operator, sectional curvature, etc.) in Chapter 3 for non-degenerate submanifolds.

But before properly addressing this issue, we are first going to show that the integral curves of the null normal vector field L for a null hypersurface H are actually null geodesics if suitably parametrized. This of course then implies that any null hypersurface is also actually ruled by null geodesics, as we have mentioned in the beginning of the section. Notably, this really aids our geometric intuition when thinking about these kind of submanifolds. We follow [Are18, Prop. 2.2.1].

Proposition 6.1.7. *Let H be a null hypersurface with a future directed null vector field $L \in \mathfrak{X}(H)$. The integral curves of L can be parametrized to be null geodesics, the so called null generators.*

Proof. We want to show that $\nabla_L L = \lambda L$ for some $\lambda \in \mathcal{C}^\infty(H)$. Then it automatically follows that the integral curves of L are pregeodesics and the claim follows. Let $X \in TH$. We have

$$\langle \nabla_L L, X \rangle = -\langle L, \nabla_L X \rangle = -\langle L, \nabla_X L \rangle - \langle L, [L, X] \rangle = -\frac{1}{2} X(\langle L, L \rangle) = 0,$$

where we have used the fact that $[L, X] \in TH$. This is true at any $p \in H$ and for any $X_p \in T_p H$, i.e. $\langle \nabla_L L|_p, X_p \rangle = 0$. Therefore $\nabla_L L|_p = \lambda(p) L_p$ for all $p \in H$ and thus $\nabla_L L = \lambda L$ as claimed. □

Remark 6.1.8. It is clear from the proof above, that this result is not in general true for arbitrary null submanifolds. For example ([Kup87, p.35]) consider 3-dimensional Minkowski space, $\mathbb{R}^{1,2}$, and the 1-dimensional submanifold given by the image of the curve $\alpha(t) = (t, \sin(t), \cos(t))$ for $t \in \mathbb{R}$, which is definitely not a null (pre)geodesic. At the same time however, this submanifold is indeed null because $\alpha'(t) = (1, \cos(t), -\sin(t))$ and therefore $\not{g}(t) = g(\alpha'(t), \alpha'(t)) = -1^2 + \cos^2(t) + (-\sin(t))^2 = 0$.

From now on, we will always consider a null hypersurface H of a spacetime M with a particular choice of null tangent vector field L.

So far, we have only concerned ourselves with properties of H that followed by extrinsic considerations using the ambient spacetime M. But of course our goal is also to makes sense of the intrinsic geometry of H itself. To this end, we have to face some pressing issues. It was mentioned before that the fundamental problem is the absence of a splitting of the tangent bundle of M. Note that the intersection $TH^\perp \cap TH =: \mathcal{L}$ is actually a trivial one dimensional null line bundle[31] spanned by L and that $L \in \Gamma(\mathcal{L})$. Consequently, we hope to resolve our

[31] This follows by the local frame criterion for subbundles, see e.g. [Lee13, Lemma 10.32].

issues by modding out $\mathcal{L}$ or rather the span of L_p at each $p \in H$. It turns out that this is the desired setting to define the classic intrinsic quantities used for non-degenerate hypersurfaces, in particular such that they also contain similar geometric information.

Therefore let $L \in \Gamma(\mathcal{L})$ be non-vanishing and $p \in H$. The vector $L_p \in T_pH$ is a tangent null vector with respect to H. Note that $\operatorname{span}\{L_p\}$ is a linear subspace of T_pH and consequently we can consider the quotient vector space $T_pH/L \coloneqq T_pH/\operatorname{span}\{L_p\}$. To simplify notation, the equivalence class of an element $X \in T_pH$ will be denoted by a bar on top, e.g. $\overline{X} \in T_pH/L$. The collection of these classes then forms a vector bundle over H.

Proposition 6.1.9. *$TH/\mathcal{L} = \bigsqcup_{p\in H} T_pH/L$ is a vector bundle of rank $n-1$ over H. Furthermore, $TH/\mathcal{L}$ is independent of the choice of L.*

Proof. We will construct the local trivializations of $TH/\mathcal{L}$ by using the fact that $\mathcal{L}$ is a subbundle of TH. Let $p \in H$ and let $U \subseteq H$ be an open neighbourhood of p. Furthermore, let L be a local frame for $\mathcal{L}$ on U. Shrinking U if necessary, one may extend L by local sections $E_1, \ldots, E_{n-1}$ of TH on U in order to obtain a local frame $\{L, E_1, \ldots, E_{n-1}\}$ on TH. Using this, it is possible to define a local trivialization $\Phi : \pi^{-1}(U) \to U \times \mathbb{R}^n$ of TH at p such that $\Phi^{-1}\left(p, (v^L, v^1, \ldots, v^{n-1})\right) = v^L L(p) + \sum_{i=1}^{n-1} v^i E_i(p)$ (cf. [Lee13, Prop. 10.19]). Note that its restriction to $\mathcal{L}$, denoted by $\Phi^{\mathcal{L}}$, also yields a local trivialization of the form $\pi^{-1}(U) \cap \mathcal{L} \to U \times \mathbb{R}$. One may now simply pass to the quotient and define $\Psi : \left(\pi^{\mathcal{L}}\right)^{-1}(U) \to U \times \mathbb{R}^n/\mathbb{R} \cong \mathbb{R}^{n-1}$ by $\overline{X} \in T_pH/L \mapsto (\mathrm{id}_U \times \tilde{q}) \circ \Phi(X)$ where $\tilde{q} : \mathbb{R}^n \to \mathbb{R}^n/\mathbb{R}$ is the quotient map with respect to the equivalence relation $v \sim v'$ if and only if $v - v' = (\lambda, 0, \ldots, 0)$ for $\lambda \in \mathbb{R}$ and where $\pi^{\mathcal{L}} : TH/\mathcal{L} \to H$ is the canonical projection map. Of course Ψ is well-defined because all the involved maps are constant on the fibers. Furthermore, Ψ is also clearly bijective and an isomorphism from T_qH/L to $\{q\} \times \mathbb{R}^{n-1}$ for each $q \in U$. In term of transition maps, note that if (Φ, U) and (Φ', V) are local trivializations of TH constructed as before with $U \cap V \neq \varnothing$, one has

$$\Phi' \circ \Phi^{-1}(x, v) = (x, \tau(x) \cdot v)\,,$$

where $\tau : U \cap V \to \mathrm{GL}(n)$ is smooth and of the form

$$\tau(x) = \begin{pmatrix} \lambda(x) & a^T(x) \\ 0 & A(x) \end{pmatrix},$$

where $\lambda : U \cap V \to \mathbb{R}$, $a : U \cap V \to \mathbb{R}^{n-1}$ and $B : U \cap V \to \mathrm{GL}(n-1)$ are smooth. The transition function for $\Psi' \circ \Psi^{-1}$ are then simply given by B.

The above construction holds for any $p \in H$ and thus is sufficient[32] to conclude that $TH/\mathcal{L}$ indeed is a vector bundle of rank $n-1$ over H. In particular it is also independent of the choice of null vector field L, as any other choice $\tilde{L}$ satisfies $L = f\tilde{L}$ for some positive $f \in \mathcal{C}^\infty(H)$,

[32] See [Lee13, Lemma 10.6] for a vector bundle construction result.

which in turn gives rise to the same quotient structure. $\square$

Denote the set of smooth sections $\Gamma(TH/\mathcal{L})$, the *mod $\mathcal{L}$ vector fields*, by $\mathfrak{X}^{\mathcal{L}}(H)$. We make the following observation: Let $\overline{X} \in \mathfrak{X}^{\mathcal{L}}(H)$ and choose $E_1, \ldots, E_{n-1} \in \mathfrak{X}(U)$ such that, together with L, they span T_pH for each $p \in U$ in $U \subseteq H$ open. Note that one can then write $\overline{X} = \sum_{i=1}^{n-1} X^i \overline{E}_i$ with $X^i \in \mathcal{C}^\infty(U)$ for each $i \in \{1, \ldots, n-1\}$. Naturally $X = \sum_{i=1}^{n-1} X^i E_i$ is a (local) representative of $\overline{X}$. In this manner, we obtain a smooth vector field representative X on all of H for any mod $\mathcal{L}$ vector field.
As usual, one may define the dual bundle $T^*H/\mathcal{L}$ of $TH/\mathcal{L}$. The set of smooth sections of $T^*H/\mathcal{L}$ will be written as $\Omega^{\mathcal{L}}(H)$. Furthermore, we can also construct various other kinds of tensor bundles over H with respect to the modded tangent space, which we will sometimes refer to as H mod $\mathcal{L}$. We will denote the (k, s) tensor bundle over H mod $\mathcal{L}$ by $T^k_s(TH/\mathcal{L})$ and its smooth sections by $\mathcal{T}^k_s(TH/\mathcal{L})$.
Having established the technical details related to the quotient structure, one may now show that the map $\not{g}$, or by extension g itself, descends to a positive definite metric h on $TH/\mathcal{L}$ as follows

Proposition 6.1.10. *Define the map $h : H \to T^0_2(TH/\mathcal{L})$ pointwise for $p \in H$ by $h_p : T_pH/L \times T_pH/L \to \mathbb{R}$ with*

$$h(p)(\overline{X}, \overline{Y}) = h_p(\overline{X}, \overline{Y}) \coloneqq \langle X, Y \rangle.$$

Then h is a positive definite metric on H mod $\mathcal{L}$.

Proof. First, we show that h is well-defined. Let $X', Y' \in T_pH$ with $\overline{X'} = \overline{X}$ and $\overline{Y'} = \overline{Y}$. Then

$$\begin{aligned} h_p(\overline{X'}, \overline{Y'}) &= \langle X', Y' \rangle = \langle X + \lambda L, Y + \mu L \rangle \\ &= \langle X, Y \rangle + \lambda \langle L, Y \rangle + \mu \langle X, L \rangle + \lambda\mu \langle L, L \rangle \\ &= \langle X, Y \rangle = h_p(\overline{X}, \overline{Y}). \end{aligned}$$

Next one needs to prove that h is smooth. For this let $\overline{X}, \overline{Y} \in \mathfrak{X}^{\mathcal{L}}(H)$ and let $X, Y \in \mathfrak{X}(H)$ be their representatives respectively. Smoothness then follows by the fact that the map $p \mapsto h(p)\left(\overline{X}(p), \overline{Y}(p)\right) = \langle X, Y \rangle|_p$ is smooth and $\overline{X}, \overline{Y}$ were arbitrary.
Since symmetry is clear, all that is left is to show that h is indeed positive definite. Let $\overline{X} \in T_pH/L$ with $\overline{X} \neq \overline{0}$. Then $X \neq \lambda L$ for any $\lambda \in \mathbb{R}$. Now Lemma 6.1.4 implies that X is spacelike. Thus $h_p(\overline{X}, \overline{X}) = \langle X, X \rangle > 0$. We conclude that h is non-degenerate and of constant index, finishing the proof. $\square$

Since we have a metric tensor on H, one may start to define the null equivalents of the usual geometric operators on non-degenerate hypersurfaces.

Proposition 6.1.11. *The null Weingarten map $b \in \mathcal{T}^1_1(TH/\mathcal{L})$ on H with respect to L, is defined at each $p \in H$ by the linear map $b_p : T_pH/L \to T_pH/L$ with*[33]

$$b(p)(\overline{X}) = b_p(\overline{X}) \coloneqq \overline{\nabla_X L}$$

is uniquely determined by L and self-adjoint with respect to h.

Proof. First we show that b is well-defined. Let $p \in H$ and $X, Y \in T_pH$ such that $\overline{X} = \overline{Y}$. Then

$$\begin{aligned} b_p(\overline{Y}) &= \overline{\nabla_Y L} = \overline{\nabla_{X+\lambda L_p} L} = \overline{\nabla_X L + \lambda \nabla_{L_p} L} \\ &= \overline{\nabla_X L + \kappa L_p} = \overline{\nabla_X L} = b_p(\overline{X}), \end{aligned}$$

where we have used the fact that $\lambda \nabla_{L_p} L = \kappa L_p$ for some $\kappa \in \mathbb{R}$, like in the proof of Proposition 6.1.7 and also that $\nabla_X L + \kappa L_p \sim_L \nabla_X L$.
Smoothness follows in similar fashion as in Proposition 6.1.10 by noting that the image $\overline{\nabla_X L}$ is a smooth mod $\mathcal{L}$ vector field.
For uniqueness, let $\tilde{L} = fL$ for some $f \in \mathcal{C}^\infty(H)$ and define the map $\tilde{b}$ using $\tilde{L}$. We show that $\tilde{b} = fb$. Let $p \in H$ and $\overline{X} \in T_pH/L$. Note that the quotient is independent of our choice of L, so also $\overline{X} \in T_pH/\tilde{L}$. We get

$$\begin{aligned} \tilde{b}_p(\overline{X}) &= \overline{\nabla_X \tilde{L}} = \overline{\nabla_X fL} = \overline{X(f)L + f(p)\nabla_X L} = \overline{\lambda L + f(p)\nabla_X L} \\ &= \overline{f(p)\nabla_X L} = b_p(f(p)\overline{X}) = f(p)b_p(\overline{X}), \end{aligned}$$

where linearity of b_p was used in the last line. Thus we have shown $\tilde{b}_p = f(p)\, b_p$ for any $p \in H$ and consequently $\tilde{b} = fb$.
Finally, we prove self-adjointness. Again, let $p \in H$ and $\overline{X}, \overline{Y} \in T_pH/L$. Extend $X, Y \in T_pH$ to any vector fields on M which are tangent to H. This yields

$$\begin{aligned} h(b_p(\overline{X}), \overline{Y}) &= \langle \nabla_X L, Y \rangle\big|_p = -\langle L, \nabla_X Y \rangle\big|_p = -\langle L, \nabla_Y X \rangle\big|_p - \langle L, [X, Y] \rangle\big|_p \\ &= -\langle L, \nabla_Y X \rangle\big|_p = \langle \nabla_Y L, X \rangle\big|_p = h(b_p(\overline{Y}), \overline{X}) = h(\overline{X}, b_p(\overline{Y})), \end{aligned}$$

where one has used that $[X, Y] \in TH$ and that h is symmetric. Therefore we have shown that b is self-adjoint at any point and so we are done. ☐

Definition 6.1.12. *The null second fundamental form B of H with respect to L, is defined at each $p \in H$ by the bilinear form $B : T_pH/L \times T_pH/L \to \mathbb{R}$ with*

$$B(p)(\overline{X}, \overline{Y}) = B_p(\overline{X}, \overline{Y}) \coloneqq h(b_p(\overline{X}), \overline{Y}) = \langle \nabla_X L, Y \rangle.$$

[33] By this we mean that for $\overline{X} \in \mathfrak{X}^{\mathcal{L}}(H)$ and $\overline{\omega} \in \Omega^{\mathcal{L}}(H)$ one has $b(\overline{\omega}, \overline{X}) = \overline{\omega}(b(\overline{X}))$ with a slight abuse of notation.

Remark 6.1.13. By symmetry of h and self-adjointness of b, one obtains for $\overline{X}, \overline{Y} \in \mathfrak{X}^{\mathcal{L}}(H)$

$$B(\overline{X}, \overline{Y}) = h(b(\overline{X}), \overline{Y}) = h(\overline{X}, b(\overline{Y})) = h(b(\overline{Y}), \overline{X}) = B(\overline{Y}, \overline{X}),$$

so B is symmetric. Note also that, by definition, B is automatically smooth and therefore $B \in \mathcal{T}_2^0(TH/\mathcal{L})$

Definition 6.1.14. *H is called totally geodesic if $B = 0$.*

Down below we will show that the property of being totally geodesic holds the same geometric information as in the case for non-degenerate submanifolds. Consequently, the null second fundamental form truly behaves quite similarly to its usual definition (for the non-degenerate case), i.e. as the normal projection of the Levi-Civita connection.

Proposition 6.1.15. *If H is totally geodesic, then any geodesic in M starting out tangent to H remains in H initially.*

Proof. We show that the restriction of the Levi-Civita connection on M restricts to a linear connection on H, i.e. $\nabla : \mathfrak{X}(H) \times \mathfrak{X}(H) \to \mathfrak{X}(H)$. Let $X, Y \in \mathfrak{X}(H)$. Then

$$\langle \nabla_X Y, L \rangle = X \langle Y, L \rangle - \langle Y, \nabla_X L \rangle = -B(\overline{X}, \overline{Y}) = 0$$

and therefore $\nabla_X Y$ is indeed tangent to H. From this follows that vector fields along curves in H have the same induced covariant derivative with respect to the connections on H or M respectively. One can see this, for example, by looking at the local coordinate expression. To this end let $c : I \to H$ be smooth and $Z \in \mathfrak{X}(c)$ with $Z(t) \in T_{c(t)}H$ for all $t \in I$. If $\{\partial_1, \dots, \partial_{n+1}\}$ is an adapted coordinate frame with respect to H, we have (see [KS21, Prop. 1.3.27])

$$Z'(t) = \frac{dZ^i}{dt} \partial_i|_{c(t)} + Z^i(t) \nabla_{\dot{c}(t)} \partial_i$$

for $Z(t) = \sum_{i=1}^n Z^i(t) \partial_i|_{c(t)}$. It is clear that one obtains the same local expression if viewed with respect to the Levi-Civita connection on M instead. We may therefore conclude that a curve in H is a H-geodesic if and only if it is a M-geodesic.
Now, let $c : I \to M$ be a M-geodesic with $\dot{c}(0) \in T_{c(0)}H$. By solvability of the geodesic equation, there exists an H-geodesic $\overline{c} : J \to H$ with $\dot{\overline{c}}(0) = \dot{c}(0)$ and $\overline{c}(0) = c(0)$ and from the above and uniqueness it follows that $c|_J = \overline{c}$, which proves the claim. ☐

Example 6.1.16. Let us revisit two prior examples, which turn out to be totally geodesic.

(i) The null hypersurfaces in Minkowski space (Example 6.1.3 (ii)) are totally geodesic, because the null tangent vector field X is constant and therefore $\nabla_Y X = 0$ for any tangent vector Y.

(ii) Recall that the event horizon of the Schwarzschild spacetime is a null hypersurface with $L = \partial_v$ (see Example 6.1.3 (iii)). We want to calculate $\nabla_X \partial_v$ for $X \in \mathfrak{X}(H)$. Note that $X = X^v \partial_v + X^\theta \partial_\theta + X^\phi \partial_\phi$. Then

$$\nabla_X \partial_v = X^v \nabla_{\partial_v} \partial_v + X^\theta \nabla_{\partial_\theta} \partial_v + X^\phi \nabla_{\partial_\phi} \partial_v = X^v \frac{1}{4m} \partial_v,$$

where we used the fact that the only relevant non-vanishing Christoffel symbol turns out to be $\Gamma^v_{vv} = \frac{1}{4m}$ (cf. [MG10, Ch. 2.2.4]). Hence, for $X, Y \in \mathfrak{X}(H)$, we obtain

$$\langle \nabla_X \partial_v, Y \rangle = \frac{X^v}{4m} \langle \partial_v, Y \rangle = B(\overline{X}, \overline{Y}) = 0$$

since $\langle \partial_v, Y \rangle = 0$ for any $Y \in \mathfrak{X}(H)$.

Finally, we introduce the null mean curvature as follows.

Definition 6.1.17. *The null mean curvature (null expansion scalar) $\theta \in \mathcal{C}^\infty(H)$ on H with respect to L is defined by $\theta = \operatorname{tr} b$.*

Remark 6.1.18. Let $f \in \mathcal{C}^\infty(H)$ be strictly positive and let $\tilde{L} = fL$. By the proof of Proposition 6.1.11 we have $\tilde{b} = fb$. Therefore, it follows that also $\tilde{\theta} = f\theta$. This means the sign of θ is invariant under positive rescaling of the null tangent vector field on H.

The null mean curvature possesses a very nice geometric interpretation. To see this, let us look at a hypersurface S of H that is transversal to L (Figure 28), i.e. $L(p) \notin T_pS$ for any $p \in S \cap H$. Define the intersection $\Sigma = H \cap S$, which is itself a submanifold of M with codimension 2 (see [Lee13, Theorem 6.30] for a proof of this fact). Observe that for any $p \in \Sigma$ the tangent space $T_p\Sigma$ is contained, both, in T_pH and T_pS and consequently $T_p\Sigma \subseteq T_pH \cap T_pS$. Since the dimensions of these vector spaces are the same, one in fact has $T_p\Sigma = T_pH \cap T_pS$ and by transversality $L(p) \notin T_p\Sigma$ and $T_p\Sigma$ can only contain the spacelike direction, which are all orthogonal to $L(p)$.

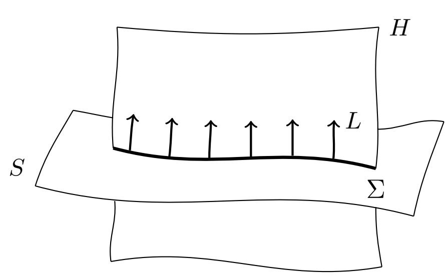


Figure 28: Transversal intersection between H and S.

In this setting, the null mean curvature θ on $H \cap \Sigma$ is related to the divergence of the null vector field L with respect to Σ. To show this, we will follow [Gal14, p. 27].

Proposition 6.1.19. *Let S be a hypersurface of M which is transversal to L. Then*

$$\theta = \operatorname{div}_\Sigma L,$$

where $\operatorname{div}_\Sigma L$ *is the divergence of L with respect to Σ.*

Proof. Since this is a local claim, let $p \in \Sigma$ and let $\{e_1, \dots, e_{n-1}\}$ be an orthonormal basis of $T_p\Sigma$. It follows that $\{\overline{e_1}, \dots, \overline{e_{n-1}}\}$ is an orthonormal basis for T_pH/L. By definition of the trace, one obtains

$$\theta(p) = \operatorname{tr} b\,(p) = \sum_{i=1}^{n-1} h(b(\overline{e_i}), \overline{e_i}) = \sum_{i=1}^{n-1} B(\overline{e_i}, \overline{e_i}) = \sum_{i=1}^{n-1} \langle \nabla_{e_i} L, e_i \rangle = \operatorname{div}_\Sigma L(p).$$

Since p was arbitrary, we conclude $\theta = \operatorname{div}_\Sigma L$. □

Remark 6.1.20. The divergence operator has the geometric interpretation of measuring the change of volume of Σ along the integral curves of L (see e.g. [Lee13, Ch. 16] and [Spi99, Ch. 9]). Therefore, one may conclude, using the above Proposition 6.1.19, that θ measures the overall expansion of the null generators of H. For example, in Minkwoski space, the future null cone $\partial I^+(p) \setminus \{p\}$ (resp. past null cone $\partial I^-(p) \setminus \{p\}$) has $\theta > 0$ (resp. $\theta < 0$) (see [Are18]), which corresponds to the growth (decline) of the volume of the cross sections along the null generators towards the future.

6.2 Null Generators and the Raychaudhuri Equation

In this second subsection, we take a closer look at the null generators of H. In particular, we want to establish the *Raychaudhuri equation*, which describes the change of θ along these null geodesics and which also plays a crucial role in proving the Incompleteness Theorem. As an additional application, we will see — assuming the underlying spacetime satisfies the so-called *null energy condition*: $\operatorname{Ric}(X, X) \geq 0$ for any null vector X — that this equation may be used to show that the null mean curvature is non-negative, provided the null generators of H exist for sufficiently long times or are in fact complete. We follow [Gal14] as the main source for this subsection.

As before, for the remainder of this section, H will denote a null hypersurface of a spacetime M with null tangent vector field L which spans the null bundle $\mathcal{L}$. Furthermore let $\eta : I \to M$ be a future directed affinely parametrized null generator of H. We denote by $\mathfrak{X}^{\mathcal{L}}(\eta)$ the set of mod $\mathcal{L}$ vector fields along η, i.e. if $\mathcal{Z} \in \mathfrak{X}^{\mathcal{L}}(\eta)$ then $\mathcal{Z}$ is a smooth map $\mathcal{Z} : I \to TH/\mathcal{L}$ such that $\pi^{\mathcal{L}} \circ \mathcal{Z} = \eta$, i.e.

$$\begin{tikzcd} TH/\mathcal{L} \arrow[r, "\pi^{\mathcal{L}}"] & H \\ I \arrow[u, "\mathcal{Z}"] \arrow[ur, dashed, "\eta"'] & \end{tikzcd}$$

commutes, where $\pi^{\mathcal{L}} : TH/\mathcal{L} \to H$ is the bundle projection of $TH/\mathcal{L}$. Note that there (at least locally) exists $Z \in \mathfrak{X}(\eta)$ such that $\overline{Z}(t) = \mathcal{Z}(t)$ for all $t \in I$ so each $\mathcal{Z}$ has a smooth local representative.
We may now define an induced covariant derivative for mod $\mathcal{L}$ vector fields along a null generator of H.

Definition 6.2.1. *Let $\eta : I \to H$ be a future directed affinely parametrized null generator of H and $\mathcal{Z} \in \mathfrak{X}^{\mathcal{L}}(\eta)$. We define the covariant derivative of Z as*

$$\mathcal{Z}'(t) = \overline{Z'(t)},$$

where Z' is the regular covariant derivative for a (local) representative $Z \in \mathfrak{X}(\eta)$ of $\mathcal{Z}$.

The above notion is well-defined. To see this, let Y be another representative of $\mathcal{Z}$, i.e. we have $Z = Y + f(L \circ \eta)$ with $f \in \mathcal{C}^\infty(I)$. Then $Z'(t) = Y'(t) + f'(t)(L \circ \eta)(t) + f(t)(L \circ \eta)'(t)$. Note now that $(L \circ \eta)'(t) = \nabla_{\dot\eta(t)} L = 0$ since $\dot\eta = L$. Therefore we have $\overline{Z'} = \overline{Y'}$.
Our next goal is to understand how the null Weingarten map changes along the null generator η. To this end, we want to have some notion of taking the derivative b' along η. In order to do this, we first have to define the notion of tensor fields along η. This can be done in the same fashion as above for vector fields. In fact, we have the following generalization

Definition 6.2.2. *A mod $\mathcal{L}$ (k,s)-tensor field along $\eta : I \to H$ is a smooth map $A : I \to T^k_s(TH/\mathcal{L})$ such that $\pi^{\mathcal{L}}_{(k,s)} \circ A = \eta$, i.e.*

$$\begin{tikzcd} T^k_s(TH/\mathcal{L}) \arrow[r, "\pi^{\mathcal{L}}_{(k,s)}"] & H \\ I \arrow[u, "A"] \arrow[ur, dashed, "\eta"'] & \end{tikzcd}$$

commutes, where $\pi^{\mathcal{L}}_{(k,s)}$ is the vector bundle projection of $T^k_s(TH/\mathcal{L})$. We denote the set of mod $\mathcal{L}$ (k,s) tensor fields along η by $\mathcal{T}^{\mathcal{L}}_{(k,s)}(\eta)$.

It is quite easy to construct explicit examples for this. For instance, let $A \in \mathcal{T}^k_s(TH/\mathcal{L})$, then it follows $A \circ \eta \in \mathcal{T}^{\mathcal{L}}_{(k,s)}(\eta)$, since it is a composition of smooth functions and the above diagram commutes because $A(\eta(t)) \in T^k_s(T_{\eta(t)}H/L)$. We therefore may now define the *Weingarten*

map along η by $b \circ \eta \in \mathcal{T}^{\mathcal{L}}_{(1,1)}(\eta)$, which we will sloppily denote by b for notational convenience. Note that $t \mapsto b(t)$ therefore forms a one-parameter family of maps $b(t)$, each of which is called the Weingarten map based on $\eta(t)$ with respect to the null tangent vector field L.
One can now extend the induced covariant derivative from Definition 6.2.1 to this one-parameter family of Weingarten maps $t \mapsto b(t)$ by demanding a natural product rule to hold. In this way we obtain a notion of derivative for the null Weingarten map b along η as required.

Definition 6.2.3. *Let $X \in \mathfrak{X}(\eta)$. The induced covariant derivative b' of $t \mapsto b(t)$ along η is defined by the Leibniz rule, i.e.*

$$b'(\overline{X}) = b(\overline{X})' - b(\overline{X'}),$$

where $b(\overline{X})'$ is the usual derivative of $b(\overline{X}) : I \to \mathbb{R}$.

Remark 6.2.4. One could actually generalize this derivative to general tensor fields along η in the same way, i.e. by requiring a Leibniz rule to hold. Note that by using *Willmore's* Theorem (cf. [Kun22a, Thm. 4.2.3]) it follows that this differential operator indeed exists and is unique. But since we do not explicitly need this here, we avoid getting into the details and also refer to [BEE96, Chapter 10.3].

Following [Kup87, Chapter 4] and [KS23, Proposition 3.11.2], one may establish the following technical Lemma.

Lemma 6.2.5. *Let M be globally hyperbolic and H a null hypersurface. Let L be the null tangent vector field on H and η one of the null generators of H. Then there exists an open neighbourhood $U \subseteq H$ of $\eta(t_0)$ for any $t_0 \in I$ such that we can scale L to be a local geodesic null vector field, i.e. $\nabla_L L = 0$ on U and with $L = \dot{\eta}$ on η.*

Proof. Let $t_0 \in I$. Pick $x, y \in M$ such that $x \ll \eta(t_0) \ll y$. Then $U := I^+(x) \cap I^-(y)$ is an open neighborhood of $\eta(t_0)$ that is itself globally hyperbolic. From this follows[34] that one can find a spacelike hypersurface S in H that is intersected by the generators η exactly once in $H \cap U$. Similarly to the proof of 5.2.12 (ii) one may now assume that L is complete in $U \cap H$ and that $(t, p) \in \mathbb{R} \times S \mapsto \mathrm{Fl}^L_p(t) \in U \cap H$ is a homeomomorphism. We will show that this map is in fact a diffeomorphism. For this let $(t_0, p_0) \in \mathbb{R} \times S$ and define $\varphi : p \mapsto \mathrm{Fl}^L_{-t_0}(p)$. Observe that φ is a diffeomorphism as it is smooth with smooth inverse $q \mapsto \mathrm{Fl}^L_{t_0}(q)$. Our goal is to demonstrate that $T_{(t_0,p_0)}(\varphi \circ \mathrm{Fl}^L)$ is surjective, from which the claim follows. Note that $(\varphi \circ \mathrm{Fl}^L)(t, p) = \mathrm{Fl}^L_{t-t_0}(p)$ and let $v \in T_{p_0}S$. Define a smooth curve $\alpha : I \to S$ with $\alpha(0) = p_0$ and $\dot{\alpha}(0) = v$ and let $\gamma : I \to \mathbb{R} \times S$ such that $\gamma(t) = (t_0, \alpha(t))$. It follows that $\gamma(0) = (t_0, p_0)$,

[34] See [Kup87, Prop. 22].

$\dot{\gamma}(0) = (0, v)$ and

$$v = \dot{\alpha}(0) = \frac{d}{dt}\bigg|_{t=0} \left(\varphi \circ \mathrm{Fl}^L \circ \gamma\right)(t) = T_{(t_0,p_0)}\left(\varphi \circ \mathrm{Fl}^L\right)(\dot{\gamma}(0)) \in \mathrm{im}\left(T_{(t_0,p_0)}\left(\varphi \circ \mathrm{Fl}^L\right)\right)$$

since $\varphi \circ \mathrm{Fl}^L \circ \gamma(t) = \mathrm{Fl}^L_{t-t_0}(\gamma(t)) = \mathrm{Fl}^L_0(\alpha(t)) = \alpha(t)$. Also for $\alpha(t) := (t_0 + t, p_0)$ one similarly has

$$T_{(t_0,p_0)}\left(\varphi \circ \mathrm{Fl}^L\right)(\dot{\alpha}(0)) = \frac{d}{dt}\bigg|_{t=0} \left(\varphi \circ \mathrm{Fl}^L\right)(\alpha(t)) = \frac{d}{dt}\bigg|_{t=0} \mathrm{Fl}^L_t(p_0) = \dot{\eta}_{p_0}(0) \notin T_{p_0}S,$$

where η_{p_0} is the null generator through p_0. We have therefore demonstrated that $T_{p_0}U = \mathrm{span}\left(\{\dot{\eta}_{p_0}(0)\} \cup T_{p_0}S\right) \subseteq \mathrm{im}\, T_{(t_0,p_0)}\left(\varphi \circ \mathrm{Fl}^L\right)$ and as a result $T_{(t_0,p_0)}\left(\varphi \circ \mathrm{Fl}^L\right)$ must indeed be surjective.
We may now identify $U \cap H \cong \mathbb{R} \times S$ and can simply rescale L to be the velocity field of the null generators η on $U \cap H$, i.e. $L(t,p) := \frac{d}{dt}\mathrm{Fl}^L_t(p) = \dot{\eta}_p(t)$, which is indeed a smooth null tangent vector field on $H \cap U$. By construction one has $\nabla_L L = 0$ and $\dot{\eta} = L$ on $U \cap H$. $\square$

While the assumption of global hyperbolicity above is indeed very restrictive, it is not a significant matter for our later arguments because there we will generally consider globally hyperbolic spacetimes anyway. Consequently, we will assume (M, g) to be globally hyperbolic from now on for the rest of this chapter. One may now show that the family $t \mapsto b(t)$ on H satisfies a certain Riccati equation. For this, we define the curvature endomorphism $R : \mathfrak{X}^{\mathcal{L}}(\eta) \to \mathfrak{X}^{\mathcal{L}}(\eta)$ by

$$R(t)(\overline{X}) = \overline{R_{X\dot{\eta}(t)}\dot{\eta}(t)},$$

where on the right hand side R denotes the Riemann tensor on M. Note that we have a natural identification of R with a mod $\mathcal{L}$ $(1,1)$ tensor field along η, i.e. $R \in \mathcal{T}^{\mathcal{L}}_{(1,1)}(\eta)$. Following [Gal14, Prop. 6.3], we now show the following

Proposition 6.2.6. *The one-parameter family of Weingarten maps $t \mapsto b(t)$ satisfies the following Riccati equation*

$$b' + b^2 + R = 0. \tag{6.1}$$

Proof. This is a local claim, so we can work at a single point. Let $t_0 \in I$ and $p = \eta(t_0)$. From the previous Lemma 6.2.5, we know that we can scale L such that $\nabla_L L = 0$ on some neighbourhood U of p and also $L \circ \eta = \dot{\eta}$. Let $X \in T_pM$ and extend it to a vector field on U in such a way that $[X, L] = \nabla_X L - \nabla_L X = 0$ (shrinking U if necessary). One has

$$R_{XL}L = \nabla_X \nabla_L L - \nabla_L \nabla_X L - \nabla_{[X,L]}L = -\nabla_L \nabla_X L = -\nabla_L \nabla_L X.$$

Along η it therefore follows that $-R_{X\dot{\eta}}\dot{\eta} = \nabla_L \nabla_L X|_\eta$ and since $\nabla_L \nabla_L X|_\eta = \nabla_{\dot{\eta}} \nabla_{\dot{\eta}} X = X''$, we conclude that $X'' = -R_{X\dot{\eta}}\dot{\eta}$, i.e. $X \circ \eta$ is a Jacobi field. By definition of the null Weingarten map along η and its induced covariant derivative, one then obtains

$$\begin{aligned} b'(\overline{X}) &= b(\overline{X})' - b(\overline{X'}) = \overline{\nabla_X L}' - b(\overline{\nabla_L X}) \\ &= \overline{\nabla_L X}' - b(\overline{\nabla_X L}) = \overline{X''} - b(b(\overline{X})) \\ &= -\overline{R_{X\dot{\eta}}\dot{\eta}} - b^2(\overline{X}) = -R(\overline{X}) - b^2(\overline{X}) \end{aligned}$$

and the claim follows because $X \in T_pM$ was arbitrary. □

Definition 6.2.7. *The shear scalar $\sigma : I \to \mathbb{R}$ is defined as*

$$\sigma = \sqrt{\operatorname{tr} \hat{b}^2},$$

where $\hat{b} = b - \frac{1}{n-1}\theta \cdot \mathrm{id}$ is the trace free part of b.

We are now able to establish the Raychaudhuri equation, which relates the null mean curvature θ along η, i.e. $\theta \circ \eta$, to the Ricci curvature of M.

Proposition 6.2.8 (Raychaudhuri Equation)**.** *The null mean curvature along a future directed, affinely parametrized null generator $\eta : I \to H$ satisfies the following equation*

$$\theta' = -\operatorname{Ric}(\dot{\eta}, \dot{\eta}) - \sigma^2 - \frac{1}{n-1}\theta^2. \tag{6.2}$$

Proof. For the proof, we just have to take the trace of the Riccati equation (6.1). Note that

$$\operatorname{tr}\left(b' + b^2 + R\right) = \operatorname{tr} b' + \operatorname{tr} b^2 + \operatorname{tr} R,$$

so we can treat each term separately. For $\operatorname{tr} b'$ note that the trace is a bounded, linear operator and therefore $\operatorname{tr} b' = (\operatorname{tr} b)' = \theta'$. For the second term we have

$$\begin{aligned} \sigma^2 + \frac{1}{n-1}\theta^2 &= \operatorname{tr} \hat{b}^2 + \frac{1}{n-1}\theta^2 = \operatorname{tr}\left(b^2 - \frac{2}{n-1}\theta \cdot b + \frac{1}{(n-1)^2}\theta^2 \cdot \mathrm{id}\right) + \frac{1}{n-1}\theta^2 \\ &= \operatorname{tr} b^2 - \frac{2}{n-1}\theta \ \operatorname{tr} b + \frac{1}{(n-1)^2}\theta^2 \ \operatorname{tr}(\mathrm{id}) + \frac{1}{n-1}\theta^2 \\ &= \operatorname{tr} b^2 - \frac{2}{n-1}\theta^2 + \frac{1}{(n-1)^2}\theta^2(n-1) + \frac{1}{n-1}\theta^2 \\ &= \operatorname{tr} b^2. \end{aligned}$$

For the last term let $t \in I$ and $\{\overline{e_1}, \ldots, \overline{e_{n-1}}\}$ be an orthonormal basis for $T_{\eta(t)}H/L$. We can define $e_{n+1}, e_n \in T_{\eta(t)}M$ with e_{n+1} timelike and $e_{n+1} + e_n = \dot{\eta}(t)$, such that $\{e_1, \ldots, e_{n+1}\}$ is an orthonormal basis of $T_{\eta(t)}M$. One then has

$$-\langle R_{e_{n+1}\dot{\eta}(t)}\dot{\eta}(t), e_{n+1}\rangle + \langle R_{e_n\dot{\eta}(t)}\dot{\eta}(t), e_n\rangle = -\langle R_{\dot{\eta}(t)-e_n,\dot{\eta}(t)}\dot{\eta}(t), \dot{\eta}(t) - e_n\rangle + \langle R_{e_n\dot{\eta}(t)}\dot{\eta}(t), e_n\rangle$$

$$
\begin{aligned}
&= -\langle R_{\dot\eta(t)\dot\eta(t)}\dot\eta(t), \dot\eta(t)\rangle + \langle R_{\dot\eta(t)\dot\eta(t)}\dot\eta(t), e_n\rangle + \langle R_{e_n\dot\eta(t)}\dot\eta(t), \dot\eta(t)\rangle \\
&\quad - \langle R_{e_n\dot\eta(t)}\dot\eta(t), e_n\rangle + \langle R_{e_n\dot\eta(t)}\dot\eta(t), e_n\rangle = 0,
\end{aligned}
$$

where we have used the skew-symmetry and skew-adjointness of the Riemann tensor. From this then follows

$$
\begin{aligned}
\operatorname{tr} R(t) &= \sum_{i=1}^{n-1} h\left(R(\overline{e_i}), \overline{e_i}\right) = \sum_{i=1}^{n-1} h\left(\overline{R_{e_i\dot\eta(t)}\dot\eta(t)}, \overline{e_i}\right) = \sum_{i=1}^{n-1}\langle R_{e_i\dot\eta(t)}\dot\eta(t), e_i\rangle \\
&= -\langle R_{e_{n+1}\dot\eta(t)}\dot\eta(t), e_{n+1}\rangle + \langle R_{e_n\dot\eta(t)}\dot\eta(t), e_n\rangle + \sum_{i=1}^{n-1}\langle R_{e_i\dot\eta(t)}\dot\eta(t), e_i\rangle \\
&= \sum_{i=1}^{n+1}\varepsilon_i\langle R_{e_i\dot\eta(t)}\dot\eta(t), e_i\rangle = \operatorname{Ric}(\dot\eta(t), \dot\eta(t)) = \operatorname{Ric}(\dot\eta, \dot\eta)(t),
\end{aligned}
$$

where $\varepsilon_i = \langle e_i, e_i\rangle$. Because $t \in I$ was arbitrary, we have $\operatorname{tr} R = \operatorname{Ric}(\dot\eta, \dot\eta)$, which finishes the proof. □

The Raychaudhuri equation may now be applied to obtain a quite powerful result on the boundedness of the null mean curvature. We follow [Gal14, Prop. 6.4].

Proposition 6.2.9. *Let M be a spacetime with $\operatorname{Ric}(X, X) \geq 0$ for any $X \in \mathfrak{X}(M)$ null. Further, let H be a null hypersurface in M with tangent null vector field L. If the null generators of L are future geodesically complete, then the null mean curvature θ of H is bounded below by zero, i.e. $\theta \geq 0$.*

Proof. Assume to the contrary that $\theta(p) < 0$ for some $p \in H$. Let $\eta : I \to H$ be the future directed, affinely parametrized null generator with $\eta(0) = p$. Consider the null mean curvature along η, i.e. $\theta = \operatorname{tr}(b \circ \eta)$. Using Raychaudhuri's equation (6.2) and the fact that $\operatorname{Ric}(\dot\eta, \dot\eta) \geq 0$ by assumption and also that $\sigma^2 \geq 0$, one obtains

$$
\theta' \leq -\frac{1}{n-1}\theta^2
$$

and therefore $\theta < 0$ for all $t \geq 0$. On the other hand, note that

$$
\frac{d}{dt}\left(\frac{1}{\theta}\right) = -\frac{1}{\theta^2}\theta' \geq \frac{1}{n-1},
$$

so in fact $\frac{1}{\theta(t)} < 0$ has to intersect 0 in finite parameter time $t_0 > 0$. But then $\theta \to \infty$ for $t \to t_0$, which contradicts the smoothness of θ. □

7 Focal Points and Null Geodesics

Back in Chapter 4.3, we introduced the theory of variation of causal curves and defined the notion of Jacobi fields. As a central result it was shown that timelike variations can be constructed for any causal curves unless it is a null pregeodesic (cf. Proposition 4.3.5 and Corollary 4.3.6). However, recall that this was not the whole story. In fact it is true that the existence of so called conjugate points along the curve is actually sufficient to find timelike paths between points on the curve. In this chapter, we want to address this issue in detail and extend our previously established results. In particular, we will define the notion of focal points, which can be viewed as a generalization of conjugate points and are defined in the setting of geodesics emanating orthogonally from some submanifold. We will see how such points are connected to the exponential map and show that their existence along null geodesics enables us to find timelike variations. It therefore comes to no surprise that we need to concern ourselves with more theory of variational analysis. Our arguments will primarily require the machinery developed in Chapter 3 and aim at gaining insights into the underlying geometric constructions. A slightly different but more common approach using *Morse Index Theory* can be found in [BEE96] or [O'N83]. The main sources for this chapter are [O'N83], [Bä20] and [KS23].

As before, let (M, g) be a spacetime with $\dim M = n+1$ and consider a geodesic $c : I \to M$. Recall that a Jacobi field is a vector field $J \in \mathfrak{X}(c)$ along c which satisfies the *Jacobi equation*

$$\frac{\nabla^2 J}{dt^2} = R_{J\dot{c}}\dot{c}.$$

In particular, any Jacobi field J then defines a geodesic variation of c (cf. Proposition 4.3.9). The simplest example are the so called *trivial Jacobi fields*, which are a family of vector fields along c of the form $J(t) = (\lambda t + \mu)\dot{c}(t)$ with $\lambda, \mu \in \mathbb{R}$. Since $R_{\dot{c}\dot{c}}\dot{c} = 0$, these types of vector fields are indeed Jacobi fields and they correspond to a geodesic variation in the tangential direction of $\dot{c}$ (thus the term trivial). Of course it is much more interesting to look at geodesic variations in the normal direction. Following [O'N83, Lemma 8.7], we can make the following observation

Lemma 7.0.1. *Let $c : I \to M$ be a geodesic and $J \in \mathfrak{X}(c)$ a Jacobi field. Then the following are equivalent*

(i) $J \perp c$.

(ii) *There exists two points $t_1, t_2 \in I$ with $t_1 \neq t_2$ such that $J(t_1) \perp c$ and $J(t_2) \perp c$.*

(iii) *There exists $t_1 \in I$ such that $J(t_1) \perp c$ and $J'(t_1) \perp c$.*

Proof. Because $\ddot{c} = 0$ we have

$$\frac{d^2}{dt^2}\langle J, \dot{c}\rangle = \left\langle \frac{\nabla^2 J}{dt^2}, \dot{c} \right\rangle = \langle R_{J\dot{c}}\dot{c}, \dot{c}\rangle = 0$$

by skew-adjointness of the Riemann tensor. Consequently, when integrating two times, we obtain $\langle J, \dot{c}\rangle = \lambda t + \mu$ for $\lambda, \mu \in \mathbb{R}$. Observe now that $J \perp c$ holds true if and only if $\lambda = \mu = 0$. In (ii), $\lambda t_1 + \mu = 0$ and $\lambda t_2 + \mu = 0$ imply the same. Finally, also for (iii) one has $\lambda t_1 + \mu = 0$ and $\lambda = 0$ and therefore each condition is equivalent and the claim follows. □

Without further ado, let us introduce the notion of conjugate points.

Definition 7.0.2. *Let $c : I \to M$ be a geodesic. Two points $c(t_1), c(t_2)$ with $t_1, t_2 \in I$ and $t_1 \neq t_2$ are called conjugate along c if there exists a nonzero Jacobi field $J \in \mathfrak{X}(c)$ such that $J(t_1) = 0$ and $J(t_2) = 0$.*

Note that the previous Lemma implies that the Jacobi fields in the above definition cannot be trivial Jacobi fields. A lot of things could be said about conjugate points, but our immediate goal will be to further generalize this notion. For this, instead of looking at curves between two points we replace the inital point by some submanifold, which is sometimes called *endmanifold*. This enables us to consider variations that are more flexible, i.e. where the initial point does not necessarily have to stay fixed, though it still has to be part of the endmanifold. In particular, we are interested in variations of geodesics which emanate orthogonally from some semi-Riemannian submanifold P (see Figure 29).

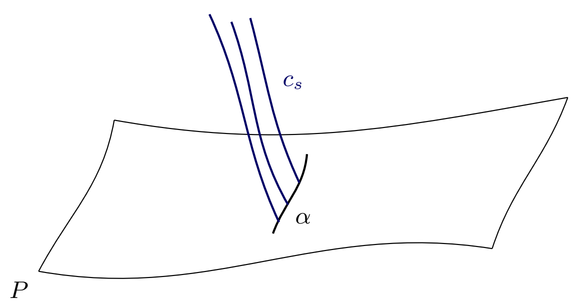


Figure 29: Geodesics c_s emanating orthogonally from a semi-Riemannian submanifold P. The black line $\alpha \coloneqq \frac{\partial c_s}{\partial s}\big|_{t=0}$ is the smooth transversal curve through the initial points of c_s.

One may now look at the family of Jacobi fields that yield geodesic variations that remain orthogonal to P, as seen in the previous figure. We follow [O'N83, Prop. 28] and [KS23, Lemma 3.2.12].

Proposition 7.0.3. *Let $P \subseteq M$ be a semi-Riemannian submanifold and let $c : I \to M$ be a geodesic normal to P, i.e. $c(t_0) \in P$ and $\dot{c}(t_0) \perp P$ with $t_0 \in I$. A Jacobi field $J \in \mathfrak{X}(c)$ is the variational vector field of a geodesic variation c_s of c with $c_s(t_0) \in P$ and $\dot{c}_s(t_0) \in T_{c(t_0)}P^{\perp}$*

for all small enough s if and only if $J(t_0) \in T_{c(t_0)}P$ and

$$\tan\left(\frac{\nabla J}{dt}(t_0)\right) = \tilde{\mathrm{I\!I}}(J(t_0), \dot{c}(t_0))$$

for $\tilde{\mathrm{I\!I}}$ from definition 3.2.9.

Proof. Without loss of generality set $t_0 = 0$ and $p := c(0)$.
" $\Rightarrow$ " By assumption there is a geodesic variation $c_s : I \times (-\epsilon, \epsilon) \to M$ with $\epsilon > 0$ and $J = \frac{\partial c_s}{\partial s}\big|_{s=0}$ such that $c_s(0) \in P$ and $\dot{c}_s(0) \in T_pM^\perp$ for all $s \in (-\epsilon, \epsilon)$. Note that if $X \in \mathfrak{X}(P)$, then for all s one has

$$\langle X(c_s(0)), \dot{c}_s(0)\rangle = 0.$$

Using Proposition 3.2.6, Lemma 2.1.6 (i) and Proposition 3.2.10, we calculate

$$\begin{aligned}
0 &= \frac{d}{ds}\bigg|_{s=0} \langle X(c_s(0)), \dot{c}_s(0)\rangle = \left\langle \frac{\nabla}{ds}\bigg|_{s=0} X(c_s(0)), \dot{c}(0)\right\rangle + \left\langle X(p), \frac{\nabla}{\partial s}\frac{\partial c_s}{\partial t}\bigg|_{s=t=0}\right\rangle \\
&= \left\langle \frac{\cancel{\nabla}}{ds}\bigg|_{s=0} X(c_s(0)) + \mathrm{I\!I}_p(J(0), X(p)), \dot{c}(0)\right\rangle + \left\langle X(p), \frac{\nabla}{\partial s}\frac{\partial c_s}{\partial t}\bigg|_{s=t=0}\right\rangle \\
&= \langle \mathrm{I\!I}_p(J(0), X(p)), \dot{c}(0)\rangle + \left\langle X(p), \frac{\nabla}{\partial t}\frac{\partial c_s}{\partial s}\bigg|_{s=t=0}\right\rangle \\
&= \langle \mathrm{I\!I}_p(J(0), X(p)), \dot{c}(0)\rangle + \left\langle X(p), \frac{\nabla J}{dt}(0)\right\rangle = -\langle \tilde{\mathrm{I\!I}}_p(J(0), \dot{c}(0)), X(p)\rangle + \left\langle \frac{\nabla J}{dt}(0), X(p)\right\rangle \\
&= -\left\langle \tilde{\mathrm{I\!I}}_p(J(0)), \dot{c}(0)) - \frac{\nabla J}{dt}(0), X(p)\right\rangle = \left\langle \tan\left(\frac{\nabla J}{dt}(0)\right) - \tilde{\mathrm{I\!I}}_p(J(0), \dot{c}(0)), X(p)\right\rangle.
\end{aligned}$$

Since X was arbitrary, the claim follows.
" $\Leftarrow$ " Let $\sigma : (-\epsilon, \epsilon) \to P$ be a smooth curve with $\dot{\sigma}(0) = J(0)$. Exactly as in the proof of Proposition 4.3.9, the map

$$c_s(t) := \exp_{\sigma(s)}(tX(s))$$

with $X \in \mathfrak{X}(\sigma)$ such that $X(s) \in T_{\sigma(s)}P^\perp$ for all s, defines a geodesic variation that satisfies $c_s(0) = \sigma(s) \in P$, $\dot{c}_s(0) = X(s) \in T_{\sigma(s)}P^\perp$ and $\frac{\partial}{\partial s}\big|_{s=0} c_s(t) = J(t)$ if $X(0) = \dot{c}(0)$ and $\frac{\nabla X}{ds}(0) = \frac{\nabla J}{dt}(0)$. Consequently, all that we need to show is that there exists such an X. For this, consider the normal parallel transport[35] of $\dot{c}(0)$ along σ denoted by $U \in \mathfrak{X}(\sigma)$. Naturally one has $U(s) \in T_{\sigma(s)}P^\perp$ and $\frac{\nabla^\perp U}{ds} = 0$ for all s. Furthermore, define $V \in \mathfrak{X}(\sigma)$ to be the normal parallel transport of $\mathrm{nor}\left(\frac{\nabla J}{dt}(0)\right)$ along σ. Now simply set

$$X(s) := U(s) + sV(s).$$

[35] Recall Lemma 3.2.13 and Proposition 3.2.14.

Note that $X(0) = U(0) = \dot{c}(0)$. It remains to show that $\frac{\nabla X}{ds}(0) = \frac{\nabla J}{dt}(0)$. One has

$$\begin{aligned}
\operatorname{nor}\left(\frac{\nabla X}{ds}(0)\right) &= \operatorname{nor}\left(\frac{\nabla U}{ds}(0) + V(0)\right) = \operatorname{nor}\left(\frac{\nabla U}{ds}(0)\right) + \operatorname{nor}(V(0)) \\
&= \frac{\nabla^{\perp} U}{ds}(0) + \operatorname{nor}\left[\operatorname{nor}\left(\frac{\nabla J}{dt}(0)\right)\right] = \operatorname{nor}\left(\frac{\nabla J}{dt}(0)\right).
\end{aligned}$$

For the tangential component we make the observation that

$$\tan\left(\frac{\nabla X}{ds}(0)\right) = \tan\left(\frac{\nabla U}{ds}(0) + V(0)\right) = \frac{\nabla U}{ds}(0)$$

because $V \perp P$. Let $Y \in \mathfrak{X}(\sigma)$ such that $\operatorname{nor}(Y) = 0$. Using the above equation and the fact that $U \perp P$, we calculate

$$\begin{aligned}
\left\langle \tan\left(\frac{\nabla X}{ds}(0)\right), Y(0)\right\rangle &= \left\langle \frac{\nabla U}{ds}(0), Y(0)\right\rangle = \left.\frac{d}{ds}\right|_{s=0} \langle U, Y\rangle - \left\langle U(0), \frac{\nabla Y}{ds}(0)\right\rangle \\
&= -\left\langle U(0), \operatorname{nor}\left(\frac{\nabla Y}{ds}(0)\right)\right\rangle = -\left\langle U(0), \mathrm{II}_p(\dot{\sigma}(0), Y(0))\right\rangle \\
&= -\langle \dot{c}(0), \mathrm{II}_p\left(J(0), Y(0)\right)\rangle = \langle \tilde{\mathrm{II}}_p(J(0), \dot{c}(0)), Y(0)\rangle \\
&= \left\langle \tan\left(\frac{\nabla J}{dt}(0)\right), Y(0)\right\rangle,
\end{aligned}$$

where the last equation follows by assumption. Since Y was arbitrary, one can indeed conclude that $\tan\left(\frac{\nabla X}{ds}(0)\right) = \tan\left(\frac{\nabla J}{dt}(0)\right)$ and consequently we are done. $\square$

Definition 7.0.4. *A Jacobi field which satisfies Proposition 7.0.3 is called P-Jacobi field on c.*

One may now define the notion of a focal point of a submanifold along some geodesic.

Definition 7.0.5. *Let $P \subseteq M$ be a semi-Riemannian submanifold and let $c : I \to M$ be a geodesic normal to P. A point $c(t)$ with $t \in I$ is a focal point of P along c if there exists a nonzero P-Jacobi field $J \in \mathfrak{X}(c)$ with $J(t) = 0$.*

Remark 7.0.6. (i) By Lemma 7.0.1 the P-Jacobi field in the definition above, must be orthogonal to c everywhere since $J(t_0) \perp \dot{c}(t_0)$ and $J(t) \perp \dot{c}(t)$ at the focal point.

(ii) Focal points indeed generalize conjugate points in the sense that if we consider $P = \{p\}$ for $p = c(t_0)$, one recovers definition 7.0.2.

(iii) Let k be the dimension of P. As solutions to a second order linear ordinary differential equation, the P-Jacobi fields form a vector space denoted by $\mathcal{J}_P$. By the previous Proposition, its dimension must be equal to $k + (n + 1 - k) = n + 1$ since one has k degrees of freedom of $J(0) \in T_pP$ and $n + 1 - k$ of $\tan(J'(0)) = \tilde{\mathrm{II}}_p(J(0), \dot{c}(0))$. Assume now that $c(t_0)$ is a focal point of P along c. The *order* of this focal point is defined as

$$\mu = \dim\{J \in \mathcal{J}_P \mid J(t_0) = 0\}$$

It follows that $\mu \leq n$ by the above and the fact that the trivial Jacobi field $J(t) = t\dot{c}$ is a P-Jacobi field that is non-vanishing at t_0.

Example 7.0.7. Let $M = \mathbb{R}^{1,n}$ and consider $P = S^n$ defined on the level set $x^0 = 0$ as shown in Figure 30. There are two different future null directions normal to P, one pointing inwards and the other one outwards. It follows that P has two focal points: At the point $(\sqrt{2}, 0, \ldots, 0)$, in the future of P, where the inward pointing geodesics cross and, by symmetry, at the point $(-\sqrt{2}, 0, \ldots, 0)$, corresponding to the outward pointing geodesics which meet in the past.

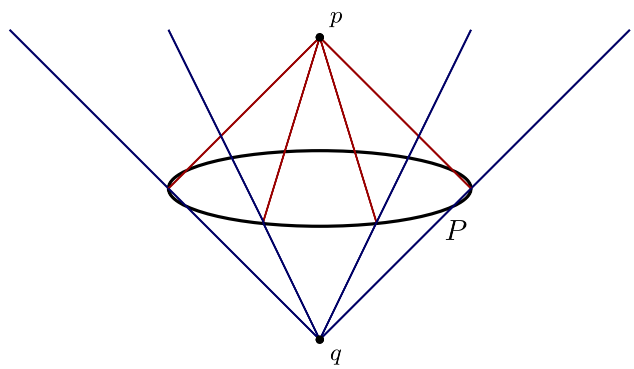


Figure 30: Two different families of null geodesics in $\mathbb{R}^{1,2}$ normal to a sphere $P = S^1$. The blue curves are outward pointing and the red curves are inward pointing. P has two focal points p, q.

The above example suggests that one should intuitively think of focal points as an almost meeting point of nearby geodesics. This in turn should have an effect on the behavior of the exponential map. We want to demonstrate that this is indeed the case. For this, one needs the subsequent Lemma that establishes a relation between Jacobi fields and the exponential map. Based on [O'N83, Prop. 8.6] and [KS21, Thm. 2.1.21] we have the following.

Lemma 7.0.8. *Let $p \in M$ and $x \in T_pM$. Further, let $c_x : [0,1] \to M$ be the radial geodesic $c_x(t) = \exp_p(tx)$. If $v \in T_x(T_pM)$, then*

$$T_x \exp_p(v) = V(1),$$

where $V \in \mathfrak{X}(c)$ is the unique Jacobi field of c such that $V(0) = 0$ and $V'(0) = v$.

Proof. Using the identification $T_x(T_pM) \cong T_pM$, one defines the map

$$\alpha(t, s) = \exp_p\left(t(x + sv)\right) = c_{x+sv}(t).$$

We want to show that α is a geodesic variation of c_x. In order to do this we first need to show that the map is defined on the product of two intervals. By continuity of $f : (t, s) \mapsto t(x+sv)$, the set $[0, 1]\times\{0\}$ maps into the open[36] domain D_p of $\exp_p$. Thus, there exist $\epsilon > 0$ and $\delta > 0$ such that α is defined on $(-\delta, 1 + \delta) \times (-\epsilon, \epsilon)$. Furthermore, α is smooth and $\alpha(t, 0) = c_x(t)$, so it is a geodesic variation of c_x following from the first equation.
The variation vector field $V \in \mathfrak{X}(c)$, which is a Jacobi field by Proposition 4.3.9, is given by

$$\begin{aligned} V(t) = \left.\frac{\partial \alpha}{\partial s}\right|_{s=0}(t) &= T_{(t,0)}\alpha\left(\frac{\partial}{\partial s}\right) = T_{(t,0)}(\exp_p \circ f)\left(\frac{\partial}{\partial s}\right) \\ &= T_{tx}\exp_p\left(\frac{\partial f}{\partial s}(t,0)\right) = T_{tx}\exp_p(tv) \end{aligned}$$

and consequently

$$V(1) = T_x \exp_p(v).$$

Note also that the curve $s \mapsto \alpha(0, s) = \exp_p(0) = p$ is constant at p and thus

$$V(0) = \left.\frac{\partial}{\partial s}\right|_{s=0} \alpha(0, s) = 0.$$

Finally, by Lemma 2.1.6 (i), one has

$$V'(0) = \frac{\nabla}{\partial t}\frac{\partial \alpha}{\partial s}(0, 0) = \frac{\nabla}{\partial s}\frac{\partial \alpha}{\partial t}(0, 0).$$

Using the fact that $\frac{\partial f}{\partial t}(0, s) = x + sv$, the same calculation as in the proof of Lemma 4.3.3 then yields

$$V'(0) = \left.\frac{\nabla}{\partial s}\right|_{s=0}\left(\frac{\partial \alpha}{\partial t}(0, s)\right) = \left.\frac{\nabla}{\partial s}\right|_{s=0}(x + sv) = v.$$

$\square$

Before tackling focal points directly, it is useful to investigate the particular case of conjugate points. As in [O'N83, Prop. 10.10], one has

Lemma 7.0.9. *Let $p \in M$ and $c : [0, t_0] \to M$ be a geodesic with $c(0) = p$. The following are equivalent*

(i) *$c(t_0)$ is a conjugate point of p along c.*

(ii) *There exists a nontrivial geodesic variation α of c starting at p such that $\frac{\partial \alpha}{\partial s} = 0$.*

(iii) *$\exp_p : T_pM \to M$ is singular at $t_0\dot{c}(0)$, i.e. there exists a non-zero vector $v \in T_{t_0\dot{c}(0)}(T_pM)$ such that $d\exp_p(v) = T_{t_0\dot{c}(0)}\exp_p(v) = 0$.*

[36]See [KS21, Prop. 2.2.3].

Proof. By reparametrization we can have without loss of generality that $t_0 = 1$.
(ii) $\Rightarrow$ *(i)* Since α is a nontrivial geodesic variation with fixed endpoints, $c(1)$ is a conjugate point by definition.
(i) $\Rightarrow$ *(iii)* Let $J \in \mathfrak{X}(c)$ be a Jacobi field with $J(0) = J(1) = 0$. Let $v \in T_{\dot{c}(0)}(T_pM) \cong T_pM$ such that $v = J'(0)$. By Lemma 7.0.8 one must have

$$T_{\dot{c}(0)} \exp_p(v) = J(1) = 0.$$

Since J is non-trivial and $J(0) = 0 = J(1)$ if follows that $v = J'(0) \neq 0$ by uniqueness of Jacobi fields.
(iii) $\Rightarrow$ *(ii)* Let $v \in T_{\dot{c}(0)}(T_pM)$ be a singular value of $\exp_p$ at $\dot{c}(0)$. Define the map

$$c_s(t, s) \coloneqq \exp_p \left(t(\dot{c}(0) + sv) \right).$$

As shown in 7.0.8, c_s defines a non-trivial geodesic variation with Jacobi field $V \in \mathfrak{X}(c)$ which satisfies $V(0) = 0$, $V(1) = T_{\dot{c}(0)} \exp_p(v) = 0$ and $V'(0) = v$. □

Remark 7.0.10. In any normal neighbourhood of a point p, the exponential map $\exp_p$ is a diffeomorphism and consequently cannot exhibit singular values. Lemma 7.0.9 therefore implies that geodesics cannot have conjugate points as long as the geodesics are sufficiently short.

Let $P \subseteq M$ be a semi-Riemannian submanifold with $\dim P = k$. We want to establish a connection between the existence of focal points and singularities of the exponential map in the same spirit to the above result. However, since the P-Jacobi fields give variations where the initial point in P can change, we need something more "global" than the exponential map at some $p \in P$. To this end, recall that one has the *global exponential map*

$$\begin{aligned} \exp : \mathcal{D} \subseteq TM &\to M \\ (p, v) &\mapsto \exp(p, v) = \exp_p(v), \end{aligned}$$

which is smooth and where $\mathcal{D}$ is its maximal domain, which is an open[37] subset of TM. Furthermore, since we are at the moment considering geodesics emanating orthogonally from P, one wants to consider the *normal exponential map* $\exp^\perp : \mathcal{D}^\perp \subseteq TP^\perp \to M$ given by the restriction of exp to the normal bundle $TP^\perp \subseteq TM$, i.e. $\mathcal{D}^\perp = \mathcal{D} \cap TP^\perp$. Naturally $\exp^\perp$ is a smooth map. In conclusion, our goal now is to relate focal points to singular behaviour of $\exp^\perp$.
In order to look for singularities of $\exp^\perp$ at some $(p, v) \in TP^\perp$, we need to work with the double tangent space $T_{(p,v)}(TP^\perp)$. To get a better picture let us recall some basics from differential geometry. Let $(U, \Psi = (x^1, \ldots, x^k, \xi^1, \ldots, \xi^{n-k}))$ be a vector bundle chart for

[37][KS21, Proposition 2.2.3].

$TP^\perp$ at (p,v). Consider a smooth curve $c : I \to TP^\perp$ with $c(0) = (p,v)$. In coordinates, one has

$$\Psi \circ c : t \mapsto (x^1(t), \ldots, x^k(t), \xi^0(t), \ldots, \xi^{n-k}(t)).$$

An element $x \in T_{(p,v)}(TP^\perp)$ can be written as[38] $x = [c]_{(v,p)}$, which is the equivalence class corresponding to $(\dot{x}^1(0), \ldots, \dot{x}^k(0), \dot{\xi}^0(0), \ldots, \dot{\xi}^{n-k}(0))$. Let $\pi^\perp : TP^\perp \to P$ be the bundle projection. Then

$$T_{(p,v)}\pi^\perp(x) = [\pi^\perp \circ c]_p$$

We have $\psi \circ \pi^\perp \circ c(t) = (x^1(t), \ldots, x^k(t))$, where $\psi = (x^1, \ldots, x^k)$ is the chart for P at p with $\Psi = (\psi, \xi^0, \ldots, \xi^{n-k})$ and where $x^i(t)$ is the shorthand notation for $x^i \circ c(t)$ with $i = 1, \ldots, k$. Therefore $[\pi^\perp \circ c]_p$ corresponds to $(\dot{x}^1(0), \ldots, \dot{x}^k(0))$. As a minor remark, it could then actually be concluded that $T\pi^\perp$ is the bundle projection of $TTP^\perp$ but we will not need this. However, we may now classify x in two different categories:

(1) x is tangential to the fiber $T_pP^\perp$, i.e. $\dot{x}^1(t) = \ldots = \dot{x}^k(t) = 0$. In this case $T_{(p,v)}\pi^\perp(x) = 0$ and x "points" only into the directions of $T_pP^\perp$.

(2) x has a horizontal component, thus the foot point in $TP^\perp$ changes along c. Here $T_{(p,v)}\pi^\perp(x) \neq 0$.

With this at hand and just as in [O'N83, Prop. 10.30], we may finally prove the following

Proposition 7.0.11. *Let $P \subseteq M$ be a semi-Riemannian submanifold. Let $c : [0, t_0] \to M$ be a geodesic orthogonal to P, i.e. $\sigma(0) \in P$ and $\dot{c}(0) \perp P$. The following are equivalent*

(i) $c(t_0)$ is a focal point of P along c.

(ii) There exists a nontrivial geodesic variation α of c with the variatonal vector field given by a P-Jacobi field and for which $\frac{\partial \alpha}{\partial s}(t_0, 0) = 0$.

(iii) $\exp^\perp : \mathcal{D}^\perp \subseteq TP^\perp \to M$ is singular at $t_0\dot{c}(0)$, i.e. there exists a non-zero vector $x \in T_{t_0\dot{c}(0)}(TP^\perp)$ such that $d\exp^\perp(x) = T_{t_0\dot{c}(0)}\exp^\perp(x) = 0$.

Proof. By reparametrization we can have without loss of generality that $t_0 = 1$.
$(i) \Leftrightarrow (ii)$ This follows from Proposition 7.0.3.
$(iii) \Rightarrow (ii)$ Let $x \in T_{\dot{c}(0)}(TP^\perp)$ be the point where $\exp^\perp$ is singular and let $p := c(0)$. There are then two cases as described previously. For the first, assume x is tangent to $T_pP^\perp$. This simply means $x \in T_{\dot{c}(0)}(T_pP^\perp) \cong T_pP^\perp$ and since $T_pP^\perp \subseteq T_pM$, Lemma 7.0.9 implies that $c(1)$ is a conjugate point of p along c and therefore a focal point. Secondly, assume that x has some horizontal component. Then $d\pi^\perp(x) = T_{\dot{c}(0)}\pi^\perp(x) \neq 0$. Let A be any smooth curve

[38] Cf. [Kun22a, Def. 2.1.6].

in $TP^\perp$ with initial velocity $A'(0) = x$ and $A(0) = \dot{c}(0)$. Define

$$c_s(t) \coloneqq \exp^\perp(tA(s)).$$

We show that this is the desired variation. For this note that the map is smooth and $t \mapsto c_s(t)$ is a geodesic for each fixed s. Furthermore

$$c_0(t) = \exp^\perp(tA(0)) = \exp^\perp(t\dot{c}(0)) = \exp_p(t\dot{c}(0)) = \sigma_{t\dot{c}(0)}(1) = \sigma_{\dot{c}(0)}(t),$$

where $\sigma_{\dot{c}(0)}(t)$ is the radial geodesic with $\sigma_{\dot{c}(0)}(0) = p$ and $\dot{\sigma}_{\dot{c}(0)}(0) = \dot{c}(0)$. Uniqueness of geodesics then implies that $c_0 = c$ and consequently c_s is a variation of c. Finally we have

$$\begin{aligned} \frac{\partial c_s}{\partial s}\bigg|_{(0,0)} &= \frac{\partial}{\partial s}\bigg|_{s=0} \exp\left(\pi^\perp(A(s)), 0\right) = \frac{\partial}{\partial s}\bigg|_{s=0} \pi^\perp(A(s)) \\ &= d\pi^\perp(A'(0)) = d\pi^\perp(x) \neq 0 \end{aligned}$$

and c_s is nontrivial.
$(ii) \Rightarrow (iii)$ Assume that α is a variation as in (ii) with $\frac{\partial \alpha}{\partial s}(1,0) = 0$ and let $\sigma(s) \coloneqq \alpha(0,s) \in P$. We define

$$\tilde{\alpha}(t,s) \coloneqq t\,\frac{\partial}{\partial t}\alpha(0,s) \in T_{\sigma(s)}P^\perp.$$

Since α is a geodesic variation, the same argument as above, using radial geodesics and uniqueness, then gives

$$\exp^\perp(\tilde{\alpha}(t,s)) = \exp^\perp(t\,\frac{\partial}{\partial t}\alpha(0,s)) = \alpha(t,s).$$

Consequently

$$d\exp^\perp\left(\frac{\partial}{\partial s}\tilde{\alpha}(1,0)\right) = T_{(1,0)}(\exp^\perp \circ \tilde{\alpha})\left(\frac{\partial}{\partial s}\right) = T_{(1,0)}\alpha\left(\frac{\partial}{\partial s}\right) = \frac{\partial \alpha}{\partial s}(1,0) = 0.$$

Note that $\frac{\partial}{\partial s}\tilde{\alpha}(1,0) \in T_{\dot{c}(0)}(TP^\perp)$, so if we can show that $\frac{\partial}{\partial s}\tilde{\alpha}(1,0) \neq 0$, the claim follows. Observe that the curve $w : s \mapsto \frac{\partial \alpha}{\partial t}(0,s)$ in $TP^\perp$ with $\pi^\perp(w(s)) = \sigma(s)$ has initial velocity

$$w'(0) = \frac{\partial}{\partial s}\frac{\partial \alpha}{\partial t}(0,0) = \frac{\partial}{\partial s}\tilde{\alpha}(1,0).$$

Let us distinguish two cases. First suppose that $\frac{\partial \alpha}{\partial s}(0,0) = \frac{\partial \sigma}{\partial s}(0) \neq 0$. Therefore, the foot point in $TP^\perp$ is changing along w and one must indeed have $\frac{\partial}{\partial s}\tilde{\alpha}(1,0) \neq 0$. On the other hand assume that $\frac{\partial \alpha}{\partial s}(0,0) = 0$, which implies that $c(1)$ is actually a conjugate point. But for a conjugate point we may then simply use Proposition 7.0.9 and reverse the argument in $(iii) \Rightarrow (ii)$ to infer that $\exp^\perp$ is singular. $\square$

The previous Proposition does not only provide the desired relationship between singular values of $\exp^\perp$ and focal points, but also confirms our intuition of focal points being almost meeting points of nearby geodesics. In particular, observe that if the orthogonal geodesic c is timelike or spacelike, continuity implies that a focal point is an almost meeting point of geodesics of the same causal character. This also remains to be true if c is null. Following [O'N83, Proposition 10.40], one has

Proposition 7.0.12. *Let P be a semi-Riemannian submanifold of M and let c be a null geodesic orthogonal to P. A P-Jacobi field $J \in \mathcal{J}_P$ is the vector field of a variation of c through null geodesics normal to P if and only if $J \perp \dot{c}$.*

Proof. Without loss of generality let $c : [0,T) \to M$ with $\dot{c}(0) \perp P$ and $p \coloneqq c(0) \in P$. Suppose first that c_s is a variation of c such that $t \mapsto c_s(t)$ is a null geodesic with $\dot{c}_s(0) \perp P$ and $c_s(0) \in P$ for any $s \in (-\epsilon, \epsilon)$. One then has $\langle \dot{c}_s, \dot{c}_s \rangle = 0$ and calculating the derivative with respects to s yields

$$0 = \frac{1}{2} \frac{\partial}{\partial s}\bigg|_{s=0} \langle \dot{c}_s, \dot{c}_s \rangle = \left\langle \frac{\partial}{\partial s}\bigg|_{s=0} \frac{\partial c_s}{\partial t}, \dot{c} \right\rangle = \left\langle \frac{\partial}{\partial t} \frac{\partial c_s}{\partial s}\bigg|_{s=0}, \dot{c} \right\rangle = \left\langle \frac{d}{dt} J, \dot{c} \right\rangle = \langle J', \dot{c} \rangle.$$

It follows that $J'(0) \perp c$ by the above and $J(0) \perp c$ by assumption. Lemma 7.0.1 then gives the claim.
Now suppose that $J \perp \dot{c}$. Let us recall the only if part of the proof of Proposition 7.0.3. There we explicitly constructed a geodesic variation of c by defining

$$c_s(t) \coloneqq \exp_{\sigma(s)}(tX(s)),$$

where $\sigma : (-\epsilon, \epsilon) \to P$ is a smooth curve on P with $\dot{\sigma}(0) = J(0)$ and $X \in \mathfrak{X}(\sigma)$ with $X(0) = \dot{c}(0)$ and $\frac{\nabla X}{ds}(0) = \frac{\nabla J}{dt}(0)$. Observe that this proof is also sufficient here as long as we are able to construct X in a way that it is null, $X(0) = \dot{c}(0)$ and such that $\operatorname{nor}(X'(0)) = \operatorname{nor}(J'(0))$. This follows from the fact that the geodesic variation c_s yields a P-Jacobi field $\tilde{J}$ with $\tilde{J}(0) = J(0)$ and $\operatorname{nor}(\tilde{J}(0)) = \operatorname{nor}(J(0))$, and therefore $\tilde{J} = J$ by the uniqueness of P-Jacobi fields. Consequently, one indeed finds that J generates the desired null geodesic variation.
The assumption $J \perp c$ implies that $\operatorname{nor}(J'(0)) \perp c$ since

$$0 = \frac{d}{dt}\bigg|_{t=0} \langle J, \dot{c} \rangle = \langle J'(0), \dot{c}(0) \rangle = \langle \operatorname{nor}(J')(0), \dot{c}(0) \rangle + \underbrace{\langle \tan(J')(0), \dot{c}(0) \rangle}_{=0}.$$

Let Λ be the lightcone at p, i.e. $\Lambda = \{ v \in T_p P^\perp \,|\, g(v,v) = 0 \} \subset T_p P^\perp$. We can identify $\operatorname{nor}(J'(0))$ with a vector at $\dot{c}(0)$ which is tangent to Λ. The reason for this is the following: The tangent space $T_{\dot{c}(0)}\Lambda$ is given by the kernel of the derivative of the map $f(v) \coloneqq g(v,v)$ (cf. [Kun22a, Thm. 2.1.1]). Now simply note that $T_{\dot{c}(0)} f(x) = 2g(\dot{c}(0), x)$ and also $g(\dot{c}(0), \operatorname{nor}(J'(0))) = 0$. Let z be a smooth curve in Λ with $z(0) = \dot{c}(0)$ and

$\dot{z}(0) = \mathrm{nor}(J'(0))$ (see Figure 31).

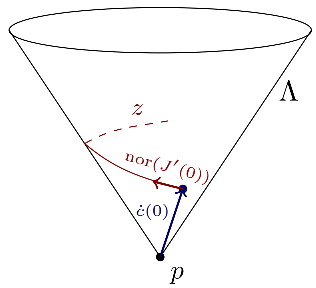


Figure 31: Smooth curve $z : (-\epsilon, \epsilon) \to \Lambda$ with $z(0) = \dot{c}(0)$ and $\dot{z}(0) = \mathrm{nor}(J'(0))$.

Now set $X(s) := \mathcal{P}^{\sigma}_{0,s}(z(s))$, the normal parallel transport of $z(s)$ along σ to $\sigma(s)$. Evidently X is a smooth null vector field along σ, and furthermore

$$X(0) = \mathcal{P}^{\sigma}_{0,0}(z(0)) = \mathrm{id}(\dot{c}(0)) = \dot{c}(0).$$

Let $k = \dim(P)$ and $\dim(M) = n$. Note that one can write $z(s) = z^i(s)e_i$, where $\{e_1, \ldots, e_{n-k}\} \subseteq T_pP^{\perp}$ is an orthonormal basis and z^i smooth. Consequently, $X(s) = z^i(s)E_i(s)$ where the E_i are the normal parallel transport of e_i respectively. Then

$$\mathrm{nor}(X'(0)) = X'(0) = \dot{z}^i(0)E_i(0) + z^i(0)\underbrace{E_i'(0)}_{=0} = \dot{z}^i(0)E_i(0) = \dot{c}^i(0)e_i = \mathrm{nor}(J'(0))$$

and we are done. □

To summarize, assuming that there is a focal point of P along a null geodesic c, which implies $J \perp c$, the above Proposition tells us that J generates a null geodesic variation. A focal point is therefore an almost meeting point of nearby geodesics with the same causal character, even in the null case.

For the final part of this section we want to discuss how focal points affect causality. The fact that radial geodesics are length maximizing curves between points[39] together with the breakdown of the exponential map when focal points are involved, raise the suspicion that null geodesics could be deformed into timelike curves in the presence of a focal point. This is indeed the case as we are going to show shortly. First however, we need to deal with another small technicality. Namely with a proof of the fact that if there are multiple focal points of a semi-Riemannian manifold P along the same geodesic c, that there still must exist a first one.

Lemma 7.0.13. *Let $P \subseteq M$ be a semi-Riemannian submanifold and $c : [0, b] \to M$ a geodesic orthogonal to P. If P has multiple focal points along c then there exists a well-defined first*

[39] Cf. [KS21, Cor. 2.3.11 & Rem. 2.4.7 (*ii*)].

focal point along c.

Proof. Let $\mathcal{F} := \{t \in (0,b] \,|\, c(t) \,\text{is a focal point of}\, P \,\text{along}\, c\}$. By assumption $\mathcal{F} \neq 0$, so let $\{t_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F}$ such that $t_n < t_m$ for $m < n$. Using the previous Proposition, there exists a corresponding sequence $\{x_n\}_{n\in\mathbb{N}}$ with $x_n \in T_{t_n\dot{c}(0)}(TP^\perp)$ such that $d\exp^\perp(x_n) = 0$ for all n. Now define $\tilde{t} := \inf \mathcal{F}$ and consider a local trivialization $\mathcal{U} \times \mathbb{R}^{n+1}$ of $TTP^\perp$ around $(\dot{c}(\tilde{t}), 0)$. By definition of $\tilde{t}$ and linearity of $d\exp^\perp$, we can assume without loss of generality that there exists a relatively compact subset $V \subset\subset \mathcal{U} \times \mathbb{R}^{n+1}$ containing $(\dot{c}(\tilde{t}), 0)$ and each x_n (otherwise consider rescaled subsequence of $\{x_n\}$, such that it is contained in some small Euclidean ball). By compactness of $\overline{V}$ there exists a subsequence $\{x_k\}_{k\in\mathbb{N}}$ which converges to some $x \in \overline{V}$. In particular $x \in T_{\tilde{t}\dot{c}(0)}(TP^\perp)$ because $\mathcal{U}$ can be arbitrarily small, i.e. if $x \in T_v(TP^\perp)$ with $v \neq \tilde{t}\dot{c}(0)$, we can shrink $\mathcal{U}$ such that $v \notin \mathcal{U}$. Taking the limit of $d\exp^\perp(x_n) = 0$, one obtains by continuity that $d\exp^\perp(x) = 0$. It follows by Proposition 7.0.11 that $c(\tilde{t})$ is a focal point of P along c, finishing the proof. □

Following [Bä20, Prop. 2.38] and [KS23, Prop. 3.2.17], we may now prove one of the main results of this chapter.

Theorem 7.0.14. *Let $P \subseteq M$ be a spacelike submanifold and $c : [0,b] \to M$ be a null geodesic orthogonal to P. If P has a focal point along c before $c(b)$, then there exists a timelike curve from P to $c(b)$ which is arbitrarily close to c.*

Proof. Set $q := c(b)$, $p := c(0)$ and let $c(t_0)$ with $t_0 \in (0,b)$ be the first focal point of P along c, which exists by Lemma 7.0.13. Let $J \neq 0$ be a P-Jacobi field along c with $J(t_0) = 0$ and $J(t) \neq 0$ for any $t \in (0,t_0)$.

Claim 1. *There exists $\delta \in (0, b - t_0)$ such that $J = fU$ on $[0, t_0 + \delta]$ for a spacelike unit vector field $U \in \mathfrak{X}(c)$ and $f \in \mathcal{C}^\infty$ with $f|_{(0,t_0)} > 0$ and $f|_{(t_0,t_0+\delta)} < 0$.*

Since J is a P-Jacobi field, Lemma 7.0.1 implies that $J \perp \dot{c}$ on $[0,b]$. But this does not exclude the possibility of J being tangential to $\dot{c}$, which is null by assumption. Consequently suppose there exists $t_1 \in (0,t_0)$ and $\lambda \in \mathbb{R}$ such that $J(t_1) = \lambda\dot{c}(t_1)$. Now set $\tilde{J}(t) := J(t) - \frac{\lambda t}{t_1}\dot{c}(t)$. This defines a P-Jacobi field

$$\tilde{J}(0) = J(0) \in T_pP,$$
$$\tan\left(\frac{\nabla\tilde{J}}{dt}(0)\right) = \tan\left(\frac{\nabla J}{dt}(0)\right) = \tilde{\mathrm{I\!I}}\left(J(0), \dot{c}(0)\right) = \tilde{\mathrm{I\!I}}\left(\tilde{J}(0), \dot{c}(0)\right).$$

We have therefore reached a contradiction as $\tilde{J}(t_1) = 0$ implies that $c(t_1)$ is a focal point before $c(t_0)$. Thus J cannot be tangent to $\dot{c}$ on $(0,t_0)$ and must be spacelike.

Let $\{E_0, \ldots, E_n\}$ be a parallel frame along c. We can write $J(t) = J^i(t)E_i(t)$. Given that $J(t_0) = 0$ it follows that $J^i(t_0) = 0$ and one has

$$J^i(t) = J^i(t_0) + \int_0^1 \frac{d}{ds} J^i(t_0 + s(t - t_0))ds = (t - t_0) \int_0^1 \frac{dJ^i}{ds}(t_0 + s(t - t_0))ds.$$

Define the vector field $Y(t) = Y^i(t)E_i(t)$ along c such that $J(t) = (t_0 - t)Y(t)$, i.e. $-Y^i(t)$ is given by the integral expression above. If $J(0) = 0$ then $Y(0) = 0$ and by the same argument as above, we can write $Y(t) = t\tilde{Y}(t)$. Define

$$\varphi(t) := \begin{cases} t(t_0 - t) & \text{if } J(0) = 0, \\ (t_0 - t) & \text{if } J(0) \neq 0. \end{cases}$$

Renaming $\tilde{Y}$ to Y, one then has

$$J(t) = \varphi(t)Y(t)$$

for all t. By smoothness and our previous considerations, Y must be spacelike on $(0, t_0)$. If we are now able to show that $Y(0)$ and $Y(t_0)$ are spacelike, we are done, because continuity of Y implies that there exists $\delta > 0$ such that $Y|_{[0,t_0+\delta]}$ is spacelike. After that, one can simply set $U := \frac{Y}{|Y|}$ and $f := \varphi|Y|$, which yields

$$J = \varphi Y = fU$$

on $[0, t_0 + \delta]$. We therefore begin by showing that $Y(0)$ and $Y(t_0)$ are both non-vanishing. For $J(0) \neq 0$ we obviously have $Y(0) \neq 0$, so suppose $J(0) = 0$. Since J is nontrivial, it follows that $J'(0) \neq 0$. Note that

$$J'(t) = \begin{cases} (t_0 - 2t)Y(t) + t(t_0 - t)Y'(t) & \text{if } J(0) = 0, \\ -Y(t) + (t_0 - t)Y'(t) & \text{if } J(0) \neq 0 \end{cases}$$

and $0 \neq J'(0) = t_0 Y(0)$ implies $Y(0) \neq 0$. Note that in any way $J'(t_0) \neq 0$ since J is nontrivial. The above then implies that $Y(t_0) \neq 0$ for both cases.

(i) <u>$Y(0)$ is spacelike:</u> Since J is a P-Jacobi field, one has $\langle J, \dot{c} \rangle = 0$ (Remark 7.0.6). Therefore

$$0 = \frac{d}{dt}\langle J, \dot{c} \rangle = \langle J', \dot{c} \rangle$$

as $\ddot{c} = 0$. We need to show that J' has no tangential components to c. Consider the case that $J(0) = 0$ and assume to the contrary that $Y(0) = \lambda \dot{c}(0)$ for $\lambda \in \mathbb{R}$. Then $J'(0) = t_0 \lambda \, \dot{c}(0)$ and consequently $J(t) = t \cdot t_0 \lambda \, \dot{c}(t)$ by uniqueness of Jacobi fields. But this is a contradiction since $J(t_0) = t_0^2 \lambda \dot{c}(t_0) \neq 0$ and by the same argument as in the proof of Lemma 6.1.4, $Y(0)$ can now only have spacelike components. Conversely, for $J(0) \neq 0$ simply note that $Y(0) = \frac{1}{t_0} J(0) \in T_pP$ is indeed spacelike.

(ii) $\underline{Y(t_0) \text{ is spacelike}}$: Again assume to the contrary that $Y(t_0)$ is proportional to $\dot{c}(t_0)$, i.e. $Y(t_0) = \lambda \dot{c}(t_0)$ for $\lambda \in \mathbb{R}$. Then

$$J'(t_0) = \begin{cases} -t_0 Y(t_0) & \text{if } J(0) = 0, \\ -Y(t_0) & \text{if } J(0) \neq 0 \end{cases} = \begin{cases} -t_0 \lambda \dot{c}(t_0) & \text{if } J(0) = 0, \\ -\lambda \dot{c}(t_0) & \text{if } J(0) \neq 0. \end{cases}$$

Again by uniqueness, we have

$$J(t) = (t_0 - t)\lambda\, \dot{c}(t) \cdot \begin{cases} t_0 & \text{if } J(0) \neq 0, \\ -1 & \text{if } J(0) = 0. \end{cases}$$

This yields a contradiction for $J(0) = 0$ and also for $J(0) \neq 0$ as $J(0)$ must be spacelike.

Claim 2. *There exists $\delta \in (0, b - t_0)$ and a vector field $V \in \mathfrak{X}(c)$ such that $V(0) = J(0)$, $V(t_0 + \delta) = 0$ and $V \perp \dot{c}$ on $[0, t_0 + \delta]$ and $\langle \frac{\nabla^2 V}{dt^2} + R_{\dot{c}V}\dot{c}, V \rangle > 0$ on $(0, t_0 + \delta)$.*

Let f, U and δ be as in Claim 1. Define $V \coloneqq (f + g)U = J + gU$ for a g that we are going to specify shortly. Note that

$$\begin{aligned} \frac{\nabla^2 V}{d^2 t} + R_{V\dot{c}}\dot{c} &= \frac{\nabla}{dt}\left(\frac{\nabla J}{dt} + g'U + g\frac{\nabla U}{dt}\right) + R_{V\dot{c}}\dot{c} \\ &= \frac{\nabla^2 J}{d^2 t} + g''U + 2g'\frac{\nabla U}{dt} + g\frac{\nabla^2 U}{d^2 t} + R_{\dot{c}J}\dot{c} + gR_{\dot{c}U}\dot{c} \\ &= g''U + 2g'\frac{\nabla U}{dt} + g\left(\frac{\nabla^2 U}{d^2 t} R_{\dot{c}U}\dot{c}\right) \end{aligned}$$

since J is a Jacobi field, i.e. $\frac{\nabla^2 J}{d^2 t} + R_{\dot{c}J}\dot{c} = 0$. Since $\langle U, U \rangle = 1$ it follows that $\langle \frac{\nabla U}{dt}, U \rangle = 0$ and thus

$$\begin{aligned} \left\langle \frac{\nabla^2 V}{d^2 t} + R_{\dot{c}V}\dot{c}, V \right\rangle &= (f + g)\left(g''\langle U, U\rangle + 2g'\left\langle \frac{\nabla U}{dt}, U \right\rangle + g\left\langle \frac{\nabla^2 U}{d^2 t} + R_{\dot{c}U}\dot{c}, U \right\rangle\right) \\ &= (f + g)(g'' + gh), \end{aligned} \tag{1}$$

where we defined $h \coloneqq \left\langle \frac{\nabla^2 U}{d^2 t} + R_{\dot{c}U}\dot{c}, U \right\rangle$. Let $\lambda > 0$ be such that

$$h \geq -\lambda^2 \qquad \text{on } [0, t_0 + \delta] \tag{2}$$

and set $g(t) \coloneqq \mu\left(e^{\lambda t} - 1\right)$ with $\mu > 0$ such that $g(t_0 + \delta) = -f(t_0 + \delta)$, which is possible since $f(t_0 + \delta) < 0$. Consequently $(f + g)(t_0 + \delta) = 0$ and since $f > 0$ on $(0, t_0)$, also $f + g > 0$ on $(0, t_0]$. Note that indeed

$$V(t_0 + \delta) = (f + g)(t_0 + \delta) \cdot U(t_0 + \delta) = 0 \quad \text{and} \quad V(0) = J(0) + g(0) \cdot U(0) = J(0).$$

Assume without loss of generality that $t_0 + \delta$ is the first positive zero of $(f + g)$ or otherwise shrink δ. One obtains

$$\left\langle \frac{\nabla^2 V}{d^2 t} + R_{\dot{c}V}\dot{c}, V \right\rangle \overset{(1)}{=} (f+g)(g'' + gh) > 0$$

on $(0, t_0 + \delta)$, because $g'' + gh = \lambda^2 g + \lambda^2 \mu + gh \geq \lambda^2 \mu > 0$ by (2). Finally, since $J = fU$, one has $U \perp \dot{c}$ for $f \neq 0$ and so $V \perp \dot{c}$ on $(0, t_0) \cup (t_0, t_0 + \delta)$. Continuity then yields the claim.

Claim 3. *There exists $A \in \mathfrak{X}(c)$ such that $A(0) = \mathrm{I\!I}(J(0), J(0))$, $A(t_0 + \delta) = 0$ and $-\langle V'' + R_{\dot{c}V}\dot{c}, V\rangle + \frac{d}{dt}(\langle V, V'\rangle + \langle A, \dot{c}\rangle) < 0$ on $[0, t_0 + \delta]$.*

Using Proposition 3.2.10 and that J is a P-Jacobi field, one finds

$$\langle \mathrm{I\!I}(J(0), J(0)), \dot{c}(0)\rangle = -\langle \tilde{\mathrm{I\!I}}(J(0), \dot{c}(0)), J(0)\rangle = -\langle \tan(J'(0)), J(0)\rangle = -\langle J'(0), J(0)\rangle. \quad \text{(A)}$$

Let V, U and g, F be as in the proofs of the previous claims. Then

$$\begin{aligned} \langle V, V'\rangle &= \langle J + \mu\left(e^{\lambda t} - 1\right)U, J' + \lambda\mu\, e^{\lambda t}U + \mu\left(e^{\lambda t} - 1\right)U'\rangle \\ &= \langle J, J'\rangle + \langle V, \lambda\mu\, e^{\lambda t}U\rangle + \mu\left(e^{\lambda t}\right)\langle U, J'\rangle, \end{aligned}$$

where we have used $\langle J, U'\rangle = f\langle U, U'\rangle = 0$. Thus, at $t = 0$

$$\langle V, V'\rangle(0) = \langle J, J'\rangle(0) + \mu\lambda\langle V(0), U(0)\rangle = \langle J, J'\rangle(0) + \mu\lambda|J(0)| \quad \text{(B)}$$

since $\langle V(0), U(0)\rangle = f(0)\langle U, U\rangle(0) = f(0) = |J(0)|$.
We distinguish three cases:

(i) $\langle \mathrm{I\!I}(\mathbf{J}(\mathbf{0}), \mathbf{J}(\mathbf{0})), \dot{\mathbf{c}}(\mathbf{0})\rangle \neq \mathbf{0}$:
Define $X_0 \in T_pP^{\perp}$ by $\alpha X_0 := \mathrm{I\!I}(J(0), J(0))$ such that $\langle X_0, \dot{c}(0)\rangle = -1$. This implies that

$$\alpha = -\langle \alpha X_0, \dot{c}(0)\rangle = -\langle \mathrm{I\!I}(J(0), J(0)), \dot{c}(0)\rangle = \langle J'(0), J(0)\rangle. \quad \text{(C)}$$

Let $X \in \mathfrak{X}(c)$ be the parallel transport of X_0. Now define $A \in \mathfrak{X}(c)$ by

$$A(t) := \left(\langle V(t), V'(t)\rangle + \frac{\lambda\mu\,|J(0)|}{t_0 + \delta}(t - t_0 - \delta)\right)X(t).$$

We may now simply calculate the following

$$\begin{aligned} A(0) &\overset{(B)}{=} \big(\langle J, J'\rangle(0) + \lambda\mu\,|J(0)| - \lambda\mu\,|J(0)|\big)X(0) = \langle J(0), J'(0)\rangle X_0 \\ &\overset{(C)}{=} \alpha X_0 = \mathrm{I\!I}(J(0), J(0)), \end{aligned}$$

$$A(t_0 + \delta) = \big(\langle V(t_0 + \delta), V'(t_0 + \delta)\rangle + 0\big)X(t_0 + \delta) = 0,$$

where we have used that $V(t_0+\delta) = 0$ by Claim 2. Finally, since $\langle X, \dot c\rangle = -1$, one finds

$$\begin{aligned}\frac{d}{dt}\left(\langle V, V'\rangle + \langle A, \dot c\rangle\right) &= \frac{d}{dt}\left(\langle V, V'\rangle - \langle V, V'\rangle - \frac{\lambda\mu\,|J(0)|}{t_0+\delta}(t-t_0-\delta)\right)\\ &= -\frac{\lambda\mu\,|J(0)|}{t_0+\delta}(t-t_0-\delta) < 0\end{aligned}$$

for $t \in [0, t_0+\delta]$. This proves the claim, since $-\langle V'' + R_{\dot c V}\dot c, V\rangle \le 0$ by Claim 2.

(ii) $\langle \mathrm{I\!I}(\mathbf{J}(\mathbf{0}), \mathbf{J}(\mathbf{0})), \dot{\mathbf{c}}(\mathbf{0})\rangle = \mathbf{0}$ **but** $\mathbf{J}(\mathbf{0}) \neq \mathbf{0}$:
Now let X be any parallel vector field along c such that $\langle X, \dot c\rangle = -1$ and let $Z \in \mathfrak{X}(c)$ be the parallel transport of $\mathrm{I\!I}(J(0), J(0))$. Define

$$A(t) := \left(\langle V(t), V'(t)\rangle + \frac{\lambda\mu\,|J(0)|}{t_0+\delta}(t-t_0-\delta)\right)X(t) + \left(1 - \frac{t}{t_0+\delta}\right)Z(t).$$

Again, analogously to the previous case, we find

$$\begin{aligned}A(0) &= \langle J(0), J'(0)\rangle X(0) + Z(0) \overset{(A)}{=} -\underbrace{\langle \mathrm{I\!I}(J(0), J(0)), \dot c(0)\rangle}_{=0} X(0) + \mathrm{I\!I}(J(0), J(0))\\ &= \mathrm{I\!I}(J(0), J(0)),\end{aligned}$$

$$A(t_0+\delta) = 0 + 0\cdot Z(t_0+\delta) = 0.$$

By definition, $Z \perp \dot c$, and as before

$$\frac{d}{dt}\left(\langle V, V'\rangle + \langle A, \dot c\rangle\right) = -\frac{\lambda\mu\,|J(0)|}{t_0+\delta}(t-t_0-\delta) < 0$$

which gives the claim.

(iii) $\mathbf{J}(\mathbf{0}) = \mathbf{0}$:
Let $T := t_0 + \delta$ and pick $\tilde f \in \mathcal{C}^\infty([0,T])$ as in Figure 32. Choose $t_1, t_2 \in [0,T]$ such that $\tilde f'\Big|_{[0,t_1]\cup[t_2,T]} \equiv 1$.

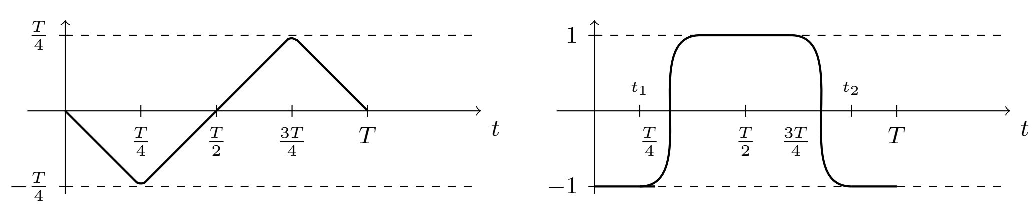


Figure 32: $\tilde f$ (left) and its derivative $\tilde f'$ (right).

By Claim 2, let $\epsilon > 0$ such that $\epsilon < \min_{t\in[t_1,t_2]}\langle V'' + R_{\dot c V}\dot c, V\rangle(t)$ and set $f := -\epsilon\tilde f$.

Define $X \in \mathfrak{X}(c)$ as previously in (ii) and set

$$A(t) := \big(\langle V, V'\rangle(t) + f(t)\big)\, X(t).$$

Using that $J(0) = 0$ and $f(0) = 0 = f(T)$, one then calculates

$$A(0) \overset{(B)}{=} \big(\langle J(0), J'(0)\rangle + \lambda\mu\, |J(0)| + f(0)\big)\, X(0) = 0 = \mathrm{I\!I}(J(0), J(0)),$$

$$A(t_0+\delta) = A(T) = 0 + f(T)X(T) = 0$$

as $V(T) = 0$ by Claim 2. Finally, since we chose X such that $\langle X, \dot c\rangle = -1$, it follows that $\frac{d}{dt}\left(\langle V, V'\rangle + \langle A, \dot c\rangle\right) = -f' = \epsilon \tilde f'$. Consequently, again by Claim 2,

$$-\langle V'' + R_{\dot c V}\dot c, V\rangle + \frac{d}{dt}\big(\langle V, V'\rangle + \langle A, \dot c\rangle\big) \le \epsilon \tilde f' = -\epsilon < 0$$

for $t \in [0, t_1] \cup [t_2, T]$ because then $\tilde f'(t) = -1$. For $t \in [t_1, t_2]$ we have $\tilde f'(t) \le 1$ and thus

$$-\langle V'' + R_{\dot c V}\dot c, V\rangle + \frac{d}{dt}\big(\langle V, V'\rangle + \langle A, \dot c\rangle\big) \le -\langle V'' + R_{\dot c V}\dot c, V\rangle + \epsilon < 0$$

by definition of ϵ and therefore $-\langle V'' + R_{\dot c V}\dot c, V\rangle + \frac{d}{dt}\left(\langle V, V'\rangle + \langle A, \dot c\rangle\right) < 0$ on all of $[0, T]$ as desired.

Claim 4. *There exists a timelike variation c_s of $c|_{[0,t_0+\delta]}$ such that $t \mapsto \dot c_s(t)$ is timelike for all $|s| > 0$ small.*

Let V and A be as in the previous claims. Define $\sigma : (-\epsilon, \epsilon) \to P$ by $\sigma(s) := \cancel{\exp}_p(sJ(0))$. Note that σ is smooth and satisfies $\sigma(0) = p$, $\dot\sigma(0) = J(0) = V(0)$ and $\frac{\cancel{\nabla}\dot\sigma}{ds} \equiv 0$. In particular,

$$\frac{\nabla\dot\sigma}{ds}(0) = \frac{\cancel{\nabla}\dot\sigma}{ds}(0) + \mathrm{I\!I}(\dot\sigma(0), \dot\sigma(0)) = A(0).$$

Pick a variation c_s of c (see Figure 33) which satisfies the following

$$c_s(0) = \sigma(s), \qquad c_s(t_0+\delta) = c(t_0+\delta), \qquad \left.\frac{\partial c_s}{\partial s}\right|_{s=0} = V, \qquad \left.\frac{\nabla}{ds}\frac{\partial c_s}{\partial s}\right|_{s=0} = A.$$

To show that such a variation exists one can, for example, introduce Fermi coordinates just as in [KS23, Prop. 3.1.8 & Prop. 3.2.17].

Let $f(t,s) := \langle \dot c_s, \dot c_s\rangle(t)$. Since $\dot c$ is null, $\ddot c = 0$ and $V \perp \dot c$, one has

$$f(t,0) = \langle \dot c(0), \dot c(0)\rangle = 0,$$

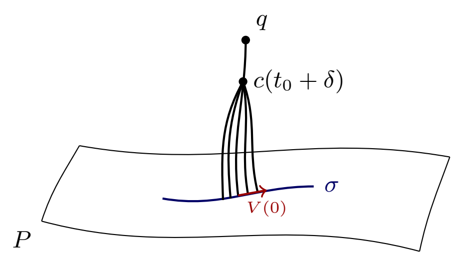


Figure 33

$$\begin{aligned}\frac{\partial f}{\partial s}(t,0) &= 2\left\langle \frac{\nabla}{\partial s}\frac{\partial c_s}{\partial t}(t,0), \dot{c}(t)\right\rangle \overset{2.1.6(i)}{=} 2\left\langle \frac{\nabla}{\partial t}\frac{\partial c_s}{\partial s}(t,0), \dot{c}(t)\right\rangle \\ &= 2\left\langle \frac{\nabla}{\partial t}V, \dot{c}\right\rangle(t) = 2\frac{d}{dt}\langle V, \dot{c}\rangle(t) - 2\langle V, \ddot{c}\rangle(t) = 0\end{aligned}$$

and furthermore

$$\begin{aligned}\frac{1}{2}\frac{\partial^2 f}{\partial s^2}(t,0) &= \frac{\partial}{\partial s}\bigg|_{s=0}\left\langle \frac{\nabla \dot{c}_s}{\partial s}, \dot{c}_s\right\rangle(t) \overset{2.1.6(i)}{=} \frac{\partial}{\partial s}\bigg|_{s=0}\left\langle \frac{\nabla}{\partial t}\frac{\partial c_s}{\partial s}, \dot{c}\right\rangle(t) \\ &= \left\langle \frac{\nabla V}{dt}, \frac{\nabla V}{dt}\right\rangle(t) + \left\langle \frac{\nabla}{\partial s}\frac{\nabla}{\partial t}\frac{\partial c_s}{\partial s}\bigg|_{s=0}, \dot{c}\right\rangle(t) \\ &\overset{2.1.6(ii)}{=} \left\langle \frac{\nabla V}{dt}, \frac{\nabla V}{dt}\right\rangle(t) + \left\langle \frac{\nabla}{\partial t}\frac{\nabla}{\partial s}\frac{\partial c_s}{\partial s}\bigg|_{s=0} + R_{\dot{c}_s \frac{\partial c_s}{\partial s}}\frac{\partial c_s}{\partial s}\bigg|_{s=0}, \dot{c}\right\rangle(t) \\ &= \left\langle \frac{\nabla V}{dt}, \frac{\nabla V}{dt}\right\rangle(t) + \left\langle \frac{\nabla A}{\partial t}, \dot{c}\right\rangle(t) - \langle R_{\dot{c}V}\dot{c}, V\rangle(t) \\ &= \left\langle \frac{\nabla V}{dt}, \frac{\nabla V}{dt}\right\rangle(t) + \frac{d}{dt}\langle A, \dot{c}\rangle(t) - \langle R_{\dot{c}V}\dot{c}, V\rangle(t) \\ &= \frac{d}{dt}\big(\langle V, V'\rangle + \langle A, \dot{c}\rangle\big)(t) - \big(\langle V'', V\rangle + \langle R_{\dot{c}V}\dot{c}, V\big)(t) \\ &= \frac{d}{dt}\big(\langle V, V'\rangle + \langle A, \dot{c}\rangle\big)(t) - \langle V'' + R_{\dot{c}V}\dot{c}, V\rangle < 0\end{aligned}$$

by Claim 3. Note that this holds for all $t \in [0, t_0 + \delta]$. Let us now consider the Taylor expansion of f with respect to s in Lagrange form, i.e.

$$f(t,s) = f(t,0) + \frac{\partial f}{\partial s}(t,0) + \frac{s^2}{2}\frac{\partial^2 f}{\partial s^2}(t,\xi),$$

where $\xi = \xi(t) \in (0,s)$. By the previous calculations and compacteness of $[0, t_0 + \delta]$, there must exists an $\epsilon > 0$ such that the above expression is strictly smaller than 0 on $[0, t_0 + \delta] \times [-\epsilon, \epsilon]$. Thus, $c_s(t)$ is timelike for $s \neq 0$. Finally, using the push-up principle, we can deform $c_s|_{[0,t_0+\delta]} \cup c|_{[t_0+\delta,b]}$ into a timelike curve from P to q with arbitrary precision to c, which concludes our proof.

□

Due to our special interest in focal points, we have worked on the presumption that c is a geodesic orthogonal to P. It turns out that dropping this assumption makes things much easier. We follow [O'N83, Lemma 10.50].

Proposition 7.0.15. *Let $P \subseteq M$ be a spacelike submanifold and $c : [0, b] \to M$ a null geodesic emanating from P but not orthogonally, i.e. $c(0) \in P$ and $\dot{c}(0) \not\perp P$. Then there exists a timelike curve from P to $c(b)$, which lies arbitrarily close to c.*

Proof. Let $p := c(0) \in P$. By assumption there exists $v \in T_pP$ such that $\langle v, \dot{c}(0) \rangle \neq 0$ and without loss of generality $\langle v, \dot{c}(0) \rangle > 0$ (otherwise change to $-v$). Let $V \in \mathfrak{X}(c)$ be the parallel transport of v and define $X \in \mathfrak{X}(c)$ by $X(t) = \left(1 - \frac{t}{b}\right) V(t)$. Because $X(0) \in T_pP$ and $X(b) = 0$, proceeding as in Lemma 4.3.3, one may construct a variation between P and $c(b)$ with variation vector field X. Note that $X' = -\frac{1}{b}V$ and therefore

$$\langle X', \dot{c} \rangle = -\frac{1}{b} \langle V, \dot{c} \rangle = -\frac{1}{b} \langle v, \dot{c}(0) \rangle < 0.$$

The claim now follows by Lemma 4.3.4. □

8 Penrose Incompleteness Theorem

Among the fundamental differences between Riemannian and Lorentzian geometry is the fact that the time-separation function $\tau : M \times M \to \mathbb{R}$ does not define a metric on M. As a consequence, one does not have a result akin to the Hopf-Rinow Theorem 2.1.8 that relates geodesic completeness to completeness of a metric space. The situation surrounding completeness is therefore a much more intricate issue in Lorentzian geometry. First and foremost, the causal structure of a Lorentzian manifold enables us to actually distinguish different kinds of notions of completeness, namely spacelike, timelike and null completeness, i.e. geodesic completeness with respect to spacelike, timelike and null geodesics respectively. In his famous 1965 paper, [Pen65], Sir Roger Penrose gave a proof of a result that provided sufficient conditions for when a Lorentzian manifold is null incomplete. The incompleteness Theorem - or more commonly the Penrose singularity Theorem - marks one of the corner stones of 20th century mathematical physics. Of course, the result itself is purely mathematical but it nevertheless is of major significance in Einstein's theory of General Relativity. The proof consists of well-known concepts from Riemannian geometry adapted to the Lorentzian setting along with the novel concept of trapped surfaces, which since then have become a crucial part of modern research. In this chapter we present a proof of the celebrated Theorem, following ideas of Gregory J. Galloway in [Gal14]. The chapter itself is split into three sections. In the first we will discuss and introduce the notion of trapped surfaces and null geodesic congruences. The second section serves as a heuristic introduction to General Relativity, in particular from an initial value problem perspective. This serves the purpose of putting Penrose's result in the right historic setting with the goal of obtaining a better intuition. Finally, in the third section, we will give a prove of the standard incompleteness Theorem as well as a variation by Gannon and Lee.

8.1 Null Geodesic Congruences and Trapped Surfaces

For our final final preparation, we consider a closed[40] spacelike submanifold $\Sigma \subseteq M$ with $\operatorname{codim}(\Sigma) = 2$. For each point $p \in \Sigma$, the space $T_p\Sigma^\perp$ then contains two distinct future null directions $\bar{l}_p, l_p \in T_p\Sigma^\perp$. For this section, we will further suppose that Σ is two-sided and admits two smooth and linearly independent null vector fields $\overline{L}$ and L with $\overline{L}(p) = \bar{l}_p$ and $L(p) = l_p$. The main sources for this section are [O'N83], [Gal14] and [Are18].

Definition 8.1.1. *Let $\Sigma \subseteq M$ be a spacelike submanifold with $\operatorname{codim}(\Sigma) = 2$ and with two linearly independent, future directed null vector fields $\overline{L}, L \in \mathfrak{X}^\perp(\Sigma)$. For each $q \in \Sigma$ denote by $\bar{c}_q : [0, \overline{T}_q) \to M$ and $c_q : [0, T_q) \to M$ the unique null geodesics emanating from q with inital velocities $\overline{L}(q)$ and $L(q)$ respectively and with $\overline{T}_q, T_q > 0$ defining the maximal future domain of definition. The future directed null geodesic congruences of Σ are then defined as the following sets*

$$\overline{C} = \bigcup_{q\in\Sigma} \operatorname{Im}(\bar{c}_q), \qquad C = \bigcup_{q\in\Sigma} \operatorname{Im}(c_q).$$

Example 8.1.2. Recall Example 7.0.7. Let $\Sigma \subseteq \mathbb{R}^{1,2}$ be a one dimensional, spacelike submanifold with $\Sigma \cong_{\text{hom.}} S^1$. We can define null vector fields $\overline{L}, L \in \mathfrak{X}^\perp(\Sigma)$ that point towards the inside and the outside of Σ respectively (see Figure 34). One suitably calls them the *ingoing* and *outgoing* null vector fields. We will similarly call the geodesic congruence generated by L (resp. $\overline{L}$) the *outgoing* (resp. *ingoing*) null congruence.

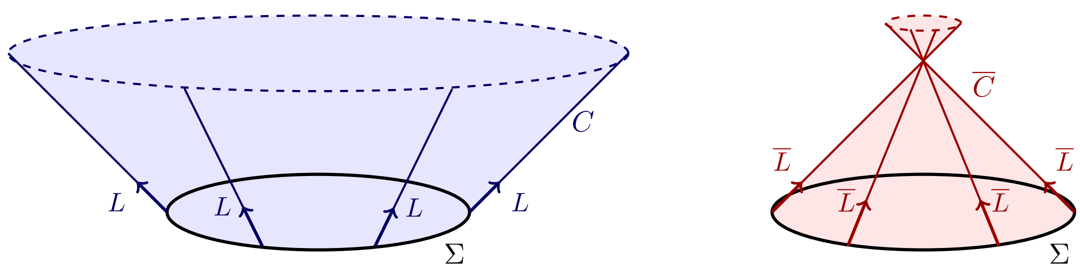


Figure 34: Outgoing null geodesic congruence C (left) and ingoing null geodesic congruence $\overline{C}$ (right) for $\Sigma = S^1 \subseteq \mathbb{R}^{1,2}$.

Remark 8.1.3. Let Σ be as in Definition 8.1.1. Following [O'N83, Lemma 7.25 & Prop. 7.27] there is a neighbourhood $U \subseteq T\Sigma^\perp$ of the zero section $T\Sigma^\perp_0 = \{(p,0) \in T\Sigma^\perp \,|\, p \in \Sigma\}$ that is mapped diffeomorphically into M via $\exp^\perp$. In particular, restricting to the null line subbundles $\mathcal{L}$ and $\overline{\mathcal{L}}$ of $T\Sigma^\perp$ which are spanned by L and $\overline{L}$ respectively, one may show that $\exp^\perp$ embeds $U \cap \mathcal{L}$ or $U \cap \overline{\mathcal{L}}$ into M. Thus the images are smooth n-dimensional submanifolds

[40] Compact and without boundary.

locally around Σ and the congruences C and $\overline{C}$ are locally smooth hypersurfaces there.
We now claim that these hypersurfaces are also null. Without loss generality consider C and denote by H the subset where C is a smooth hypersurface. Let $p \in H$ and $a \in \mathbb{R}$ such that $p = c_q(a)$, where c_q is the null geodesic emanating from $q \in \Sigma$ and contained in H. At $q \in \Sigma$ choose an orthonormal basis $\{e_1, \ldots, e_{n-1}\} \subset T_q\Sigma$ consisting of spacelike vectors. For each $i \in \{1, \ldots, n-1\}$ choose a curve $\sigma_i(s)$ on Σ with $\sigma_i(0) = q$ and $\dot{\sigma}_i(0) = e_i$. Define a variation through null geodesics by setting

$$\alpha(t, s) := \exp^{\perp}_{\sigma_i(s)}(tL(\sigma_i(s))).$$

By Propositions 4.3.9, 7.0.3 and 7.0.12 each α_i yields a Σ-Jacobi field $J_i \in \mathcal{J}_\Sigma$ for which $J_i \perp c_q$. In particular $J_i(a) \perp c_q(a)$ and it follows that $J_i(a) \in T_pH$ can be written as $J_i(a) = v + w$ for $v, w \in T_pH$ with v spacelike and w null. Since w is necessarily proportional to $\dot{c}_q(a)$, uniqueness of the Jacobi equations implies that a null component must come from the trivial Jacobi field $J_{\text{triv}}(t) = t\dot{c}_q(t)$. Consequently, assume without loss of generality[41] that $w = 0$. We need to show that the $J_i(a)$ are linearly independent. To this end, suppose to the contrary that there are $i, j \in \{1, \ldots, n-1\}$ with $i \neq j$ such that $J_i(a) = \lambda J_j(a)$ for some $\lambda \in \mathbb{R}$. The geodesic variation with the Jacobi field $\overline{J} := J_i - \lambda J_j$ then yields $\overline{J}(a) = 0$, which contradicts the fact that c_q cannot feature a focal point as otherwise $\exp^{\perp}$ must be singular by Proposition 7.0.11. This finishes the proof as we have found that T_pH is spanned by $n-1$ linearly independent spacelike directions together with the null vector $\dot{c}_q(a)$, i.e. T_pH is a null vector space.

Note that one can generally not expect all of C or $\overline{C}$ to be smooth hypersurfaces. Considering Example 8.1.2 again, the ingoing geodesic congruence of a standard (spacelike) sphere in Minkwoski space is converging to the apex of a cone and thus fails to be smooth (see also Figure 35). In particular, observe that the apex of the cone is in fact a focal point of Σ and thus $\exp^{\perp}$ fails to be an immersion. Therefore, the map cannot be an embedding and our previous argument fails.

Note that we will revisit some of these arguments later in the proof of the Incompleteness Theorem, but for now let us return to Σ itself. As already mentioned in the introduction of Chapter 3, the Penrose Incompleteness Theorem assumes the existence of a so-called *trapped surface*. By this we mean a closed spacelike surface Σ in M for which both null generators of C and $\overline{C}$ are "converging" towards the future. To define this more precisely, it is useful to introduce the following notions

Definition 8.1.4. *Let Σ be a closed spacelike surface in M with future directed null vector*

[41] Otherwise consider a new variation through null geodesics with Jacobi field $J_i - \lambda J_{\text{triv}}$, where $\lambda \in \mathbb{R}$ is chosen such that $\lambda a \dot{c}_q(a) = w$.

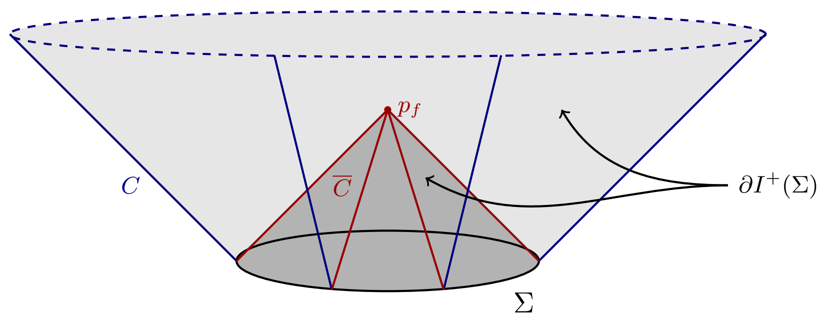


Figure 35: The null congruences C and $\overline{C}$ for a spacelike circle Σ in Minkwoski space. The inner null congruence $\overline{C}$ features a focal point p_f of Σ.

fields $\overline{L}, L \in \mathfrak{X}(\Sigma)^\perp$. The associated null second fundamental forms are defined as the maps

$$\overline{\chi} : \mathfrak{X}(\Sigma) \times \mathfrak{X}(\Sigma) \to \mathcal{C}^\infty(\Sigma), \qquad \overline{\chi}(X,Y) = g(\nabla_X \overline{L}, Y)$$

and

$$\chi : \mathfrak{X}(\Sigma) \times \mathfrak{X}(U) \to \mathcal{C}^\infty(\Sigma), \qquad \chi(X,Y) = g(\nabla_X L, Y).$$

The associated null mean curvatures $\overline{\theta}, \theta$ are then defined by[42]

$$\overline{\theta} = \mathrm{tr}_{\not g}\, \overline{\chi} = \mathrm{div}_\Sigma\, \overline{L}, \qquad \theta = \mathrm{tr}_{\not g}\, \chi = \mathrm{div}_\Sigma\, L,$$

where $\not g$ is the induced metric on Σ.

We can see that $\overline{\chi}$ and χ are $(0,2)$-tensors on Σ and that the associated null mean curvatures are smooth. Furthermore, they are closely related to their respective counterparts in Chapter 6. We may now define a trapped surface in the following way

Definition 8.1.5. *Let Σ be a closed spacelike surface in M with null directions $\overline{L}, L \in \mathfrak{X}^\perp(\Sigma)$. Σ is called*

(i) ***inner trapped*** *if $\overline{\theta} < 0$,*

(ii) ***outer trapped*** *if $\theta < 0$,*

(iii) ***trapped*** *if it is both, outer and inner trapped, i.e. $\overline{\theta}, \theta < 0$.*

For non-negative Ricci null curvature spaces, i.e. $\mathrm{Ric}(X,X) \geq 0$ for any null vector $X \in \mathfrak{X}(M)$, one finds that for trapped surfaces the null mean curvature θ satisfies $\theta < 0$ along any of the null generators of C and $\overline{C}$ starting from a trapped surface. This follows from the Raychaudhuri equation exactly as before in Proposition 6.2.9. There, we obtained the

[42]Following Proposition 6.1.19.

inequality

$$\theta' \leq -\frac{1}{n-1}\theta^2$$

for the change of null mean curvature along a generator $\eta : [0, T) \to M$ of C or $\overline{C}$ respectively. Because $\theta(0) = \not\theta(\eta(0)) < 0$, it follows that $\theta(t) < 0$ for any t. Of course, the same holds true for inner and outer trapped surfaces and the generators on $\overline{C}$ and C respectively.
Recall now that back in Chapter 6.1 we have commented on the fact that the divergence, and in fact θ, measures the overall expansion of the null generators towards the future. The intuitive geometric picture of a trapped surface Σ and both its null congruences C and $\overline{C}$ one should therefore keep in mind is seen in Figure 36 down below.

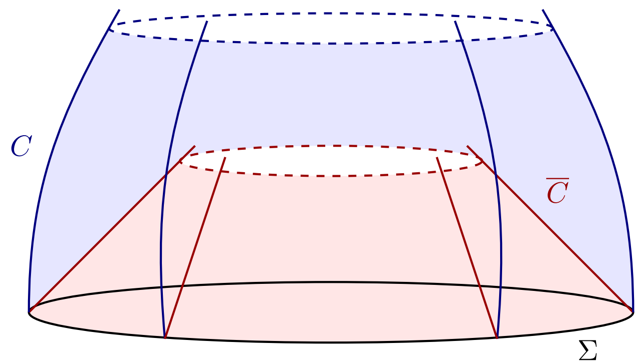


Figure 36: A future trapped surface Σ. The area of Σ transported along the ingoing and outgoing null geodesics is shrinking towards the future.

8.2 Elements of General Relativity

To put the Penrose's Theorem into its natural environment, which is General Relativity, we will give a brief overview of the theory. As a disclaimer, it should be noted that we cover a lot of ground here and are imprecise at times. A detailed discussion is far beyond the scope of this work (which is concentrated on the mathematical background and proof of the Theorem) and we will thus only sketch the main elements of the theory. We will also primarily follow [Chr99], [Daf21] and [Rin09].
The setting of General Relativity is that of a four dimensional spacetime (M, g) which solves the Einstein field equations

$$\mathrm{Ric} - \frac{1}{2}Sg + \Lambda g = 8\pi T. \tag{EEQs}$$

Here S denotes the scalar curvature which is obtained as the contraction of Ric. T is the stress-energy-momentum tensor, a symmetric $(0, 2)$-tensor field that is divergence free and encodes the physical matter content of the system under consideration. Furthermore, the third term on the left hand side includes the cosmological constant Λ that is required in

standard cosmological models. For our discussion however, it is sufficient to set $\Lambda = 0$.
The most remarkable feature of the Einstein equations is that they render General Relativity a deterministic theory. In particular, equation (EEQs) takes the form of an initial value problem: solutions evolve from prescribed initial data and are completely determined by it. In what follows, we will consider solely the Einstein vacuum equations where no matter is present, i.e. $T = 0$. In this case, Equation (EEQs) then takes the simple form[43]

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

which can be viewed as a non-linear wave equation for the metric components (cf. [Rin09, Chapter 14]).
We must now pose the question what it means for a spacetime (M, g) to solve the above equations for specified initial data $(\mathcal{H}, \not{g}, \mathrm{I\!I})$. In fact, the first question at hand is what even constitutes as initial data. Therefore, we loosely define the following

Definition 8.2.1. *An initial data set for equation (VEEQs) is a triple $(\mathcal{H}, \not{g}, \mathrm{I\!I})$, where*

(i) $(\mathcal{H}, \not{g})$ *is a three-dimensional Riemannian manifold,*

(ii) $\mathrm{I\!I}$ *is a* $(0, 2)$*-tensor field on* $\mathcal{H}$*,*

such that the constraint equations[44] are satisfied.

Remark 8.2.2. The requirement that the initial data satisfy some form of constraint equations reflects the fact that a solution (M, g) of the vacuum Einstein equations induces $\mathcal{H}$ as a smooth spacelike hypersurface, and in fact as a Cauchy surface. In particular, as the notation suggests, $\mathrm{I\!I}$ will turn out to be the second fundamental form of $\mathcal{H}$ in (M, g). Consquently, $\not{g}$ and $\mathrm{I\!I}$ cannot be specified freely because they must be extrinsically compatible with (M, g) via the Gauß and Codazzi equations (See e.g. [KS23, Theorem 1.3.14, Theorem 1.6.6]). Combining these equations with (VEEQs), one may show that $(\mathcal{H}, \not{g}, \mathrm{I\!I})$ must satisfy a set of *constraint equations*, which are intrinsic to $\mathcal{H}$. This is crucial, since in the initial-value picture the spacetime (M, g) *does not yet exist* and one works purely from an initial data perspective.

For our purposes, the exact form of the constraint equations is not essential. The important picture to keep in mind is rather that $\mathcal{H}$ will emerge as a well-defined Cauchy surface of (M, g). With this, we shall now define what it means for (M, g) to solve (VEEQs) for prescribed initial data $(\mathcal{H}, \not{g}, \mathrm{I\!I})$.

Definition 8.2.3. *The future (resp. past) development of initial data $(\mathcal{H}, \not{g}, \mathrm{I\!I})$ is a four-dimensional spacetime (M, g) with boundary, such that*

(i) (M, g) *satisfies equation (VEEQs),*

[43] See [KS21, p. 86].
[44] For the specific details of these equations see [Rin09, Chapter 13].

(ii) $\partial M \cong_{\textit{diff.}} \mathcal{H}$,

(iii) $g|_{\mathcal{H}} = \not{g}$,

(iv) $\mathrm{I\!I}$ *is the second fundamental form on* $\mathcal{H}$,

(v) $\mathcal{H}$ *is a past (resp. future) Cauchy surface of* M*, i.e. any past (resp. future) directed causal curve from any* $p \in M$ *intersects* $\mathcal{H}$ *at a unique point.*

The Cauchy development or globally hyperbolic development of $(\mathcal{H}, \not{g}, \mathrm{I\!I})$ *is then defined as the four-dimensional spacetime* (M, g) *together with an embedding* $\iota : \mathcal{H} \to M$*, such that* $(i), (iii)$ *and* (iv) *are satisfied and* $\iota(\mathcal{H})$ *is a Cauchy surface of* M*.*

Remark 8.2.4. If (M, g, ι) is the Cauchy development of initial data $(\mathcal{H}, \not{g}, \mathrm{I\!I})$, one naturally has $D(\mathcal{H}) = M$, where $D(\mathcal{H})$ is the domain of dependence (or maximal Cauchy development) of $\mathcal{H}$ from Definition 5.2.13. Furthermore, $D^+(\mathcal{H})$ must be a future development and $D^-(\mathcal{H})$ a past development of this initial data respectively.

Example 8.2.5. We now consider two spacetimes solutions of (VEEQs).

(i) Minkowski space $(\mathbb{R}^{1,3}, \eta)$ is the Cauchy development of *trivial initial data* $(\mathcal{H}, \not{\eta}, \mathrm{I\!I})$, where $\mathcal{H}$ is a complete spacelike hypersurface in $\mathbb{R}^{1,3}$ and $\not{g}, \mathrm{I\!I}$ its induced metric and second fundamental form.

(ii) The Schwarzschild solution of the Einstein vaccum equations was discovered shortly after the advent of Einstein's theory of General Relativity. In Example 6.1.3, we encountered this solution in the form of the so-called Eddington-Finkelstein extension. Recall that Figure 27 depicted this in the form of a Penrose diagram, where the extension was represented as only the subset of yet another larger spacetime manifold. This maximal Schwarzschild solution is given by the *Kruskal extension*[45] $M = \{(U, V) \in \mathbb{R} \,|\, UV < 1\} \times S^2$ with

$$g = -\frac{32m^3 e^{-r(U,V)/2m}}{r(U,V)} dU \otimes dV + r(U,V)^2 d\Omega \otimes d\Omega,$$

where $m \neq 0$, $r(U, V)$ is a smooth function with values in $(0, \infty)$ that is to be interpreted as a radial coordinate and $d\Omega \otimes d\Omega$ being the unit round metric on S^2, i.e. $d\Omega \otimes d\Omega = d\theta \otimes d\theta + \sin^2\theta \, d\phi \otimes d\phi$. (M, g) then is realized as the Cauchy development of spherically symmetric data which is asymptotically flat with 2 ends (see [Wal84, Chapter 11] for details). Figure 37 now depicts the full Penrose diagram of the maximal Schwarzschild spacetime. In particular, Birkhoff's Theorem (cf. [Jeb21] and [Bir23]) states that any spherically symmetric solution to (VEEQs) is necessarily isometric to a subset of (M, g) or $(\mathbb{R}^{1,3}, \eta)$.

[45] See e.g. [Rea24, Chapter 2].

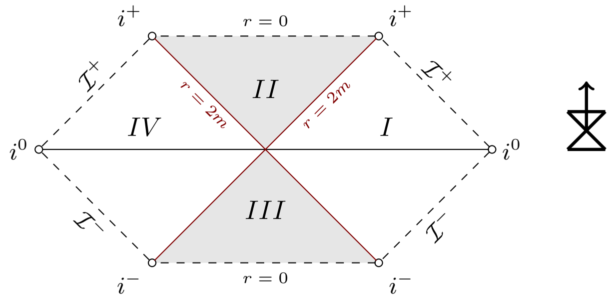


Figure 37: The Penrose diagram of the maximal Cauchy development of Schwarzschild initial data (cf. 27). Regions II and III are called the black hole and white hole region respectively. For a detailed discussion of this Figure we again refer to [Rea24, Chapters 2 & 5], [Wal84, Ch. 6,11 and 12] and [DR05, Appendix C].

Having introduced the vacuum Einstein equations as an initial value problem, a natural first question to ask is whether one has local existence and uniqueness of solutions. This was answered in the positive by Choquet-Bruhat in 1952 (see [CB52]). Later in 1969, Choquet-Bruhat together with Geroch further proved[46] the following refinement, as seen in [CBG69]

Theorem 8.2.6 (Choquet-Bruhat, Geroch)**.** *Let* $(\mathcal{H}, \not{g}, \mathbb{I})$ *be an initial data set of the Einstein vacuum equations. Then there exists a unique (up to diffeomorphism) maximal Cauchy development* (M, g) *of* $(\mathcal{H}, \not{g}, \mathbb{I})$*, i.e. any other Cauchy development is isometric to a subset of* (M, g).

Note that analogous results to Theorem 8.2.6 hold for suitable matter models and equation (EEQs) (see e.g. [Rin09]).
At this point we want to draw attention to the fact that the maximal Schwarzschild solution from Example 8.2.5 (*ii*) exhbitis a *singularity*. Of course this term requires clarification, but to actually provide a satisfactory and general definition turns out to be a notoriously difficult problem (cf. [Ger68]). Despite this, it is reasonable from a physical perspective to relate it to the existence blow-up's of curvature-related scalar quantity. In the Schwarzschild spacetime this is indeed the case: the *Kretschmann scalar* $K = R^{abcd}R_{abcd}$ diverges as $r \to 0$. Accordingly, we will refer to the singular boundary defined by $r = 0$ in Figure 37 as the singularity of the Schwarzschild spacetime. Notably, it may be shown that any causal geodesic that travels towards this singularity would "arrive" there in finite parameter time. In conclusion, the maximal Schwarzschild solution is causally incomplete in the sense of Definition 2.2.20.
This is surprising because causal incompleteness may generally be viewed as an undesirable property for a physical theory. Indeed, assuming that the Cauchy development is maximal, the failure to predict the fate of an incomplete geodesic, e.g. the trajectory of a particle, signals that the underlying theory may be insufficient. However, in the present example, the problem

[46]It is worth noting that their proof appeals to Zorn's Lemma. A more explicit argument was given much later by Sbierski, [Sbi16].

of incompleteness may be more academic than physical in nature. Naively, the Schwarzschild spacetime serves as a physical model for the vacuum region outside of spherically symmetric stellar objects and usually these kind of objects have radii of $r \gg 2m$. This is why, historically speaking, the singularity in Schwarzschild was simply thought of as a remnant of symmetry, not to be taken too seriously. To add to this, causal incompleteness is not a universal phenomenon: In the case of trivial data, we obtain a perfectly complete Minkowski spacetime. In fact, Christodoulou and Klainerman proved that one can also obtain complete solutions for certain asymptotically flat non-trivial data, [CK93].
After the issue of incompleteness in the fully extended Schwarzschild solution was therefore largely dismissed on physical grounds[47], Oppenheimer and Snyder presented in 1939 a physical model of spherically symmetric gravitational collapse (cf. [OS39]). In their work they considered a massive star collapsing under the influence of its own gravitational field in spherical symmetry. Mathematically, they solved the Einstein equations (EEQs) for a perfect fluid with no pressure, i.e.

$$T = \rho\, u^\flat \otimes u^\flat,$$

where ρ is the rest mass density of the star and $u^\flat$ is the metric equivalent of the fluids 4-velocity (see [Wal84, Chapters 4.2 & 4.3]). By Birkhoff's Theorem (see Example 8.2.5 (ii)), the vacuum region outside of the matter is given by the Schwarzschild solution. Thus, the spacetime must exhibit the same singularity encountered before in the Schwarzschild solution. It was then Sir Roger Penrose, who was able to prove a much more general result related to causal incompleteness. In his famous and concise 1965 paper, [Pen65], he examined similar collapsing scenarios akin to the one above. More precisely, he began by considering the Cauchy development (M, g) of initial data $(\mathcal{H}, \not{g}, \mathrm{I\!I})$, where $(\mathcal{H}, \not{g})$ is a complete Riemannian manifold that is asymptotically flat[48] and represents the initial matter distribution, where $\mathrm{supp}(T)$ is initially contained in a ball of radius $R_0 > 2m$ inside $\mathcal{H}$. Consider Figure 38 for a depiction of the scenario.

Let us analyse this figure more carefully (see also [Daf21]). The boundary of the blue cone is given by the radius $r = R(v)$ of the collapsing matter, e.g. a star. The radius itself depends on the advanced time coordinate v of the Eddington-Finkelstein extension of the Schwarzschild spacetime (cf. Example 6.1.3 (ii)). Let $V_{\text{crit.}} \in \mathbb{R}$ be such that $R(v) \to 0$ as $v \to V_{\text{crit.}}$. For $v > V_{\text{crit.}}$ one now recovers the Schwarzschild solution, which features a black hole region denoted by II. In this region, any future directed causal curve must travel towards the singularity. In fact, one may show that if $t \mapsto c(t)$ is such a curve that there exists $T \in \mathbb{R}$ such that $r(c(t)) \to 0$ as $t \to T$. The crucial observation is that any two-sphere $\Sigma \cong S^2$ in the vaccum region with $v = \text{const.}$ and $r_S > r = \text{const.}$ has the property that both its null

[47] Note that this also took place before any serious experimental evidence for black holes was available.

[48] In loose terms this means "*looks like trivial data at large distances*", i.e. $\not{g} \to \eta$ for $r \to \infty$, where r is the radial coordinate. For more details see [Wal84, Chapter 11].

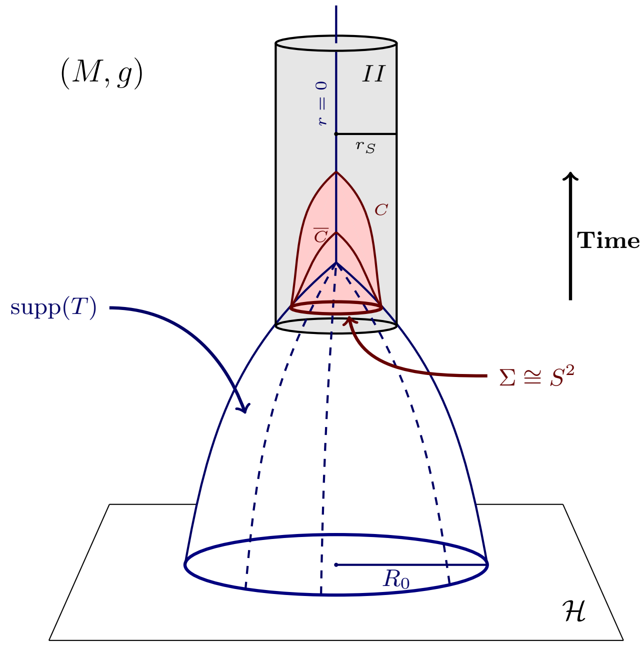


Figure 38: Spacetime model (M, g) for spherically symmetric gravitational collapse with initial Cauchy surface $\mathcal{H}$. The energy-matter content, i.e. supp(T), is contained in the blue cone. After the surface radius contracts below the Schwarzschild radius $r_S = 2m$, the black hole region II forms (cf. Figure 37) followed by the appearance of the Schwarzschild singularity at $r = 0$. In II one has a trapped surface Σ with its outgoing/ingoing null geodesic congruences C and $\overline{C}$ respectively. Note that one angular direction is suppressed in this picture. See also [Pen65, Fig. 1] for a more detailed sketch.

congruences converge locally in the future, i.e. Σ is a trapped surface.

Penrose demonstrated that —much more general— the presence of such a trapped surface is sufficient to guarantee (null) incompleteness. Remarkably, the argument makes no use of symmetry assumptions. In particular, when viewed in the context of the Oppenheimer-Snyder model, this means that perturbations away from spherical symmetry do not eliminate the breakdown of geodesic extendibility[49]. The significance of this cannot be understated: The Penrose Incompleteness Theorem has established that incompleteness is an unavoidable feature of General Relativity.

8.3 The Incompleteness Theorem

While the history surrounding the incompleteness Theorem is certainly deeply rooted in General Relativity, it is nevertheless a rigorous mathematical result of Lorentzian geometry. As a matter of fact, ever since its introduction, the Theorem has become an integral part of ongoing mathematical research and has been subjected to numerous generalizations including

[49]This reflects the fact that trapped surfaces are in some sense *stable* objects, since their definition involve strict inequalities.

to settings of lower regularity (cf. [Ste23]). This final section is dedicated to the introduction and the proof of the classic Incompleteness Theorem as well as a one sided version. We will primarily follow [Gal14, Chapter 7] and also [Daf21].
We may now state the main Theorem of this work as follows

Theorem 8.3.1 (Penrose)**.** *Let (M, g) be a globally hyperbolic spacetime such that it*

(i) satisfies the null energy condition, i.e. $\mathrm{Ric}(X, X) \geq 0$ *for any null vector X,*

(ii) has a non-compact Cauchy surface,

(iii) contains a trapped surface Σ,

then M is future null incomplete.

Remark 8.3.2. In view of Chapter 8.2, the assumptions in the above Theorem are all very natural. First of all, any spacetime (M, g) that is the solution to the Einstein equations for some initial data $(\mathcal{H}, \not{g}, \mathrm{I\!I})$, features, by definition, $\mathcal{H}$ as a Cauchy surface. By Theorem 5.2.11, this is equivalent to (M, g) being globally hyperbolic. Moreover, in General Relativity one usually considers isolated systems with asymptotically flat initial data, i.e. the Cauchy surface will be non-compact.
The *null energy condition* is originally motivated by physics and defined as a constraint on the stress-energy-momentum tensor T (cf. [Rea24, Chapter 4.9]). It is a special case of the *weak energy condition*, $T(X, X) \geq 0$ for any causal X, and reads $T(X, X) \geq 0$ for X null instead. Observe that this assumption now simply emerges from the Einstein equation (EEQs) itself since

$$Ric(X, X) - \frac{1}{2} S \underbrace{g(X, X)}_{=0} + \Lambda \underbrace{g(X, X)}_{=0} = 8\pi\, T(X, X) \geq 0$$

for any X which is null. Finally, as discussed in the previous section, the existence of a trapped surface may now be viewed as the essential geometric assumption in the Penrose Incompleteness Theorem.

The proof of the above theorem is set up by contradiction. Assuming null completeness, one shows that the achronal boundary $\partial I^+(\Sigma)$ is compact using results from the variation of null geodesics and geometry of null hypersurfaces. At the same time, the fact that the null geodesic in $\partial I^+(\Sigma)$ stretch out far enough, implies that this set is in fact a Cauchy surface. After that, the result immediately follows since Cauchy surfaces are homeomorphic, i.e. there cannot exists compact and non-compact Cauchy surfaces simultaneously. The advantage of Galloway's proof, when compared to the classic arguments in the standard textbook literature (see e.g. [O'N83, Theorem 14.61]), is that it maintains the geometric intuition of the claim as one may see below.

Proof of Theorem 8.3.1 (Galloway). Assume to the contrary that M is future null complete. We want to show that the achronal boundary $\partial I^+(\Sigma)$ is compact. After that, Corollary 5.1.10

together with Proposition 5.2.23 imply that $\partial I^+(\Sigma)$ must be a compact Cauchy surface and applying Theorem 5.2.12 (i) then yields a contradiction.
Suppose now that $\partial I^+(\Sigma)$ is not compact. Because it is closed there exists a sequence $\{q_n\}_{n\in\mathbb{N}}$ with $q_n \in \partial I^+(\Sigma)$ for all $n \in \mathbb{N}$ such that there exists no convergent subsequence in M. Since M is globally hyperbolic, it follows that $J^+(\Sigma)$ is closed by Proposition 5.2.8(i). Furthermore, using Propositions 2.2.15 and 2.2.16, one finds

$$\partial I^\pm(\Sigma) = \partial J^\pm(\Sigma) = J^\pm(\Sigma) \setminus I^\pm(\Sigma). \tag{8.1}$$

Following from Proposition 4.3.5, there exists a sequence of future directed null geodesics $\{c_n\}_{n\in\mathbb{N}}$ between points $p_n \in \Sigma$ and q_n respectively, which may be parametrized such that $c_n : [0, T_n] \to \partial I^+(\Sigma)$. By Proposition 7.0.15, one has that the c_n emanate orthogonally from Σ at p_n for each $n \in \mathbb{N}$. Normalizing the initial velocities $c_n(0)$ in a suitable way[50] and using the compactness of Σ, it follows that the sequence $\{(p_n, c_n(0))\}_{n\in\mathbb{N}}$ is contained in a compact subset of the null normal bundle of Σ. Thus, there exists a subsequence $\{(p_m, c_m(0))\}_{m\in\mathbb{N}}$ converging to (p, v) with $p \in \Sigma$ and $v \in T_p\Sigma^\perp$ null. In this way, the sequence $\{c_m\}_{m\in\mathbb{N}}$ converges, in the sense of geodesics, to a future directed null geodesic $c : [0, \infty) \to M$ with $c(0) = p$ and $\Sigma \perp c$ which is contained in $\partial I^+(\Sigma)$.
Assume without loss of generality that c is outward-pointing with $\dot{c}(0) = L(p)$, where L denotes the outward pointing null vector field of the outgoing null geodesic congruence C. We claim that no other outward pointing null geodesic emanating from Σ can intersect c. Otherwise, by contradiction, let γ be a different outward pointing null geodesic that intersects c at some point $q \in \partial I^+(\Sigma)$. Define the null curve α by concatenating γ with c after q. It follows that α is contained in $\partial I^+(\Sigma)$. However, α cannot be a null pregeodesic because it features a corner at q, i.e. the null directions of γ and c have to be different by uniqueness of geodesics. Equation (8.1) and Proposition 4.3.5 imply that this is impossible.
Our next goal is to show that $c|_{[0,T]}$ is contained in a smooth null hypersurface for any $T > 0$ (see also Figure 39). To this end, consider the normal exponential map $\exp^\perp$, restricted to the one-dimensional null line bundle $\mathcal{L} \subset T\Sigma^\perp$ spanned by L.
By Theorem 7.0.14, Σ cannot have a focal point along c and thus, Proposition 7.0.11 implies that the map is an immersion in some neighbourhood of the ray $R := \{(p, tL(p)) \in \mathcal{L} \,|\, t \in [0, T]\}$. Using a local trivialization for simplicity, we may now assume that $\exp^\perp : U \to M$ is an immersion, where $U \subset \mathbb{R}^{n-1} \times \mathbb{R}$ open and $c|_{[0,T]} \subset \exp^\perp(U)$.
Following this, we aim to demonstrate that $\exp^\perp$ is also injective in some neighbourhood of R contained in U. Again, for the sake of contradiction, assume that no such neighbourhood exists. Let V be a relatively compact neighbourhood of R in U. One may now construct a sequence $\{z_n\}_{n\in\mathbb{N}}$ with $z_n \to c(t)$ for some $t \in [0, T]$ together with sequences $\{(x_n, v_n)\}_{n\in\mathbb{N}}$ and $\{(y_n, w_n)\}_{n\in\mathbb{N}}$ in V such that $\exp^\perp(x_n, v_n) = z_n = \exp^\perp(y_n, w_n)$ and $(x_n, v_n) \neq (y_n, w_n)$ for each $n \in \mathbb{N}$.

[50] In a local trivialization, we could, for example, consider $\|c_n(0)\| = 1$ with respect to the Euclidean norm.

This may be achieved as follows: Let h be an auxiliary complete Riemannian metric on M. For $\epsilon > 0$ let N_ϵ be the ϵ-neighbourhood of $c([0,T])$ with respect to h, which is defined as

$$N_\epsilon := \{q \in M \mid \inf_{t\in[0,T]} d_h(q, c(t)) < \epsilon\}$$

and consider the family $\{N_{1/n}\}_{n\in\mathbb{N}}$. By assumption, for each $n \in \mathbb{N}$ one has that $\exp^\perp$ is non-injective on the domain $V \cap \left(\exp^\perp\right)^{-1}(N_{1/n})$. Hence, there exist $z_n \in N_{1/n}$ and $(x_n, v_n), (y_n, w_n) \in V$ with $(x_n, v_n) \neq (y_n, w_n)$ and $\exp^\perp(x_n, v_n) = z_n = \exp^\perp(y_n, w_n)$. Since $N_{1/(n+1)} \subset N_{1/n}$, the sequence is contained in $\overline{N}_1$. Because $c([0,T])$ is compact, $\overline{N}_1$ is bounded and using Hopf-Rinow also compact. Consequently, we may pass to a convergent subsequence, still denoted $\{z_n\}_{n\in\mathbb{N}}$ and with $z_n \to z \in \overline{N}_1$. By continuity and the fact that $\bigcap_{n\in\mathbb{N}} N_{1/k} = c([0,T])$, it follows that $\inf_{t\in[0,T]} d_h(z_n, c(t)) \to \inf_{t\in[0,T]} d_h(z, c(t)) = 0$. Therefore, there exists $t \in [0,T]$ with $z = c(t)$.
By compactness of $\overline{V} \subset U$, we pass to subsequences $\{(x_k, v_k)\}_{k\in\mathbb{N}}$ and $\{(y_k, w_k)\}_{k\in\mathbb{N}}$ with $(x_k, v_k) \to (x, v) \in U$ and $(y_k, w_k) \to (y, w) \in U$. By continuity one has $\exp^\perp(x, v) = c(t) = \exp^\perp(y, w)$ and since $c(t) = \exp^\perp(p, tL(p))$, we may consider the following two cases

(i) If $(x, v) = (p, tL(p)) = (y, w)$, then any neighbourhood of $(p, tL(p))$ will admit infinitely many points $(x_k, v_k) \neq (y_k, v_k)$ with $\exp^\perp(x_k, v_k) = \exp^\perp(y_k, w_k)$. But because $\exp^\perp\big|_U$ is an immersion, it is locally injective and we arrive at a contradiction.

(ii) Otherwise, at least one of (x, v) or (y, w) is different from $(p, tL(p))$. This is impossible because no other outward pointing null geodesic emanating from Σ can intersect c following from the argument before.

In conclusion, shrinking U if necessary, one has that $\exp^\perp\big|_U$ is an injective immersion.
Finally, we prove that $\exp^\perp : U \to \exp^\perp(U)$ is also a homeomorphism. Since it is a bijective continuous map into a first countable Hausdorff space, it is sufficient[51] to show that it is also proper. We may assume without loss of generality that U is itself relatively compact but still includes R (otherwise shrink U again). In this case, every preimage $\left(\exp^\perp\right)^{-1}(K) \subset \overline{U}$ of a compact set $K \in \exp^\perp(U)$ is closed by continuity and therefore compact.
Thus, $\exp^\perp$ is proper and in fact an embedding. Consequently, $c([0,T])$ is contained in the smooth submanifold $H := \exp^\perp(U)$ of dimension n, i.e. H is a smooth hypersurface. Furthermore, repeating the exact same argument from Remark 8.1.3, it follows that H is null.

Now, because Σ is trapped, $\not\theta(p) < 0$. Picking $T > \frac{n-1}{|\not\theta(p)|}$, we may now argue exactly as in the proof of Proposition 6.2.9. Using the null energy condition and the Raychauchdhuri equation (6.2), one again obtains the following inequality for the null mean curvature along c,

$$\theta' \leq -\frac{1}{n-1}\theta^2.$$

[51] Cf. [Lee11, Corollary 4.97].

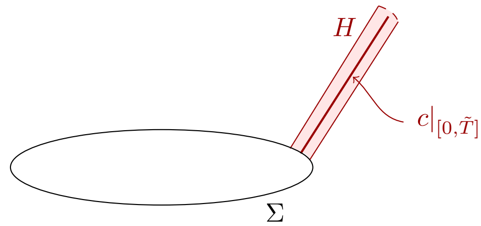


Figure 39: A segment of the null generator $c \subset \partial I^+(\Sigma)$ is contained in a (possibly very narrow) null hypersurface H, which is constructed by the image of the normal exponential map restricted to the null direction $\dot{c}(0)$ in a neighbourhood of $c(0)$.

Since $\theta(0) = \not{\theta}(p) < 0$, the above implies that $\theta(t) < 0$ for any $t \in [0, T]$. Also

$$\frac{d}{dt}\left(\frac{1}{\theta}\right) = -\frac{1}{\theta^2}\theta' \geq \frac{1}{n-1}$$

and $\frac{1}{\theta(t)} < 0$ must intersect 0 in finite parameter time $t_0 > 0$. Integrating the above inequality from 0 to t_0, one finds

$$-\frac{1}{\theta(0)} \geq \frac{t_0}{n-1} \qquad \Leftrightarrow \qquad t_0 \leq \frac{n-1}{|\theta(0)|}.$$

Thus $t_0 < T$, which contradicts the smoothness of H. Consequently $\partial I^+(\Sigma)$ must be in fact compact and the claim follows as discussed at the beginning. □

Remark 8.3.3. From the General Relativity viewpoint of Chapter 8.2, the above argument features a certain drawback. Namely, that it does not tell us anything about the nature of the incompleteness. Looking at the Oppenheimer-Snyder model, it would seem rather tempting to connect incompleteness directly to the formation of singularities in some generic way. However, this is not the only option. It could also be the case that the incompleteness is a result of (M, g) being extendible as a $\mathcal{C}^2$-spacetime. By this we mean that there could exist a larger spacetime $(\overline{M}, \overline{g})$ that features a smaller subset isometric to (M, g). Note that this implies that M has a Cauchy horizon in $\overline{M}$ (recall Definition 5.2.16 and Figure 18). This is troublesome from a physics perspective because any point $\overline{p} \in \overline{M} \setminus M$ is not determined by prescribed initial data and what happens outside of M cannot be predicted by the theory. This situation is most notably featured in the famous *Kerr* solution to the Einstein vacuum equations which models a rotating black hole. No future directed causal geodesic in the Kerr spacetime encounters a singularity and one could just pass over to any suitable extension. Even so, Penrose himself somewhat found a remedy for this in the form of a blue-shift that shows that the Cauchy horizon is unstable to perturbations (cf. [Pen68, p. 222f] and [Wal84, Chapter 12]). Nevertheless, the issue of extendibility remains an open problem and is subject to one of the so-called *Cosmic Censorship Conjectures* (cf. [Rin09, Chapter 17.2]).

For the final result of this work, we aim to establish a one-sided version of the Penrose Incompleteness Theorem. Unlike the classical version 8.3.1, the following theorem assumes that Σ is only outer trapped but additionally separates a smooth spacelike Cauchy surface into two disjoint subsets (see Figure 40). In comparison to the previous proof, here we will only obtain compactness in one direction, which leads to some additional technical difficulties. We follow [AMMS09] and in turn [Tot94].

Theorem 8.3.4. *Let (M, g) be a globally hyperbolic spacetime with a smooth spacelike Cauchy surface V and suppose that M*

(i) satisfies the null energy condition, i.e. $\operatorname{Ric}(X, X) \geq 0$ *for any null vector X,*

(ii) contains a closed smooth hypersurface Σ in V which separates V into an "inside" U and "outside" W, i.e. $V \setminus \Sigma = U \cup W$ with $U, W \subseteq V$ connected and $U \cap W = \varnothing$

If $\overline{W}$ is non-compact and Σ is outer trapped then M is future null geodesically incomplete.

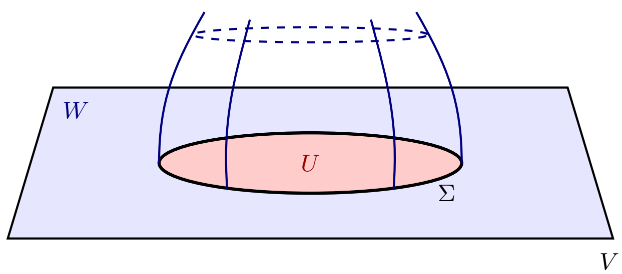


Figure 40: The situation in Theorem 8.3.4.

Proof. Let U, W, Σ be as described. Assume that M is null complete and let $L, \overline{L} \in \mathfrak{X}^{\perp}(\Sigma)$ be the outgoing and ingoing null normals of Σ. We want to argue similarly as before. However, since Σ is only outer trapped, only part of the achronal boundary $\partial I^+(\Sigma)$ is expected to be compact. Our goal is to use a projection argument to obtain a contradiction using the non-compactness of $\overline{W}$. To this end, we first want to show that $\partial I^+(\Sigma) \setminus \Sigma$ splits into two disjoint sets. Since Equation (8.1) holds, it follows that $\partial I^+(\Sigma) \setminus \Sigma = H \cup \overline{H}$ where H and $\overline{H}$ are subsets of the null geodesic congruences C and $\overline{C}$, generated by L and $\overline{L}$ respectively. Let $p \in H$ and let $c_1 : [0, 1] \to \partial I^+(\Sigma)$ be the null geodesic with $c_1(1) = p$, $c_1(0) \in \Sigma$ and $\dot{c}_1(0)$ tangent to L. Suppose that there exists a different null geodesic $c_2 : [0, 1] \to \partial I^+(\Sigma)$ with $c_2(0) \in \Sigma$, $\dot{c}_2(0)$ tangent to $\overline{L}$ and such that $c_2(1) = p$ as well. Define the continuous curve $c : [0, 2] \to \partial I^+(\Sigma)$ by

$$c(t) = \begin{cases} c_1(t) & t \in [0, 1], \\ c_2(2 - t) & t \in [1, 2]. \end{cases}$$

At this point, let $T \in \mathfrak{X}(M)$ be the timelike vector field that defines the time orientation on

M and consider the projection operator $\rho^T : M \to V$ as defined[52] in Theorem 5.2.12. Note that because H is outward pointing and $\overline{H}$ is inward pointing, one has that $\overline{H} \subseteq J^+(U)$ and $H \subseteq J^+(W)$. Thus, defining $\tilde{c}(t) \coloneqq \rho^T(c(t))$, there exists an $\epsilon > 0$ such that $\tilde{c}(t) \in W$ for $t \in (0, \epsilon)$ and $\tilde{c}(t) \in U$ for $t \in (2-\epsilon, 2)$. Because Σ separates U and W, continuity of $\tilde{c}$ implies that there must exist a $t_0 \in [\epsilon, 2-\epsilon]$ with $\tilde{c}(t_0) \in \Sigma$. But this is impossible, since then $c(t_0) \in \partial I^+(\Sigma) \cap I^+(\Sigma)$, which must be empty following from Equation (8.1). Therefore, we conclude that $\partial I^+(\Sigma) \setminus \Sigma$ splits into the two disjoint sets H and $\overline{H}$.

Since Σ is outer trapped, one may now use the same argument from the previous proof to infer that the set $H \cup \Sigma$ is compact. Note now that since V is smooth, the same argument as in Lemma 6.2.5 implies that $\mathrm{Fl}^T : \mathbb{R} \times V \to M$ is a diffeomorphism and that ρ^T is in fact smooth. Using the previous identification, one may write $\rho^T(t, y) = y$. Calculating the Jacobian at any q, we see that the kernel is given by $\mathrm{span}(T(q))$, which contains only timelike vectors.

On the other hand, consider the restriction $\rho \coloneqq \rho^T\big|_H : H \to V$. Note that $\rho(H) \subseteq W$ because $H \subseteq J^+(W)$. Continuity of ρ^T implies that $\rho^T(H \cup \Sigma) = \rho(H) \cup \Sigma \subseteq \overline{W}$ is compact. But since $\overline{W}$ is non-compact one then must have $\rho(H) \subset W$. Connectedness of W now implies that $\rho(H)$ has boundary points different from Σ in W as follows: For the sake of contradiction suppose that $\partial\rho(H) \cap W = \varnothing$. Thus, $\rho(H)$ is open in W and also closed because $\rho(H) = W \cap (\rho(H) \cup \Sigma)$. Since $\rho(H)$ is nonempty and W is connected, it follows that $W = \rho(H)$. However, this is impossible because $\overline{W} = W \cup \Sigma$ is by assumption non-compact. Consider now such a boundary point $x \in \partial\rho(H) \cap W$ and let $q \in H$ be such that $\rho(q) = x$. Since any neighbourhood of x will contain points not in $\rho(H)$, it is not possible for ρ to be a local diffeomorphism[53] at $q \in H$. This implies that $T_q\rho$ is singular and that there must exist a nontrivial $v \in T_qH$ in the kernel, i.e. $T_q\rho(v) = 0$. But since ρ is simply the restriction of ρ^T, whose kernel at q is spanned by $T(q)$, one obtains a contradiction as v is either null or spacelike. Consequently, M must be null incomplete. □

[52]Recall that ρ^T projects points onto V via the integral curves of T.

[53]While H is not globally smooth a priori, it is in fact a (local) Lipschitz manifold (cf. [Min19, Thm. 2.87]. By Rademacher's Theorem, the tangent spaces T_qH exist for any regular $q \in H$, which together form a dense subset of H, i.e. we may treat H as $\mathcal{C}^1$ (or in fact smooth) here (see also [GMMM11, Appendix A]).

References

[AMMS09] Lars Andersson, Marc Mars, Jan Metzger, and Walter Simon. The time evolution of marginally trapped surfaces. *Classical and Quantum Gravity*, 26(8):085018, 2009. doi: 10.1088/0264-9381/26/8/085018.

[Are18] Stefanos Aretakis. General relativity, 2018. Lecture Notes from June 30th, https://www.math.toronto.edu/aretakis/General%20Relativity-Aretakis.pdf, Last opened: 04.09.2025.

[Arn92] Vladimir Igorewitsch Arnold. *Ordinary Differential Equations*. Springer-Verlag, 3st edition, 1992. Translated by Roger Cooke.

[BEE96] John K. Beem, Paul E. Ehrlich, and Kevin L. Easley. *Global Lorentzian Geometry*. Marcel Dekker, Inc., 2nd edition, 1996.

[Bir23] George David Birkhoff. *Relativity and Modern Physics*. Harvard University Press, Cambridge, MA, 1923.

[BS05] Antonio N. Bernal and Miguel Sánchez. Smoothness of time functions and the metric splitting of globally hyperbolic spacetimes. *Communications in Mathematical Physics*, 257(1):43–50, April 2005. doi: 10.1007/s00220-005-1346-1.

[BS07] Antonio N. Bernal and Miguel Sánchez. Globally hyperbolic spacetimes can be defined as 'causal' instead of 'strongly causal'. *Classical and Quantum Gravity*, 24(3):745–749, January 2007. doi: 10.1088/0264-9381/24/3/n01.

[Bä20] Christian Bär. Lorentzian geometry, 2020. Lecture Notes, Summer Term 2004 at University of Potsdam, https://www.math.uni-potsdam.de/fileadmin/user_upload/Prof-Geometrie/Dokumente/Lehre/Veranstaltungen/WS0405-SS08/LorentzianGeometryEnglish13Jan2020.pdf, Last opened: 04.09.2025.

[CB52] Yvonne Choquet-Bruhat. Théorème d'existence pour certains systèmes d'équations aux dérivées partielles non linéaires. *Acta Mathematica*, 88:141–225, 1952. doi: 10.1007/BF02392131.

[CBG69] Yvonne Choquet-Bruhat and Robert Geroch. Global aspects of the cauchy problem in general relativity. *Communications in Mathematical Physics*, 14:329–335, December 1969. doi: https://doi.org/10.1007/BF01645389.

[Chr99] Demetrios Christodoulou. On the global initial value problem and the issue of singularities. *Classical and Quantum Gravity*, 16(12A):A23, December 1999. doi: 10.1088/0264-9381/16/12A/302.

[CK93] Demetrios Christodoulou and Sergiu Klainerman. *The Global Nonlinear Stability of the Minkowski Space*, volume 41 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 1993.

[Daf21] Mihalis Dafermos. London mathematical society newsletter. 493:27–34, March 2021. https://www.lms.ac.uk/sites/lms.ac.uk/files/files/NLMS_493_for%20web2.pdf, Last opened: 04.09.2025.

[dC92] Manfredo Perdigão do Carmo. *Riemannian Geometry*. Birkhäuser, 1st edition, 1992. Translated by Francis Flaherty.

[DR05] Mihalis Dafermos and Igor Rodnianski. A proof of price's law for the collapse of a self-gravitating scalar field. *Inventiones mathematicae*, 162:381–457, July 2005. doi: https://doi.org/10.1007/s00222-005-0450-3.

[EG15] Lawrance C. Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, revised edition, 2015.

[Gal14] Gregory J. Galloway. Notes on lorentzian geometry, 2014. ESI-EMS-IAMP Summer School on Mathematical Relativity, https://www.math.miami.edu/~galloway/vienna-course-notes.pdf, Last opened: 04.09.2025.

[Ger68] Robert Geroch. What is a singularity in general relativity? *Annals of Physics*, 48(3):526–540, July 1968. doi: https://doi.org/10.1016/0003-4916(68)90144-9.

[Ger70] R. Geroch. Domain of dependence. *Journal of Mathematical Physics*, 11(2), February 1970. doi: https://doi.org/10.1063/1.1665157.

[GMMM11] F. Gesztesy, I. Mitrea, D. Mitrea, and M. Mitrea. On the nature of the laplace–beltrami operator on lipschitz manifolds. *Journal of Mathematical Sciences*, 172:279–346, 2011. doi: 10.1007/s10958-010-0199-0.

[Gre75] Werner Greub. *Linear Algebra*. Springer-Verlag New York, Inc., 4th edition, 1975.

[Hat02] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 1st edition, 2002.

[HE73] S. W. Hawking and G. F. R. Ellis. *The Large Scale Structure of Space-Time*. Cambridge University Press, 1st edition, 1973.

[Hic63] Noel Hicks. Submanifolds of semi-riemannian manifolds. *Rendiconti del Circolo Matematico di Palermo*, 12:137–149, May 1963. doi: https://doi.org/10.1007/BF02843960.

[HM19] R. A. Hounnonkpe and E. Minguzzi. Globally hyperbolic spacetimes can be defined without the 'causal' condition. *Classical and Quantum Gravity*, 36(19):197001, 2019. doi: 10.1088/1361-6382/ab3f11.

[Jeb21] Jørg Tofte Jebsen. Über die allgemeinen kugelsymmetrischen lösungen der einsteinschen gravitationsgleichungen im vakuum. *Arkiv for Matematik, Astronomi og Fysik*, 15(18):1–9, 1921.

[Kri99] Marcus Kriele. *Spacetime, Foundations of General Relativity and Differential Geometry*. Springer, 1st edition, 1999.

[KS18] Michael Kunzinger and Clemens Sämann. Lorentzian length spaces. *Annals of Global Analysis and Geometry*, 54(3):399–447, 2018. doi: 10.1007/s10455-018-9633-1.

[KS21] Michael Kunzinger and Roland Steinbauer. Lecture notes in riemannian geometry, 2021. https://www.mat.univie.ac.at/~mike/teaching/ws2223/rg.pdf, Last opened: 04.09.2025.

[KS23] Michael Kunzinger and Roland Steinbauer. Lecture notes in lorentzian geometry, 2023. https://www.mat.univie.ac.at/~mike/teaching/ss23/lorentzian.pdf, Last opened: 04.09.2025.

[Kun22a] Michael Kunzinger. Lecture notes in analysis on manifolds, 2022. https://www.mat.univie.ac.at/~mike/teaching/ss22/amf.pdf, Last opened: 04.09.2025.

[Kun22b] Michael Kunzinger. Lecture notes in analysis on manifolds, 2022. https://www.mat.univie.ac.at/~mike/teaching/ss22/amf.pdf, Last opened: 04.09.2025.

[Kup87] Demir N. Kupeli. On null submanifolds in spacetimes. *Geometriae Dedicata*, 23:33–51, May 1987. doi: https://doi.org/10.1007/BF00147389.

[Lee11] John M. Lee. *Introduction to Topological Manifolds*. Springer, 2nd edition, 2011.

[Lee13] John M. Lee. *Introduction to Smooth Manifolds*. Springer, 2nd edition, 2013.

[Lee18] John M. Lee. *Introduction to Riemannian Manifolds*. Springer, 2nd edition, 2018.

[MG10] Thomas Mueller and Frank Grave. Catalogue of spacetimes. 2010. arXiv:0904.4184 [gr-qc], https://arxiv.org/abs/0904.4184.

[Min08] E. Minguzzi. Limit curve theorems in lorentzian geometry. *Journal of Mathematical Physics*, 49(9), September 2008. doi: https://dx.doi.org/10.1063/1.2973048.

[Min19] E. Minguzzi. Lorentzian causality theory. *Living Reviews in Relativity*, 22(3), 2019. doi: 10.1007/s41114-019-0019-x.

[MS08] Ettore Minguzzi and Miguel Sánchez. The causal hierarchy of spacetimes. 2008. arXiv:gr-qc/0609119, version 3, https://arxiv.org/abs/gr-qc/0609119.

[Mun14] James Munkres. *Topology*. Pearson Education Limited, 2nd edition, 2014.

[NO61] Katsumi Nomizu and Hajime Ozeki. The existence of complete riemannian metrics. *Proceedings of the American Mathematical Society*, 12(6):889–891, 1961. doi: 10.1090/S0002-9939-1961-0133785-8.

[O'N83] Barrett O'Neill. *Semi-Riemannian Geometry*. Academic Press, 1st edition, 1983.

[OS39] J. R. Oppenheimer and H. Snyder. On continued gravitational contraction. *Physical Review*, 56(5):455–459, 1939. doi: 10.1103/PhysRev.56.455.

[Pen65] Roger Penrose. Gravitational collapse and space-time singularities. *Physical Review Letters*, 14(3):57–59, 1965. doi: 10.1103/PhysRevLett.14.57.

[Pen68] Roger Penrose. Structure of space-time. In C. M. DeWitt and J. A. Wheeler, editors, *Battelle Rencontres: 1967 Lectures in Mathematics and Physics*, pages 121–235. W. A. Benjamin, New York, 1968.

[Rea24] Harvey Reall. Lecture notes on black holes, March 2024. https://www.damtp.cam.ac.uk/user/hsr1000/black_holes_lectures_2024.pdf, Last opened: 04.09.2025.

[Rin09] Hans Ringström. *The Cauchy Problem in General Relativity*. European Mathematical Society, 1st edition, 2009.

[Sbi16] Jan Sbierski. On the existence of a maximal cauchy development for the einstein equations: a dezornification. *Annales Henri Poincaré*, 17:301–329, 2016. doi: 10.1007/s00023-015-0401-5.

[Spi99] Michael Spivak. *A Comprehensive Introduction to Differential Geometry, Volume Four*. Publish or Perish, Inc., Houston, TX, 3 edition, 1999.

[Ste23] Roland Steinbauer. The singularity theorems of general relativity and their low regularity extensions. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 125(2):73–119, 2023. doi: 10.1365/s13291-022-00263-7.

[Tot94] G. Totschnig. Ein singularitätentheorem für outer trapped surfaces. Master's thesis, Universität Wien, 1994.

[Wal84] Robert M. Wald. *General Relativity*. Univ. of Chicago Press, 1st edition, 1984.