Universitat de Barcelona

Facultat de Matemàtiques i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

Emergence of Causality from the Geometry of Spacetimes

Autor: Roberto Forbicia León

Directora: Dra. Joana Cirici

Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 21 de Juny de 2020

Abstract

In this work, we study how the notion of causality emerges as a natural feature of the geometry of spacetimes. We present a description of the causal structure by means of the causality relations and we investigate on some of the different causal properties that spacetimes can have, thereby introducing the so-called causal ladder. We pay special attention to the link between causality and topology, and further develop this idea by offering an overview of some spacetime topologies in which the natural connection between the two structures is enhanced.

Resum

En aquest treball s'estudia com la noció de causalitat sorgeix com a característica natural de la geometria dels espaitemps. S'hi presenta una descripció de l'estructura causal a través de les relacions de causalitat i s'investiguen les diferents propietats causals que poden tenir els espaitemps, tot introduint l'anomenada escala causal. Es posa especial atenció a la connexió entre causalitat i topologia, i en particular s'ofereix un resum d'algunes topologies de l'espaitemps en què aquesta connexió és encara més evident.

2020 Mathematics Subject Classification. 83C05,58A05,53C50

Acknowledgements

First and foremost, I would like to thank Dr. Joana Cirici for her guidance, advice and valuable suggestions. I would also like to thank my friends for having been willing to listen to me talk about spacetime geometry. And finally, I would like to thank my family for having always looked after my education.

Contents

1 Introduction

Causality is a main feature of human cognition. We are familiar with cause-effect relationships and we continuously experience them in our everyday life. They help us understand our surroundings and make decisions based on desired outcomes. If it rains and we need to go out, we take an umbrella because we know there is a causal relationship between walking under the rain without an umbrella and getting wet. We use them to infer knowledge that is beyond our immediate perception: if it has been raining and we left our clothes out to dry, we know that they will be wet, even though we might not be at home to actually see it. The issue of how knowledge may be obtained from cause-effect relationships -and the limitations of this method- has been a major philosophical concern for centuries, and still is nowadays (on [Ebe09] one can find an interesting review on this topic). These discussions, that enter the domains of epistemology, have led to causal theories of knowledge (for instance, the one provided in [Gol67]).

The notion of causality is at the heart of the defining characteristic of any scientific theory: predictability, and therefore is present in all branches of Science, to a greater or lesser extent. However, in most cases this presence consists mainly on studying which processes are causally related and why. The fact that causal relationships simply exist is taken as an empirical truth (which in fact is). But one could actually wonder why causality exists in the first place. In a first approach, one realises that there is a strong connection between causality and the intuitive notion of the flow of time. For example, one is familiar with the following cause-effect relation: "at constant temperature, increasing the pressure of an ideal gas causes a decrease in its volume". However, we could have equally said: "at constant temperature, decreasing the volume of an ideal gas causes an increase of its pressure". The ambiguity of this sort of causal relationship is noted in [GPS05] and helps illustrate an essential point: cause-effect relationships cannot be fully described without a notion of *time orientation*. Thus, understanding causality is intrinsically linked to the arduous task of understanding the nature of time itself.

Physicists have established, at least in our region of the Universe, the existence of an *arrow of time* determined by the direction of increase of entropy, that discriminates between the concepts of *future* and *past*. This notion is assumed in most physical theories and is used to model the Universe, or a part of it. There have been many attempts to do so, but the standard ones nowadays all rely on the concept of *spacetime*, introduced by H. Minkowski in 1908. The essential idea is to merge the 3 usual space dimensions with the time one in such a way that the particular character of the latter, namely the existence of an arrow of time, is preserved. The spacetime formalism developed by Minkowski, known as Minkowski spacetime, offered an elegant and useful way of presenting A. Einstein's theory of Special Relativity (SR), that had been published in 1905. This model, as well as all that have followed, is based in the description of *events* which occur in the Universe and the study of the relationships between them. The term "event" is to be understood, in an idealised sense, as a physical occurrence that has no spatial extension or duration in time. One can imagine, for example, an instantaneous collision or an instant in the trajectory of some particle.

As the framework of SR, Minkowski spacetime is a model of the Universe that does not account for gravitational phenomena. Due to this fact, it has a quite easy to deal with mathematical structure, namely that of a 4-dimensional Lorentzian vector space. The theory of SR was generalised some years later by A. Einstein himself in order to describe gravitational phenomena as well, leading to the publication in 1915 of the theory of General Relativity (GR). The latter is based in two principles. One is the Principle of Equivalence, which states that at every point of spacetime one can choose a locally itertial reference frame according to which the effects of gravity are absent and therefore spacetime behaves locally as Minkowski's. The other one is the Principle of General Relativity, which asserts that the laws of Physics are the same for all reference frames. From these two postulates follows a geometric theory of spacetime in which the latter is regarded as a 4-dimensional Lorentzian manifold on which gravity acts by means of the metric tensor (more precisely, the metric tensor is physically interpreted as the gravitational potential). Therefore, GR inevitably links Physics to differential and semi-Riemannian geometry.

It is one of the main goals of this work to understand why the particular mathematical structure of a Lorentzian manifold is the most suitable for the purpose of representing the physical spacetime. We shall show that it is precisely the necessity to account for causality that mostly motivates this choice. This will lead to the definition of familiar concepts such as future, past and causality itself in purely mathematical terms. By doing so, we will be able to address the following question: given a certain spacetime, can we determine the nature of the causal relationships that may take place between its events? To answer this, we will shape its *causal structure*, which is a mathematical feature inherent to any Lorentzian manifold based on the classification of its tangent vectors as timelike, null or spacelike, that is, on their *causal character*. Curves on Lorentzian manifolds may also have a causal character which determines whether they can be a good candidate to represent the evolution of physical particles. In this way, the study of which events in the Universe can be causally related is reduced to the study of which points in a spacetime can be joined by what we will call a *causal curve*.

At some point we will have to address the issue of how to avoid pathological causal behaviours in a spacetime, such as the possibility to travel in time. A way to do so, for example, is to exclude spacetimes with *closed causal curves*. This will motivate the introduction of the so-called *causality conditions*, and we shall see how they will naturally classify spacetimes in a *causal ladder* according to how physical their causal behaviour is.

The ambition to present a mathematical description of causality and the will to do it in a self-contained way sets out another goal of this work: that of offering a mathematically rigorous presentation of the geometry of spacetimes. Thus, we shall give further mathematical insight into GR that will allow us, for example, to give a mathematical formulation of the Principle of Equivalence. In the sake of brevity, however, we have decided not to include the notions of the Riemann and Ricci tensors and the corresponding discussion about curvature. Although it is part of any standard textbook on semi-Riemannian geometry or offering a mathematical approach to GR, this discussion has no direct application, at the level of this work, in the presentation of causality that we want to carry out.

The study of causal properties of general relativistic spacetimes was first addressed in the 1970's by R. Penrose in [Pen72] and by S. W. Hawking and G. F. R. Ellis in [HE73], driven by their motivation of describing spacetime singularities and black holes. Their work resulted in the so-called Singularity Theorems and laid the groundwork for the study of causality in any spacetime by introducing the *causality relations* and the *causality conditions*. Our discussion on causality will be mainly focused on defining these concepts with a particular interest in how they are linked to the spacetime topology. It is precisely the study of this connection that has motivated the introduction of new topologies for spacetimes ([Zee66], [HKM76], [Ful92], among others) that are intimately related to the causal structure. We would also like to stress that the study of causality on spacetimes is a very vast topic and a current field of research. As an example, many efforts have been put in the last years in defining causal relations from a purely topological or even order-theoretical approach, without having to rely on the Lorentzian metric tensor. This approach has led to the so-called *causal set theory* (see for instance [GPS05] or [SJ14]) and its motivation is that of accounting for causality in the framework of quantum gravity theories, most of which are free of the metric tensor.

Finally, let us briefly comment on how this work is organised. In Section 2, a mathematical description of Minkowski spacetime is offered, with special attention on the emergence of its causal structure in terms of the *causal cones.* Section 3 is a review of the topics on differential and semi-Riemannian geometry that are required for a proper understanding of the geometry of spacetimes, with a focus on those that are essential in the description of causality. This includes the notion of geodesic, normal coordinates and the exponential map. Essentially these three tools will allow us to study the causal properties of general spacetimes in Section 4, by relying heavily on the causal structure of Minkowski spacetime. Finally, Section 5 is devoted to a general overview of some spacetime topologies that are physically more appealing and that are strongly related to the causal structure.

All together, it is our hope that this work will offer a solid introduction to the geometry of spacetimes and their causal structure and properties, without requiring further previous knowledge on the topic than basic point-set topology.

2 Minkowski spacetime

The goal of this section is to offer a mathematically rigorous description of Minkowski spacetime, with a focus on its causal structure.

As we have introduced, Minkowski spacetime is generally regarded as the appropriate setting within which to formulate those laws of Physics that do not refer specifically to gravitational phenomena. From a purely mathematical perspective, Minkowski spacetime is basically a real 4-dimensional Lorentzian vector space. The motivations for the choice of this particular structure have a profound physical meaning that we will try to expose. The starting point in the construction of Minkowski spacetime is to consider an abstract set representing the collection of all possible events. To this aim, it seems reasonable to consider $\mathbb{R}^4$ as the simplest candidate, since according to our experience events are characterised by one time coordinate and three spatial coordinates. Then, we shall provide a mathematical structure that allows to satisfactorily describe the results of experimental physics and to reproduce the main physical features of the universe. It must reflect, for instance, the apparent existence of an arrow of time discriminating between the human concepts of "future" and "past", thereby giving rise to the notion of causality. In fact, it is precisely this necessity that entirely motivates the choice of a Lorentzian vector space structure for $\mathbb{R}^4$. As we will see, the properties of Lorentzian vector spaces that are discussed in Section 2.1, and that differ substantially from those of Euclidean vector spaces, allow the classification of events depending on whether they can be causally related or not, thus endowing Minkowski spacetime with a causal structure.

The mathematical approach to Minkowski spacetime is a widely covered topic in the literature (see for instance [Nab12] and [O'N83]). Our approach, specially regarding the introduction of the causal structure, will be that of [Nab12], that is perhaps more elegant from a mathematical point of view. Standard references for this topic are usually accompanied by the introduction of the *Lorentz group*, namely the group of isometries of Minkowski spacetime. Although it is a very interesting discussion, again we have not included it here for brevity and because of its lack of direct application for our purposes.

2.1 Lorentzian vector spaces

Let us begin by reviewing some basic notions of semi-Euclidean geometry. These are basically analogous to the ones in a standard course of Euclidean geometry, with the difference that the positive-definiteness of the inner product is not required. We shall only present the results that will be used throughout the work, in the generality needed, and skipping some proofs for the sake of brevity. For a deeper discussion of the topic we refer the reader to the main references for this section, which are [Nab12], [SW77] and [O'N83].

In what follows, E will denote an arbitrary real vector space of dimension $n \geq 1$.

Definition 2.1. A *bilinear form* on E is a map $g : E \times E \to \mathbb{R}$ such that

$$g(av + bu, w) = ag(v, w) + bg(u, w) \text{ and } g(v, aw + bu) = ag(v, w) + bg(v, u)$$

for every $v, w, u \in E$ and every $a, b \in \mathbb{R}$. Such a bilinear form g is said to be:

(i) *symmetric* if $g(v, w) = g(w, v)$ for all $v, w \in E$.

(ii) *non-degenerate* if $g(v, w) = 0$ for all $w \in E$ implies $v = 0$.

Definition 2.2. An *inner product on* E is a bilinear form $g : E \times E \to \mathbb{R}$ that is symmetric and non-degenerate. An inner product g is said to be *positive-definite* (resp. *negative-definite*) if $g(v, v) > 0$ (resp. $g(v, v) < 0$) for every $v \neq 0$. If g is neither positive-definite nor negative-definite, it is said to be *indefinite*.

Remark 2.3. Many authors include the condition of positive-definiteness in the definition of inner product. We will however relax this hypothesis, following [Nab12] and [SW77].

Definition 2.4. An *inner product space* is a pair (E, g) where E is a real vector space and g is an inner product on E.

Definition 2.5. Let (E_1, g_1) and (E_2, g_2) be two inner product spaces. A linear map $\phi : E_1 \to E_2$ is said to be a *linear isometry* if it is an isomorphism of vector spaces satisfying

$$g_1(v, w) = g_2(\phi(v), \phi(w)), \text{ for all } v, w \in E_1.$$

We then say that ϕ preserves inner products. Let g be a positive-definite (resp. negative-definite) inner product on E and $F \subset E$ a vector subspace. Then, the restriction $g_{|F} : F \times F \to \mathbb{R}$ is a positive-definite (resp. negative-definite) inner product on F. If g is indefinite, however, the restriction $g_{|F}$ may be a positive-definite, negative-definite or indefinite inner product, or it may be a degenerate symmetric bilinear form (and thus not an inner product).

Definition 2.6. The *index* ν of an inner product g on E is the highest dimension of a subspace $F \subset E$ for which $g_{|F}$ is negative-definite.

Example 2.7. The standard euclidean inner product on $\mathbb{R}^n$ defined by

$$g(v, w) := v_1 w_1 + \cdots + v_n w_n$$

for $v = (v_1, \ldots, v_n)$ and $w = (w_1, \ldots, w_n)$ has index 0.

Note that, more generally, any positive-definite (resp. negative-definite) inner product on an n-dimensional vector space has index 0 (resp. n). The converse is also true and thus the notion of index provides an equivalent characterisation for the definiteness of an inner product.

Definition 2.8. Let (E, g) be an inner product space. Two vectors $u, v \in E$ are said to be *orthogonal* if $g(u, v) = 0$. A vector $v \in E$ is called a *unit vector* if $g(v, v) = \pm 1$. A basis $\{e_1, \ldots, e_n\}$ for E consisting of mutually orthogonal unit vectors is called an *orthonormal basis* for E.

The following result (see for example Theorem 1.1.1 in [Nab12]) states that such a basis always exists. This is a semi-Euclidean version of the Gram-Schmidt orthogonalisation process.

Theorem 2.9. *Let (E, g) be an inner product space of dimension n. Then there exists a basis $\{e_1, \ldots, e_n\}$ for E such that $g(e_i, e_j) = 0$ if $i \neq j$ and $g(e_i, e_i) = \pm 1$ for each $i, j = 1, \ldots, n$. Moreover, the number of basis vectors e_i for which $g(e_i, e_i) = -1$ is the same for any such basis.*

The last statement in Theorem 2.9 tells us that the number of vectors e_i in any orthonormal basis for E satisfying $g(e_i, e_i) = -1$ is precisely the index ν. From now on we will assume that all orthonormal bases are indexed in such a way that these e_i appear at the beginning:

$$\{e_1, \ldots, e_\nu, e_{\nu+1}, \ldots, e_n\}$$

where $g(e_i, e_i) = -1$ for $i = 1, \ldots, \nu$ and $g(e_i, e_i) = 1$ for $i = \nu + 1, \ldots, n$. If relative to such basis we have vectors $v = (v^1, \ldots, v^n)$ and $w = (w^1, \ldots, w^n)$ then their inner product will be given by:

$$g(v, w) = -v^1 w^1 - \cdots - v^\nu w^\nu + v^{\nu+1} w^{\nu+1} + \cdots + v^n w^n$$

In the subsequent discussion we will restrict our attention to a particular case of inner product which is of main interest for our purposes.

Definition 2.10. An inner product is called *Lorentzian* if it has index $\nu = 1$. A *Lorentzian vector space* is a real vector space of dimension $n \geq 2$ together with a Lorentzian inner product.

Henceforth, E will denote an n-dimensional Lorentzian vector space with Lorentzian inner product g.

Let $F \subset E$ be a vector subspace of dimension k. The notion of *orthogonal complement* $F^\perp$ for a subspace F in a Lorentzian vector space is the obvious one

$$F^\perp := \{v \in E \mid g(v, w) = 0 \ \forall w \in F\}$$

and it satisfies analogous properties as in Euclidean vector spaces, namely: $F^\perp$ is a subspace of dimension $\dim E - \dim F$, the double orthogonal is itself $F^{\perp\perp} = F$ and there is a direct sum vector space decomposition

$$E = F \oplus F^\perp$$

if and only if the restriction of the inner product of E to F is non-degenerate.

As we anticipated before, the restriction of inner products of arbitrary index to different subspaces may have different properties depending on the subspace one considers. In the Lorentzian case, there are three mutually exclusive options that give rise to the following classification:

Definition 2.11. A subspace $F \subset E$ is said to be:

1. *timelike* if $g_{|F}$ is non-degenerate of index 1,
2. *null* or *lightlike* if $g_{|F}$ is degenerate,
3. *spacelike* if $g_{|F}$ is positive-definite.

The type into which F falls is called its *causal character*, for reasons that will later become clear.

Definition 2.12. We say that a vector $v \in E$ is

1. *timelike* if $g(v, v) < 0$,
2. *null* or *lightlike* if $g(v, v) = 0$ and $v \neq 0$,
3. *spacelike* if $g(v, v) > 0$ or $v = 0$.

Finally, we say that v is *causal* if it is not spacelike.

Again, the type into which v falls is called its *causal character*.

Remark 2.13. Note that both definitions are consistent in the sense that the causal character of a non-zero vector coincides with that of the subspace it spans. Furthermore, a subspace is spacelike if and only if all its vectors are spacelike, lightlike if and only if it contains a lightlike vector but no timelike vector and timelike if and only if it contains a timelike vector. Lastly, note that in any Lorentzian space there is at least one vector (and hence one subspace) of each type.

2.2 Physical interpretation of Minkowski spacetime

We can now formalise the definition of Minkowski spacetime.

Definition 2.14. We define *Minkowski spacetime* $\mathcal{M}$ as a pair $(\mathbb{R}^4, \eta)$, where η is a Lorentzian inner product.

We will also use $\mathcal{M}$ to denote the vector space $\mathbb{R}^4$ itself, the inner product η being implied from now on. The elements of $\mathcal{M}$ will be called *events*. According to Theorem 2.9, there exists an orthonormal basis $\{e_1, e_2, e_3, e_4\}$ on $\mathcal{M}$ for which η has the following matrix representation:

$$\eta(e_i, e_j) = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

As we anticipated, Minkowski spacetime is more than an abstract mathematical entity and cannot be fully understood without the profound physical meaning it possesses. To see how this meaning arises the first step is to establish a correspondence between events, as elements of $\mathcal{M}$, and actual "physical" events. Once a reference frame is fixed, the latter are characterized by the measure t of one time coordinate and the measure (x, y, z) of three spatial coordinates provided by an observer over the reference frame in question. If we multiply the time coordinate by the speed of light in the vacuum c we obtain four coordinates (ct, x, y, z) all having units of distance, which is physically more appealing. Now, assume an event $x \in \mathcal{M}$ has coordinates (x_1, x_2, x_3, x_4) with respect to an orthonormal basis. If we identify $\{e_1, e_2, e_3, e_4\}$ with a certain reference frame $\mathcal{R}$ then we can identify $x = (x_1, x_2, x_3, x_4) \in \mathcal{M}$ with the physical event that in $\mathcal{R}$ is characterised by the four coordinates (x_1, x_2, x_3, x_4). As we will see later, the correspondence between orthonormal basis for $\mathcal{M}$ and reference frames is actually more subtle as it involves the choice of a spatial and a time orientations. For the moment, however, this suffices for our purposes.

The following example may help understand the previous discussion.

Example 2.15. Assume an orthonormal basis, that we identify with a certain reference frame $\mathcal{R}$, is fixed. Consider two events $v_1, v_2 \in \mathcal{M}$ such that $v = v_2 - v_1$ is lightlike (see Definition 2.12). Condition $\eta(v, v) = 0$ implies

$$-(v^1)^2 + (v^2)^2 + (v^3)^2 + (v^4)^2 = -(v_2^1 - v_1^1)^2 + (v_2^2 - v_1^2)^2 + (v_2^3 - v_1^3)^2 + (v_2^4 - v_1^4)^2 = 0 \quad (1)$$

According to the previous considerations, we identify v_1 with event (ct_1, x_1, y_1, z_1) and v_2 with event (ct_2, x_2, y_2, z_2) in $\mathcal{R}$, and (1) now reads

$$\begin{gathered} -c^2(t_2 - t_1)^2 + (x_2 - x_1)^2 + (x_2 - x_1)^2 + (x_2 - x_1)^2 = 0 \\ \Updownarrow \\ c|t_2 - t_1| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \end{gathered} \quad (2)$$

Physically, (2) states that the spatial distance between the physical events represented by v_1 and v_2 coincides with the distance light would travel during the time lapse between them. This means that they can be connected by a light ray, or equivalently that both events can be experienced by the same photon. Whether this light ray is directed from v_1 to v_2 or vice-versa depends on the physical evidence of the existence of an arrow of time, or in other words, that "time only moves forward". Hence, if $t_2 - t_1 > 0$ one can for instance imagine a photon being emitted at v_1 and later received at v_2 whereas if $t_2 - t_1 < 0$ the situation is reversed.

In general, if v_1 and v_2 are two events in $\mathcal{M}$ we will refer to $v = v_2 - v_1 \in \mathcal{M}$ as the *displacement vector from v_1 to v_2*. Moreover, whenever we use the expression *v_1 and v_2 can be connected by v* we will mean that v is either the displacement vector from v_1 to v_2 or the displacement vector from v_2 to v_1. The example above incidentally explains the use of the name "lightlike" for null vectors. Indeed, lightlike vectors in Minkowski spacetime connect events that can be experienced by the same photon.

As we have said, we think of Minkowski spacetime as the collection of all possible events in the universe. In this way, the existence of a particle is represented by the continuous sequence of events that it experiences, what we shall call its *worldline*. To understand what we mean by "continuous" we first need to fix a topology on $\mathcal{M}$. The most natural way to do so, at least from a mathematical point of view, is of course to consider on $\mathcal{M}$ the Euclidean 4-dimensional topology, namely the topology generated by the Euclidean balls

$$B_\epsilon(x) = \{y \in \mathbb{R}^4 |\, d(x,y) < \epsilon\},$$

where

$$d(x,y) = \sqrt{(x^1-y^1)^2 + \cdots + (x^4-y^4)^2}.$$

is the usual Euclidean distance. As will be discussed in Section 5, one may define different topologies for $\mathcal{M}$, but until then we will assume that it has the Euclidean topology. A simple example of worldline is that of a photon. In fact, we already referred to it in Example 2.15, but now we can provide a formal definition.

Definition 2.16. A *light ray on* $\mathcal{M}$ is a subset λ of $\mathcal{M}$ defined by

$$\lambda = \{x_0 + t(x - x_0)|\, t \in \mathbb{R}\},$$

for any $x, x_0 \in \mathcal{M}$ such that $x - x_0$ is lightlike.

More generally, wordlines will be described by curves on $\mathcal{M}$ satisfying certain conditions. Let I denote a real interval.

Definition 2.17. A *curve on* $\mathcal{M}$ is a continuous map $\alpha : I \to \mathcal{M}$.

Relative to any orthonormal basis $\{e_i\}_{i=1,\ldots,n}$ for $\mathcal{M}$ we can write $\alpha(t) = \sum_{i=1}^{4} x^i(t)e_i$ for each $t \in I$. We will assume that α is smooth, i.e., that each component function is infinitely differentiable. Its *velocity vector* will be given by

$$\alpha'(t) = \sum_{i=1}^{4} \frac{dx^i}{dt} e_i.$$

Definition 2.18. A curve $\alpha : I \to M$ is said to be *timelike, null* or *spacelike* if its velocity vector $\alpha'(t)$ has that causal character for all $t \in I$.

According to our previous discussion the last three components of α correspond to the tree spatial coordinates of the physical event it represents. If we think of this event as being experienced by some particle, then the last three components of α' correspond to the three spatial components of the particle's instantaneous velocity v. The same reasonining used in Example 2.15 shows that

$$\alpha'(t_0) \text{ timelike } \iff v(t_0) < c,$$
$$\alpha'(t_0) \text{ spacelike } \iff v(t_0) > c.$$

Now, as it is well known in the domains of Physics, the postulates of SR inevitably lead to the conclusion that it is impossible for information or energy (and hence matter) to travel faster than light. It follows that a curve $\alpha : I \to M$ such that $\alpha'(t_0)$ is spacelike for some $t_0 \in I$

is non-admissible physically, in the sense that it cannot represent the worldline of a physical particle. Therefore, only timelike and null curves describe worldlines. Timelike curves represent the wordline of material particles (that move at a velocity strictly smaller than c) and may have arbitrarily complicated shapes provided that the timelikeness condition is everywhere satisfied. A particular case is that of *free moving* material particles, the wordlines of which are timelike lines.

Definition 2.19. A *timelike line* is a subset τ of $\mathcal{M}$ defined by

$$\tau = \{x_0 + t(x - x_0) |\, t \in \mathbb{R}\},$$

for any $x, x_0 \in \mathcal{M}$ such that $x - x_0$ is timelike.

Null curves, on the other hand, represent the wordlines of photons (massless particles). In this case, the condition that $v(t) = c$ force any null curve to take the form of a light ray, in the sense of Definition 2.16.

To end this section, let us point out that from now on we shall make use of Einstein's summation convention, according to which a repeated index, one subscript and one superscript, indicates a sum over the range of values the index can assume. For example, if $i = 1, \dots, 4$, then $x^i e_i = \sum_{i=1}^{4} x^i e_i$.

2.3 Causal structure

In the following discussion we aim to describe how the causal structure of Minkowski spacetime arises from the results concerning Lorentzian vector spaces that were introduced in Section 2.1. Therefore, we will focus on the 4-dimensional case but we want to stress that the following definitions and results can be naturally generalised to the case of arbitrary dimension $n \geq 2$, since no particular use of the dimension will be made. Only in some particular cases will the generalisation require some brief comment, that we shall do at due time.

Definition 2.20. Let $x_0 \in \mathcal{M}$. We define the *null cone* (or *light cone) $\mathcal{C}_N(x_0)$ at x_0* as

$$\mathcal{C}_N(x_0) := \{x \in \mathcal{M} |\, \eta(x - x_0, x - x_0) = 0\}.$$

Note that, relative to any orthonormal basis $\{e_i\}, (i = 1, \dots, 4)$, if $x_0 = x_0^i e_i$, then events $x = x^i e_i$ laying in the light cone at x_0 satisfy the equation

$$-(x^1 - x_0^1)^2 + (x^2 - x_0^2)^2 + (x^3 - x_0^3)^2 + (x^4 - x_0^4)^2 = 0, \tag{3}$$

completely analogous to (1). In fact, as follows from Example 2.15, $\mathcal{C}_L(x_0)$ corresponds to the set of all physical events that can be connected to x_0 by a light ray. This, together with the fact that (3) can be thought as the equation of a cone in $\mathbb{R}^4$, explains the name light cone.

This geometrical interpretation of $\mathcal{C}_L(x_0)$ is of utmost importance for the mathematical description of causality. For example, one can see that an event $x \in \mathcal{M}$ lays inside $\mathcal{C}_L(x_0)$ if and only if its coordinates satisfy the equation

$$-(x^1 - x_0^1)^2 + (x^2 - x_0^2)^2 + (x^3 - x_0^3)^2 + (x^4 - x_0^4)^2 < 0, \tag{4}$$

namely, if $x - x_0$ is timelike. This motivates the following definition.

Definition 2.21. We define the *time cone $\mathcal{C}_T(x_0)$ at x_0* as

$$\mathcal{C}_T(x_0) := \{x \in \mathcal{M} |\, \eta(x - x_0, x - x_0) < 0\}.$$

Remark 2.22. It follows from (3) and (4) that $\mathcal{C}_N(x_0)$ and $\mathcal{C}_T(x_0)$ are, respectively, closed and open in $\mathcal{M}$ with the Euclidean topology.

Similarly, an event $x \in \mathcal{M}$ lays outside $\mathcal{C}_L(x_0)$ if and only if its coordinates satisfy the equation

$$-(x^1 - x_0^1)^2 + (x^2 - x_0^2)^2 + (x^3 - x_0^3)^2 + (x^4 - x_0^4)^2 > 0, \tag{5}$$

namely, if $x - x_0$ is spacelike. In general, the set of events that satisfy (5) is not explicitly defined. Some authors, however, call it the *space cone at* x_0 and denote it by $\mathcal{C}_S(x_0)$, in analogy to Definitions 2.20 and 2.21 (see for example [Zee66]). The physical interpretation of the time cone $\mathcal{C}_T(x_0)$ is that x_0 and any event $x \in \mathcal{C}_T(x_0)$ can be connected by a timelike line or, equivalently, that x_0 and $x \in \mathcal{C}_T(x_0)$ can be experienced by the same free moving material particle. Furthermore, any timelike curve α passing through x_0 must lie entirely in $\mathcal{C}_T(x_0)$ for in order to leave the time cone its velocity vector should be non-timelike at some point. Thus, $C_T(x_0)$ is the set of all points that can be joined by the worldline of some material particle experiencing x_0. Since no information or energy can travel faster than light, all physically admissible wordlines passing through x_0 are contained in $\mathcal{C}_L(x_0) \cup \mathcal{C}_T(x_0)$. Since no cause-effect relationship may be established between two events without some kind of information exchange taking place between them, we can affirm that for a given event $x_0 \in \mathcal{M}$ the set $\mathcal{C}_L(x_0) \cup \mathcal{C}_T(x_0)$ consists of all the events that can be causally related to x_0. On the other hand, all events outside $\mathcal{C}_L(x_0) \cup \mathcal{C}_T(x_0)$ are causally disconnected from x_0.

In order to complete the description of the causal structure of Minkowski spacetime, however, we need to know in which direction the causal relationship between two events can be established. That is, for every $x_0 \in \mathcal{M}$ we need to distinguish between the events that can *causally affect* x_0 and those that can be *causally affected by* x_0. Physically, this question is solved by the existence of an arrow of time and the human notions of future and past. Our next goal is then to study how these notions may arise from the mathematical structure of $\mathcal{M}$.

Consider the set $\mathcal{T}$ of all timelike vectors in $\mathcal{M}$, which is open in $\mathcal{M}$ by an argument analogous to that of Remark 2.22. Now, define on $\mathcal{T}$ the following relation:

$$v \sim w \quad \Longleftrightarrow \quad \eta(v, w) < 0$$

We want to show that $\sim$ is an equivalence relation. For this purpose, let us first prove the following result.

Lemma 2.23. *Suppose that v is timelike and $w \neq 0$ is either timelike or lightlike. Let $\{e_i\}$ be an orthonormal basis for $\mathcal{M}$ with $v = v^i e_i$ and $w = w^i e_i$. Then either*

(i) $v^1 w^1 > 0$ and $\eta(v, w) < 0$, or

(ii) $v^1 w^1 < 0$ and $\eta(v, w) > 0$.

Proof. By assumption we have

$$\eta(v, v) = -(v^1)^2 + (v^2)^2 + (v^3)^2 + (v^4)^2 < 0 \text{ and } \eta(w, w) = -(w^1)^2 + (w^2)^2 + (w^3)^2 + (w^4)^2 \leq 0.$$

Therefore it follows that

$$(v^1 w^1)^2 > \left((v^1)^2 + (v^2)^2 + (v^3)^2\right)\left((w^1)^2 + (w^2)^2 + (w^3)^2\right) \geq (v^1 w^1 + v^2 w^2 + v^3 w^3)^2,$$

where the last inequality follows from Schwartz's inequality in $\mathbb{R}^3$. We then have that

$$|v^1 w^1| > |v^2 w^2 + v^3 w^3 + v^4 w^4|,$$

which in particular implies that $\eta(v, w) \neq 0$. Now suppose that $v^1 w^1 > 0$, then

$$v^1 w^1 = |v^1 w^1| > |v^2 w^2 + v^3 w^3 + v^4 w^4| \geq v^2 w^2 + v^3 w^3 + v^4 w^4.$$

Therefore we obtain

$$-v^1 w^1 + v^2 w^2 + v^3 w^3 + v^4 w^4 = \eta(v, w) < 0.$$

The case $v^1 w^1 < 0$ follows analogously. $\square$

The next immediate corollary of the previous result will be useful in a few lines.

Corollary 2.24. *If a nonzero vector in E is orthogonal to a timelike vector, then it must be spacelike.*

Proposition 2.25. *The relation $\sim$ is an equivalence relation with precisely two equivalence classes.*

Proof. Reflexivity and symmetry of $\sim$ follow directly from the definition of timelike vector and the symmetry of η, respectively. For transitivity, consider $v, w, u \in \mathcal{M}$ and assume $v \sim w$ and $w \sim u$, i.e., $\eta(v, w) < 0$ and $\eta(w, u) < 0$. By Lemma 2.23 we then have $v^1 w^1 > 0$ and $w^1 u^1 > 0$. Hence $v^1 (w^1)^2 u^1 > 0 \Rightarrow v^1 u^1 > 0$, which by Lemma 2.23 means $\eta(v, u) < 0 \iff v \sim u$. Finally, for a given $w \in \mathcal{T}$, again by Lemma 2.23 either $\eta(v, w) < 0$ and so w is in the equivalence class $[\,v\,]$ of v or $\eta(v, w) > 0 \Rightarrow \eta(-v, w) < 0$ and so w is in the equivalence class $[-v]$ of $-v$. $\square$

Corollary 2.26. *The set $\mathcal{T}$ has two connected components.*

Proof. Let $v \in \mathcal{T}$, then by Lemma 2.23 $[\,v\,] = \{w \in \mathcal{T} |\, v^1 w^1 < 0\}$ and $[-v] = \{w \in \mathcal{T} |\, v^1 w^1 > 0\}$. It follows that $[\,v\,]$ and $[-v]$ are open in $\mathcal{T}$ and since $\mathcal{T} = [\,v\,] \sqcup [-v]$, as shown in the previous proof, we have that $[\,v\,]$ and $[-v]$ are the two connected components of $\mathcal{T}$. $\square$

The last two results, yet simple, are essential in our description of causality, as they allow us to give a mathematical definition to the physical concepts of future and past.

Definition 2.27. A *time orientation for $\mathcal{M}$* is an (arbitrary) labelling of the two components of $\mathcal{T}$ as $\mathcal{T}^+$ (called the *future*) and $\mathcal{T}^-$ (called the *past*). We will refer to elements in $\mathcal{T}^+$ (resp. $\mathcal{T}^-$) as *future-directed* (resp. *past-directed*) timelike vectors.

Remark 2.28. It follows from the properties of an inner product that $\mathcal{T}^+$ and $\mathcal{T}^-$ are cones, in the sense that if v and w are elements of $\mathcal{T}^+$ (resp. $\mathcal{T}^-$) and λ is a positive real number, then λv and $v + w$ are also in $\mathcal{T}^+$ (resp. $\mathcal{T}^-$).

As we suggested at the end of Section 2.1, the identification of an orthonormal basis with a reference frame is not that immediate. In order for the latter to be "physically admissible" we must first fix a time and a spatial orientation for the basis.

Definition 2.29. We say that an orthonormal basis $\{e_1, e_2, e_3, e_4\}$ is an *admissible basis* for $\mathcal{M}$ if e_1 is future-directed and timelike and $\{e_2, e_3, e_4\}$ is spacelike and "right-handed", i.e., satisfying $e_2 \times e_3 \cdot e_4 = 1$.

Note that since the restriction of η to $\langle e_2, e_3, e_4 \rangle$ is the usual Euclidean inner product on $\mathbb{R}^3$, the cross product and inner product here are the familiar ones from vector calculus.

The distinction between a future and a past direction for timelike vectors leads to the following definition.

Definition 2.30. For each $x_0 \in \mathcal{M}$, we define the *future time cone $\mathcal{C}_T^+(x_0)$* and the *past time cone $\mathcal{C}_T^-(x_0)$ at x_0* as

$$\mathcal{C}_T^+(x_0) := \{x \in \mathcal{M} |\; x - x_0 \in \mathcal{T}^+\} = \mathcal{C}_T(x_0) \cap \mathcal{T}^+,$$

and

$$\mathcal{C}_T^-(x_0) := \{x \in \mathcal{M} |\; x - x_0 \in \mathcal{T}^-\} = \mathcal{C}_T(x_0) \cap \mathcal{T}^-.$$

The following figure helps illustrate these ideas:

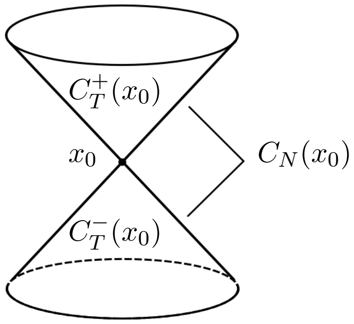

Figure 1: The null cone and the future and past time cones at x_0.

In order to complete the formal description of causality in $\mathcal{M}$ we need to extend the notion of future-directed and past-directed to null vectors as well. Consider the set $\mathcal{N}$ of null vectors of $\mathcal{M}$, then the following result holds.

Lemma 2.31. *For every $w \in \mathcal{N}$, $\eta(w, v)$ has the same sign for all $v \in \mathcal{T}^+$.*

Proof. Suppose that there exist $v_1, v_2 \in \mathcal{T}^+$ such that $\eta(w, v_1) < 0$ and $\eta(w, v_2) > 0$. We may assume $|\eta(w, v_1)| = \eta(w, v_2)$ since this is not the case we can set $\lambda = \eta(w, v_2)/|\eta(w, v_1)|$ and replace v_1 by λv_1, which is still in $\mathcal{T}^+$ by Remark 2.28 and satisfies $\eta(w, \lambda v_1) = \lambda \eta(w, v_1) = -\eta(w, v_2)$. Thus, $\eta(w, v_1) = -\eta(w, v_2)$ and therefore $\eta(w, v_1 + v_2) = 0$. Again by Remark 2.28, $v_1 + v_2 \in \mathcal{T}^+$ and so, in particular, is timelike. Since w is null (and hence nonzero), this contradicts Corollary 2.24. $\square$

Therefore, we have shown that for every $w \in \mathcal{N}$, either $\eta(w, v) < 0$ or $\eta(w, v) > 0$ for every $v \in \mathcal{T}^+$. Equivalently, either $\eta(w, v) < 0$ for every $v \in \mathcal{T}^+$ or $\eta(w, v) < 0$ for every $v \in \mathcal{T}^-$. We can now define the sets $\mathcal{N}^+ := \{w \in \mathcal{N} |\, \eta(w, v) < 0,\ \forall\, v \in \mathcal{T}^+\}$ and $\mathcal{N}^- := \{w \in \mathcal{N} |\, \eta(w, v) < 0,\ \forall\, v \in \mathcal{T}^-\}$, which are open in $\mathcal{N}$. All together, we have proved a result analogous to Corollary 2.26 in the case of null vectors.

Corollary 2.32. *The set $\mathcal{N}$ has two connected components.*

Proof. By Lemma 2.31, if $w \in \mathcal{N}$, either $\eta(w, v) < 0$ or $\eta(w, v) > 0$ for every $v \in \mathcal{T}^+$. Equivalently, either $\eta(w, v) < 0$ for every $v \in \mathcal{T}^+$ or $\eta(w, v) < 0$ for every $v \in \mathcal{T}^-$. If we now define

$$\mathcal{N}^+ := \{w \in \mathcal{N} |\, \eta(w, v) < 0,\ \forall\, v \in \mathcal{T}^+\} \quad \text{and} \quad \mathcal{N}^- := \{w \in \mathcal{N} |\, \eta(w, v) < 0,\ \forall\, v \in \mathcal{T}^-\},$$

we have that $\mathcal{N} = \mathcal{N}^+ \sqcup \mathcal{N}^-$ and the result follows from the fact that $\mathcal{N}^+$ and $\mathcal{N}^-$ are open in $\mathcal{M}$. $\square$

Remark 2.33. The last result does not hold for $n = 2$, in which case the set $\mathcal{N}$ splits into 4 connected components. However, they can be grouped in pairs: $\mathcal{N}^+ = \mathcal{N}_1^+ \cup \mathcal{N}_2^+$ and $\mathcal{N}^- = \mathcal{N}_1^- \cup \mathcal{N}_2^-$ in such a way that the subsequent discussion is essentially the same.

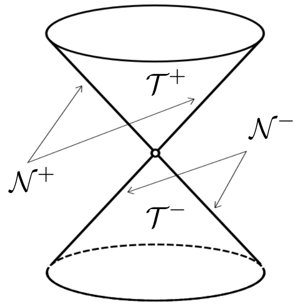

Figure 2: The connected components of $\mathcal{T}$ and $\mathcal{N}$.

Remark 2.34. It can be shown that each of the two components of $\mathcal{T}$ is homeomorphic to $\mathbb{R}^4$ and each of the two components of $\mathcal{N}$ is homeomorphic to $\mathbb{R} \times S^2$ (see Figure 2).

Definition 2.35. A null vector w is *future-directed* if $w \in \mathcal{N}^+$ and *past-directed* if $w \in \mathcal{N}^-$.

Definition 2.36. For any $x_0 \in \mathcal{M}$ we define the *future null* (or *light*) *cone* $\mathcal{C}_N^+(x_0)$ and the *past null* (or *light*) *cone* $\mathcal{C}_N^-(x_0)$ *at* x_0 as

$$\mathcal{C}_N^+(x_0) := \{x \in \mathcal{M} |\ x - x_0 \text{ is future-directed}\},$$

and

$$\mathcal{C}_N^-(x_0) := \{x \in \mathcal{M} |\ x - x_0 \text{ is past-directed}\}.$$

We can now interpret the previous discussion from a physical point of view. First, if we trust our experience and rule out the possibility to move back in time, only future-directed timelike or null curves may describe the wordline of a particle. Then, $\mathcal{C}_T^+(x_0)$ consists of all events that may be experienced by some material particle that has already experienced x_0 in the past while $\mathcal{C}_T^+(x_0)$ consists of all events that may have been experienced in the past by some material particle experiencing x_0. The same applies to $\mathcal{C}_N^+(x_0)$ and $\mathcal{C}_N^-(x_0)$ but for photons instead of material particles. For example, for every $x \in \mathcal{C}_N^+(x_0)$, x_0 and x can be regarded as the emission and the reception of a photon, respectively. Thus, $\mathcal{C}_N^+(x_0)$ may be thought of as the history in spacetime of a spherical electromagnetic wave emitted at x_0.

We can finally complete our description of causality by asserting that every $x_0 \in \mathcal{M}$ can only be *causally affected by* events $x \in \mathcal{C}_N^-(x_0) \cup \mathcal{C}_T^-(x_0)$ and can only *causally affect* events $x \in \mathcal{C}_N^+(x_0) \cup \mathcal{C}_T^+(x_0)$.

The results of this section will be essential in Section 4, where we will study how the causal structure arises for spacetimes other than Minkowski's.

3 Semi-Riemannian geometry

In order to describe arbitrary spacetimes where gravity is present a much more complex geometrical structure than that provided by Lorentzian vector spaces is required, namely that of a Lorentzian manifold. The goal of this section is to offer a description of the geometry of spacetimes by reviewing some standard topics on differential and semi-Riemannian geometry. Special attention will be paid to those aspects that are relevant for the subsequent description of causality in spacetimes, such as vector fields, geodesics or the exponential map.

Differential geometry deals with the study of smooth manifolds. These are, roughly speaking, mathematical objects that behave locally as Euclidean spaces, in such a way that one can apply on them the techniques of differential calculus. This allows to generalise the standard study of surfaces in $\mathbb{R}^3$ to an arbitrary dimension by defining the notion of tangent vectors and tangent space without requiring an ambient Euclidean space. Once this has been settled, semi-Riemannian geometry allows to generalise the notion of inner product from vector spaces to smooth manifolds by the introduction of the metric tensor.

For this section, we have mainly followed [O'N83].

3.1 Smooth manifolds

Let M be a topological space.

Definition 3.1. An *n-dimensional local chart* on M is a pair (U, φ) where U is an open subset in M and $\varphi : U \to \varphi(U) \subset \mathbb{R}^n$ is a homeomorphism. The functions $x^i = u^i \circ \varphi : U \to \mathbb{R}$ $(1 \leq i \leq n)$ where $u^i : \mathbb{R}^n \to \mathbb{R}$ denote the canonical coordinate functions $u^i(x^1, \ldots, x^n) = x^i$, are called the *coordinate functions of* φ. In this case, we will write $\varphi = (x^1, \ldots, x^n)$. The functions x^i determine a *local coordinate system* $\{U; x^1, \ldots, x^n\}$ *in* U for which a point $p \in U$ is said to have *coordinates* $(x^1(p), \ldots, x^n(p))$.

Remark 3.2. If (U, φ) and (V, ψ) are two local charts on M such that $U \cap V \neq \emptyset$, the following diagram:

$$\begin{tikzcd} U \cap V \arrow[r, "\varphi"] \arrow[d, "\psi"] & \varphi(U \cap V) \arrow[dl, "\psi\circ\varphi^{-1}"] \\ \psi(U \cap V) & \end{tikzcd}$$

is commutative. Therefore the transition map $\psi \circ \varphi^{-1} : \varphi(U \cap V) \to \psi(U \cap V)$ is a homeomorphism as well.

Definition 3.3. An *atlas* on M is a set $\mathcal{A} = \{(U_i, \varphi_i)\}_{i \in I}$ of local charts on M such that $M = \bigcup\limits_{i \in I} U_i$. An *n-dimensional topological manifold* is a Hausdorff, second-countable topological space M for which there is a family $\{(U_i, \varphi_i)\}_{i \in I}$ of n-dimensional local charts such that $M = \bigcup\limits_{i \in I} U_i$.

Definition 3.4. An atlas $\mathcal{A} = \{(U_i, \varphi_i)\}_{i \in I}$ on M is *smooth* if the transition maps

$$\varphi_{ij} := \varphi_j \circ \varphi_i^{-1} : \varphi_i(U_i \cap U_j) \to \varphi_j(U_i \cap U_j)$$

are of class $\mathcal{C}^\infty(\mathbb{R}^{dimM}, \mathbb{R}^{dimM})$ for all $i, j \in I$ such that $U_i \cap U_j \neq \emptyset$.

Definition 3.5. Let $\mathcal{A} = \{(U_i, \varphi_i)\}_{i \in I}$ be a smooth atlas on M. A local chart (V, ψ) on M is said to be *compatible* with $\mathcal{A}$ if and only if $\mathcal{A} \cup \{(V, \psi)\}$ is a smooth atlas. Two smooth atlases $\mathcal{A}$ and $\mathcal{A}'$ on M are *compatible* if and only if every local chart on $\mathcal{A}'$ is compatible with $\mathcal{A}$, and vice-versa.

Compatibility of smooth atlases is an equivalence relation and equivalence classes of smooth atlases on M are called *smooth structures*.

Definition 3.6. An *n-dimensional smooth manifold* is a pair $(M, [\mathcal{A}])$, where M is an n-dimensional topological manifold and $[\mathcal{A}]$ is a smooth structure on M.

In order to simplify the notation we will use $\mathcal{A}$ instead of $[\mathcal{A}]$ whenever it is understood from the context which is the one we are referring to. We may also refer to a smooth manifold $(M, [\mathcal{A}])$ simply by M, the smooth structure thus being implied although not specified. If this is the case, when considering two different local charts on the same smooth manifold, it will go without saying that they belong to the same smooth structure. From now on, whenever we say "manifold" we will mean "smooth manifold".

Examples 3.7. For every integer $n \geq 1$:

1. $\mathbb{R}^n$ is a smooth manifold with atlas $\{(\mathbb{R}^n, id_{\mathbb{R}^n})\}$.
2. Any open set $U \subset \mathbb{R}^n$ is a smooth manifold with atlas $\{(U, id_U)\}$.
3. Any n-dimensional real vector space V is a smooth manifold with atlas $\{(V, \phi)\}$, where $\phi : V \to \mathbb{R}^n$ is an isomorphism.
4. Let $S^n = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid x_1^2 + \cdots + x_{n+1}^2 = 1\}$ denote the n-dimensional sphere. Consider its open subsets $U_N = S^n \setminus \{(0,0,1)\}$ and $U_S = S^n \setminus \{(0,0,-1)\}$, and the stereographic projections $\varphi_N : U_N \to \mathbb{R}^n$ and $\varphi_S : U_S \to \mathbb{R}^n$. Then, S^n is a smooth manifold with atlas $\{(U_N, \varphi_N), (U_S, \varphi_S)\}$.

In what follows, unless otherwise specified, M will be an n-dimensional manifold.

Definition 3.8. Let $(M, \mathcal{A})$ and $(N, \mathcal{A}')$ be two smooth manifolds of dimension m and n, respectively. A map $f : M \to N$ is said to be *smooth at $p \in M$* if for every $(U, \varphi) \in \mathcal{A}$ such that $p \in U$ and for every $(V, \psi) \in \mathcal{A}'$ such that $f(p) \in V$, the map

$$\psi \circ f \circ \varphi^{-1} : \varphi(U \cap f^{-1}(V)) \longrightarrow \psi(V)$$

is of class $\mathcal{C}^\infty(\mathbb{R}^n, \mathbb{R}^m)$. Then, f is said to be *smooth* if it is smooth for all $p \in M$.

Two of the most frequent particular cases of smooth maps are curves and functions on a manifold.

Definition 3.9. Let I be an open interval in $\mathbb{R}$. A *curve on M* is a smooth map $\gamma : I \to M$.

For an interval $J \subset \mathbb{R}$ not necessarily open, one can still define a curve $\alpha : J \to M$ on M by requiring that there exists some open interval I and some curve $\gamma : I \to M$ such that $J \subset I$ and $\alpha = \gamma|_J$. This is made so that differentiability makes sense at the endpoints.

Definition 3.10. A *function on M* is a smooth map $f : M \to \mathbb{R}$. We denote by $\mathcal{F}(M)$ the set of all functions on M.

The set $\mathcal{F}(M)$ has the structure of a real vector space with the point-wise operations:

$$(f+g)(p) := f(p) + g(p) \ ; \ (\lambda f)(p) := \lambda \cdot f(p),$$

as well as a ring structure, with the multiplication $(fg)(p) := f(p) \cdot g(p)$, for all $p \in M$.

Definition 3.11. A smooth map f : $M \to N$ is said to be a *diffeomorphism* if it is bijective and its inverse f^{-1} is also smooth.

Remark 3.12. Not every homeomorphism is a diffeomorphism, even if it is smooth. For instance, $f : \mathbb{R} \to \mathbb{R}$ defined by $t \mapsto t^3$ is smooth. Its inverse is continuous but not smooth.

Our next goal is to study how some subsets of a smooth manifold M, called submanifolds, inherit its smooth structure in a natural way and become smooth manifolds on their own. To do that, we shall give a definition of submanifold and then prove that it fulfills the required conditions, namely that submanifolds thus defined are indeed smooth manifolds and that their smooth structure is obtained from the restriction of that in M.

Recall that if X is a topological space, then its topology $\mathcal{T}$ naturally induces on any subset $A \subset X$ a topology $\mathcal{T}_A$, called the *subspace topology*, by letting $\mathcal{T}_A = \{U \cap A |\ U \in \mathcal{T}\}$. In this case A is said to be a *topological subspace of* X and is in particular a topological space on its own.

Definition 3.13. A subset $S \subset M$ is a *k-dimensional smooth submanifold of* M if for every $p \in S$ there is a chart (U, φ) of M around p such that

$$\varphi(U \cap S) = \varphi(U) \cap (\mathbb{R}^k \times \{0\}) = \{x \in \varphi(U) |\, x^{k+1} = \cdots = x^n = 0\}.$$

Examples 3.14.
1. Any open subset $U \subset M$ is a smooth submanifold with the same dimension as M.
2. Any k-dimensional subspace F of a real vector space E is a k-dimensional smooth submanifold of E.

A k-dimensional submanifold S of M can be given a smooth structure as follows. First, assume the subspace topology on S. This implies that S is second-countable and Hausdorff, since these properties are inherited by subspaces. Now, consider the maps

$$\begin{aligned} \pi : \mathbb{R}^n \to \mathbb{R}^k, &\quad (x^1, \ldots, x^n) \mapsto (x^1, \ldots, x^k), \\ j : \mathbb{R}^k \hookrightarrow \mathbb{R}^n, &\quad (x^1, \ldots, x^k) \mapsto (x^1, \ldots, x^k, 0, \ldots, 0). \end{aligned}$$

Let (U, φ) be a local chart of M at $p \in S$ as in Definition 3.13. Then, $\psi = \pi \circ \varphi|_{U \cap S}$ has inverse $\psi^{-1} = \varphi^{-1} \circ j$ and defines a k-dimensional chart $(U \cap S, \psi)$ on S at $p \in S$. Again by the definition of smooth submanifold we have that S can be covered by such charts. Hence S is a k-dimensional topological manifold. Finally, all such charts are compatible since the transition maps satisfy

$$\psi_{ik} = \psi_k \circ \psi_i^{-1} = \pi \circ \varphi_k \circ \varphi_i^{-1} \circ j = \pi \circ \varphi_{ik} \circ j,$$

and therefore are smooth.

3.2 Tangent vector space

One can attach to every point of a smooth manifold a tangent space. The latter is an n-dimensional real vector space that intuitively contains all the possible directions in which one can tangentially pass through p. This definition relies on a manifold's ability to be embedded into an ambient vector space. However, it is more convenient to define the notion of a tangent space depending only on the manifold. Let us first begin by introducing the notion of vector tangent to M at $p \in M$ and then the tangent space will be defined naturally as the set of all such vectors.

Definition 3.15. Let M be a smooth manifold and $p \in M$. A *vector tangent to M at p* is a map $v : \mathcal{F}(M) \to \mathbb{R}$ such that

(i) $v(af + bg) = av(f) + bv(g)$,

(ii) $v(fg) = v(f)g(p) + f(p)v(g)$,

for all $f, g \in \mathcal{F}(M)$ and $a, b \in \mathbb{R}$.

Definition 3.16. We define the *tangent space* T_pM *of* M *at* p as the set of all vectors tangent to M at p.

The tangent space T_pM is a real vector space with the point-wise addition and multiplication by a scalar, namely,

$$(v + w)(f) := v(f) + w(f) \quad \text{and} \quad (\lambda v)(f) := v(\lambda f)$$

for all $v, w \in T_pM$ and $\lambda \in \mathbb{R}$.

Remark 3.17. Any smooth curve $\gamma : (-\epsilon, \epsilon) \to M$ with $\gamma(0) = p$ defines a tangent vector v_γ at p by letting

$$v_\gamma(f) = \left(\frac{d(f \circ \gamma)}{dt}\right)(0).$$

In fact, one may alternatively define vectors tangent to M at p as equivalence classes (see for example [Nak90]), denoted by $\dot{\gamma}(0)$, of smooth curves $\gamma : (-\epsilon, \epsilon) \to M$ with $\gamma(0) = p$, where γ_1 and γ_2 are said to be equivalent if the derivatives of $\varphi \circ \gamma_1$ and $\varphi \circ \gamma_2$ at 0 coincide for some chart (U, φ) with $p \in U$. The tangent space is then defined as the set of all such equivalence classes.

Tangent vectors are to be regarded as local objects, as they satisfy the localisation principle: if two smooth functions f and g coincide in a neighborhood of a point $p \in M$, then their images coincide for all $v \in T_pM$.

Our next step is to provide an adequate coordinate description for the vector space T_pM. Let (U, φ) be a chart on M with coordinate functions $x^1, \ldots, x^n$. For each $i \in \{1, \ldots, n\}$ consider the function $\frac{\partial}{\partial x^i}\big|_p : \mathcal{F}(M) \longrightarrow \mathbb{R}$ defined by

$$\frac{\partial}{\partial x^i}\bigg|_p(f) = \frac{\partial(f \circ \varphi^{-1})}{\partial u^i}\left(\varphi(p)\right),$$

where $u^1, \ldots, u^n$ are the natural coordinate functions of $\mathbb{R}^n$. It is easy to see that $\frac{\partial}{\partial x^i}\big|_p$ is a vector tangent to M at p, in the sense of Definition 3.15.

Definition 3.18. The vector $\frac{\partial}{\partial x^i}\big|_p \in T_pM$ is called the *vector tangent to* M *at* $p \in U$ *in the* x^i *coordinate direction.*

Whenever there is no confusion with respect to which chart is being considered, we will denote the vector $\frac{\partial}{\partial x^i}\big|_p$ simply by $\partial_i|_p$. The following result establishes a fundamental link between coordinates and tangent vectors (see for instance Theorem 1.12 in [O'N83] for a proof).

Theorem 3.19. *Let* (U, φ) *be a chart on* M *with coordinate functions* x^i *for* $i \in \{1, \ldots, n\}$. *Then,* $\{\partial_i|_p\}_{i=1,\ldots,n}$ *is a basis of* T_pM *in terms of which every* $v \in T_pM$ *can be written as*

$$v = v(x^i)\partial_i|_p.$$

Corollary 3.20. *The vector space* T_pM *has the same dimension as* M.

The numbers $v(x^i)$ are then the coordinates of $v \in T_pM$ in the basis $\{\partial_i|_p\}_{i=1,\ldots,n}$. We will denote them by v^i.

Remark 3.21. In the particular case where M is a real vector space with a certain orthonormal basis $\{e_i\}_{i=1,\ldots,n}$, there is a natural linear isomorphism sending every $v_p = v^i\partial_i|_p \in T_pM$ to $v = v^ie_i \in M$.

It follows from Corollary 3.20 that if $S \subset M$ is a k-dimensional submanifold of M, then for every $p \in S$ the tangent space T_pS is a k-dimensional real vector space that can be regarded as a subspace of the n-dimensional tangent space T_pM.

Definition 3.22. Let $\phi : M \to N$ be a smooth map between manifolds. For each $p \in M$ we define the *differential map of* ϕ *at* p by

$$\begin{aligned} d\phi_p : T_pM &\to T_{\phi(p)}N \\ v &\mapsto d\phi_p(v) = v_\phi, \end{aligned}$$

where v_ϕ is given by the rule $v_\phi(g) = v(g \circ \phi)$ for every $g \in \mathcal{F}(N)$.

One can easily check that v_ϕ is indeed a vector tangent to N at $\phi(p)$ in the sense of Definition 3.15. It follows also that $d\phi_p$ is a linear map between vector spaces.

Remark 3.23. The differential map has perhaps a more intuitive description when considering tangent vectors as equivalence classes. In this case, it can be defined by

$$\begin{aligned} d\phi_p :\ T_pM &\longrightarrow T_{\phi(p)}N \\ \dot\gamma(0) &\mapsto d\phi_p(\dot\gamma(0)) = (\phi \dot\circ \gamma)(0), \end{aligned}$$

which of course does not depend on the representative γ one chooses for $\dot\gamma(0)$.

Proposition 3.24. *Let $\phi : M \to N$ be a smooth map. If (U, φ) is a chart on M around some $p \in M$ with coordinate functions $x^1, \ldots, x^n$ and (V, ψ) is a chart on N around $\phi(p) \in N$ with coordinate functions $y^1, \ldots, y^m$, then:*

$$d\phi_p\left(\left.\frac{\partial}{\partial x^i}\right|_p\right) = \sum_{j=1}^m \frac{\partial(y^j \circ \phi)}{\partial x^i}(p)\left.\frac{\partial}{\partial y^j}\right|_{\phi(p)}, \qquad (i = 1, \ldots, n).$$

Proof. Let $w \in T_{\phi(p)}N$ be the left hand side of the previous equality. By Theorem 3.19 we may write

$$w = \sum_{j=1}^m w(y^j)\frac{\partial}{\partial y^j}|_{\phi(p)}.$$

But by the definition of differential map:

$$w(y^j) = d\phi_p\left(\left.\frac{\partial}{\partial x^i}\right|_p\right)(y^j) = \frac{\partial(y^j \circ \phi)}{\partial x^i}(p). \qquad \square$$

In view of the above result, the matrix of $d\phi_p$ relative to the basis $\left\{\frac{\partial}{\partial x^i}|_p\right\}_{i=1,\ldots,n}$ and $\left\{\frac{\partial}{\partial y^j}|_{\phi(p)}\right\}_{j=1,\ldots,m}$ of T_pM and $T_{\phi(p)}N$, respectively, is

$$\left(\frac{\partial(y^j \circ \phi)}{\partial x^i}(p)\right)_{1\le i\le n, 1\le j\le m},$$

called the *Jacobian matrix of ϕ at p relative to φ and ψ.*

Example 3.25. Any linear map $\phi : E \to F$ between real vector spaces is a smooth map. By the previous result, its differential can be expressed in terms of the notation introduced in Remark 3.21 as

$$d\phi(v_p) = (\phi(v))_{\phi(p)},$$

where we have dropped the indices in $d\phi_p$ for simplicity.

Proposition 3.26. *Let $\phi : M \to N$ and $\psi : N \to P$ be smooth maps. Then, for each $p \in M$,*

$$d(\psi \circ \phi)_p = d\psi_{\phi(p)} \circ d\phi_p.$$

Proof. If $v \in T_pM$ and $g \in \mathcal{F}(P)$, then

$$d(\psi \circ \phi)_p(v)(g) = v(g \circ \psi\phi) = d\phi_p(v)(g \circ \psi) = \left(d\psi_{\phi(p)}(d\phi_p(v))\right)(g). \qquad \square$$

The differential map allows us to generalise the notion of velocity of a curve. Consider an open interval I and a curve $\gamma : I \to M$ on M. As a manifold, I has the identity $Id|_I$ as a global chart. In order to clarify the notation let us denote by u the (only) coordinate function of the chart $Id|_I$. Then, according to Theorem 3.19, for every $t \in I$ we can regard the coordinate vector $\left(\frac{d}{du}\right)_t$ as the unit vector in the positive u direction in T_tI.

Definition 3.27. Let $\gamma : I \to M$ be a curve. For every $t \in I$ we define its *velocity vector* $\gamma'(t)$ *at* t by

$$\gamma'(t) = d\gamma_t\left(\left.\frac{d}{du}\right|_t\right) \in T_{\gamma(t)}M.$$

Remark 3.28. The velocity vector $\gamma'(t)$ applied to some $f \in \mathcal{F}(M)$ gives

$$\gamma'(t)(f) = d\gamma_t\left(\left.\frac{d}{du}\right|_t\right)(f) = \left.\frac{d}{du}\right|_t (f \circ \gamma) = \frac{d(f \circ \gamma)}{du}(t).$$

Also, according to Proposition 3.24, the coordinate expression of $\gamma'(t)$ on a local chart (U, φ) with coordinate functions $x^1, \ldots, x^n$ is

$$\gamma'(t) = \sum_{i=1}^{n} \frac{d(x^i \circ \gamma)}{du}(t)\, \partial_i|_{\gamma(t)}.$$

The following result is the generalisation in terms of manifold theory of the usual inverse function theorem.

Theorem 3.29. *Let $\phi : M \to N$ be a smooth map and $p \in M$. Then, the differential map $d\phi_p$ is an isomorphism if and only if there exists an open neighbourhood U of p such that $\phi|_U : U \to \phi(U)$ is a diffeomorphism.*

The last result motivates the following definition.

Definition 3.30. A smooth map $\phi : M \to N$ is called a *local diffeomorphism* if $d\phi_p$ is an isomorphism for every $p \in M$.

One can see that if a local diffeomorphism is also injective and onto, then it is a diffeomorphism. As we will see, the possibility to establish a local diffeomorphism between a manifold and its tangent space will provide an essential tool for our purposes in this work.

3.3 Vector and tensor fields

All the tangent spaces T_pM of a manifold M may be glued together to form a new smooth manifold.

Definition 3.31. The *tangent bundle* TM of M is defined as the disjoint union of the tangent spaces of M:

$$TM = \bigsqcup_{p \in M} T_pM.$$

Therefore, an element of TM can be thought of a pair (p, v) where p is a point in M and v a vector tangent to M at p. There is of course a natural projection map $\pi : TM \to M$ such that $\pi(p, v) = p$. Its topology and smooth structure are defined as follows:

Given a chart (U, φ) of M with coordinate functions $\varphi = (x^1, \dots, x^n)$, let

$$\Psi_U : \pi^{-1}(U) \longrightarrow \mathbb{R}^{2n}$$

be defined by

$$(p, v) \mapsto (x^1(p), \dots, x^n(p), v^1, \dots, v^n),$$

where $\varphi(p) = (x^1(p), \dots, x^n(p))$ and $v = v^i \partial_i|_p$. The topology of TM is generated by the preimages of Ψ_U for all open sets of $\mathbb{R}^{2n}$ and all charts of M. If $\{(U_i, \varphi_i)\}$ is an atlas of M, then $\{(\pi^{-1}(U_i), \Psi_{U_i})\}$ is an atlas of TM. Hence, the tangent bundle TM is a $2n$-dimensional manifold.

The tangent bundle is the prototypical example of *vector bundle*, which is in turn a particular type of *fibre bundle*. From this point of view, the preimage $\pi^{-1}(\{p\})$, that we will denote by M_p, is called the *fibre of TM at p* and is canonically identified with T_pM:

$$T_pM \cong M_p = \{(p, v)|\ v \in T_pM\} \subset TM.$$

Definition 3.32. A *vector field on* M is a smooth map $X : M \to TM$ such that $\pi \circ X = Id$.

A vector field X is then given by $X(p) = (p, X_p)$ where $X_p \in T_pM$ and therefore assigns to each point on the manifold a vector of its tangent space. We denote by $X(f) \in \mathcal{F}(M)$ the function sending each $p \in M$ to $X_p(f) \in \mathbb{R}$, which is smooth. In the language of fibre bundles, a vector field is a *section* of the tangent bundle. We denote by $\mathcal{X}(M)$ the set of vector fields of M. One can define on $\mathcal{X}(M)$ an addition and a multiplication by real numbers by

$$(X + Y)_p := X_p + Y_p \quad \text{and} \quad (\lambda X)_p := \lambda X_p$$

for all $p \in M$, all $X, Y \in \mathcal{X}(M)$ and all $\lambda \in \mathbb{R}$. Moreover, one can also define on $\mathcal{X}(M)$ a multiplication by functions on M by

$$(fX)_p := f(p)X_p,$$

for all $p \in M$ and all $f \in \mathcal{F}(M)$. In this way, $\mathcal{X}(M)$ is a real vector space and a module over the ring $\mathcal{F}(M)$.

Let $\varphi = (x^1, \dots, x^n)$ be a chart on $U \subset M$, then for each $i = 1, \dots, n$ one can define a vector field ∂_i sending every $p \in U$ to the tangent vector $\partial_i|_p$. The vector field ∂_i is called the *coordinate vector field of φ in the x^i direction.* It follows from Theorem 3.19 that any vector field can then be expressed as

$$X = X^i \partial_i,$$

where the functions $X^i = X(x^i) : M \to \mathbb{R}$ are called *coordinate functions of* X.

Example 3.33. A very interesting example of vector field is the *Lie Bracket.* For all vector fields X, Y the Lie Bracket $[X, Y]$ of X and Y is defined as the unique vector field such that

$$[X,Y](f) = X(Y(f)) - Y(X(f)).$$

Locally, on a chart (U, φ) with $\varphi = (x_1, \ldots, x_n)$, the Lie Bracket $[X, Y]$ can be expressed in terms of the coordinate functions of X and Y as follows

$$[X,Y] = \left(X^j \partial_j Y^i - Y^j \partial_j X^i\right) \partial_i.$$

Definition 3.34. A *derivation on* $\mathcal{F}(M)$ is a map $D : \mathcal{F}(M) \to \mathcal{F}(M)$ satisfying

(i) $D(af + bg) = aD(f) + bD(g)$,
(ii) $D(fg) = D(f)g + fD(g)$,

for all $a, b \in \mathbb{R}$ and all $f, g \in \mathcal{F}(M)$.

Remark 3.35. Note how the definition of derivation resembles that of tangent vector. In fact, the latter implies that every vector field $X \in \mathcal{X}(M)$ defines a derivation on $\mathcal{F}(M)$ by setting $f \mapsto X(f)$. Conversely, any derivation D on $\mathcal{F}(M)$ defines a vector field X by letting $X_p(f) = D(f)(p)$.

Definition 3.36. Let M be a smooth manifold and $p \in M$. The dual space T_p^*M of T_pM is called the *cotangent space of* M *at* p. Its elements are called *linear forms* or *covectors.* Similarly to TM, the *cotangent bundle* T^*M *of* M is defined by $T^*M = \bigsqcup_{p \in M} T_p^*M$.

As in the case of TM, there is also a projection map $\pi : T^*M \to M$ defined by $\pi(p, \alpha) = p$. The cotangent bundle has a natural description as a smooth manifold obtained in the same way as that of the tangent bundle.

Definition 3.37. A *one-form on* M is a smooth map $\omega : M \to T^*M$ such that $\pi \circ \omega = Id$.

A one-form is then given by $\omega(p) = (p, \omega_p)$ where $\omega_p \in T_p^*M$ and therefore assigns to each point on the manifold a linear form of its cotangent space. The cotangent bundle is also an example of vector bundle, whose fibre at each point $p \in M$ is the cotangent space T_p^*M. One-forms can then be regarded as sections of T^*M, in the same way that vector fields are sections of TM. Following this analogy, we denote by $\mathcal{X}^*(M)$ the set of all one-forms on M, and again with the natural operations

$$\begin{aligned}
(\omega + \theta)_p &:= \omega_p + \theta_p, \\
(\lambda\omega)_p &:= \lambda\omega_p, \\
(f\omega)_p &:= f(p)\omega_p,
\end{aligned}$$

for all $p \in M$, all $\omega, \theta \in \mathcal{X}^*(M)$, all $\lambda \in \mathbb{R}$ and all $f \in \mathcal{F}(M)$, it is a real vector space and a module over $\mathcal{F}(M)$.

Functions, vector fields and one-forms on a manifold can be thought of as particular cases of more general objects called tensor fields. Tensor fields therefore provide the mathematical means of describing more complicated objects on a manifold. In particular, they are an essential tool in G.R. Although tensor fields may occur in very different ways, their characteristic property is multilinearity. Now, we do not intend to cover the topic exhaustively but only to introduce those notions that are essential for our work. In particular, how tensor fields provide a generalisation of the notion of inner product to smooth manifolds that is the origin of semi-Riemannian geometry. For a thorough approach to the topic, we refer the reader to [O'N83], Chapter 2.

We will first introduce the notion of tensor over an arbitrary module and then see how the notion of tensor field follows immediately. Consider a module V over a ring R and the set V^* of R-linear maps from V to R. Then V^* with the usual addition and multiplication by elements of R is also a module over R called the *dual module of* V. Note that this is only a generalisation of the results we have seen for the modules $\mathcal{X}(M)$ and $\mathcal{X}^*(M)$ over the ring $\mathcal{F}(M)$. Then, the usual component-wise operations make $(V^*)^r$ and V^s also modules over R, for all integers $r, s \geq 0$.

Definition 3.38. Let $r, s \geq 0$ be two integers, not both zero. A *tensor of type* (r, s) *over* V is an R-multilinear map

$$A : (V^*)^r \times V^s \longrightarrow R.$$

Here, we understand $A : (V^*)^r \longrightarrow R$ if $s = 0$ and $A : V^s \longrightarrow R$ if $r = 0$. A *tensor of type* $(0, 0)$ *over* V is simply an element of R.

The R-multilinearity of A means that A is R-linear in each slot, that is, that for $\alpha \in V^*$ and $v \in V$ the maps

$$\alpha \longmapsto A(\alpha_1, \dots, \alpha_{i-1}, \alpha, \alpha_{i+1}, \dots, \alpha_r, v_1, \dots v_s),$$

and

$$v \longmapsto A(\alpha_1, \dots, \alpha_r, v_1, \dots v_{j-1}, v, v_{j+1}, \dots, v_s),$$

are R-linear for all $i = 1, \dots, r$ and $j = 1, \dots, s$.

Example 3.39. Suppose V is a real vector space and V^* its dual. Then a $(0, 0)$ tensor over V is just a real number λ. A $(0, 1)$ tensor is simply a linear form $\alpha \in V^*$ since $\alpha(v) \in \mathbb{R}$ for all $v \in V$. Similarly, a $(1, 0)$ tensor over V can be regarded as a vector $v \in V$ by letting $v(\alpha) = \alpha(v) \in \mathbb{R}$, for all linear forms $\alpha \in V^*$. Finally, any bilinear form g on V is a $(0, 2)$ tensor over V. In particular, inner products on V are non-degenerate symmetric tensors of type $(0, 2)$ over V.

We denote by $\mathfrak{T}^r_s(V)$ the set of all tensors of type (r, s) over V. Defining in a natural way an addition and a multiplication by elements of R, one can see that $\mathfrak{T}^r_s(V)$ is also a module over R.

At this point, we can say that tensor fields on a manifold are simply tensors over the module of its vector fields. More precisely:

Definition 3.40. For all integers $r, s \geq 0$, a *tensor field of type (r, s) on* M is a tensor of type (r, s) over the $\mathcal{F}(M)$-module $\mathcal{X}(M)$.

This is to say that a tensor field of type (r, s) is an $\mathcal{F}(M)$-multilinear map

$$A : (\mathcal{X}^*(M))^r \times (\mathcal{X}(M))^s \longrightarrow \mathcal{F}(M).$$

Therefore, it produces smooth functions on M when evaluated over r one-forms and s vector fields: $A(\omega_1, \dots, \omega_r, X_1, \dots, X_s) \in \mathcal{F}(M)$.

Example 3.41. Smooth functions are $(0, 0)$ tensors fields, vector fields are $(1, 0)$ tensors fields and one-forms are $(0, 1)$ tensors fields.

The set of all tensors fields of type (r, s) on M is denoted by $\mathfrak{T}^r_s(M)$, and as we have seen it is a module over $\mathcal{F}(M)$.

The last issue regarding tensor fields that we want to address is how any tensor field A on M can indeed be regarded as a field on M, in the sense that it assigns a certain value A_p to each point $p \in M$, just as vector fields and one-forms do. Indeed, the value at $p \in M$ of the smooth function $A(\omega_1, \ldots, \omega_r, X_1, \ldots, X_s)$ produced by A depends not on the entirety of each one-form and each vector field evaluated, but only on their values $\omega_{1p}, \ldots, \omega_{rp}$ and $X_{1p}, \ldots, X_{sp}$ at $p \in M$. Therefore, a tensor field $A \in \mathfrak{T}^r_s(M)$ assigns to each point $p \in M$ a map

$$A_p : (T^*_p M)^r \times T_p M^s \longrightarrow \mathbb{R}$$

defined as follows. If $\alpha_1, \ldots, \alpha_r \in T_pM^*$ and $v_1, \ldots, v_s \in T_pM$, let

$$A_p(\alpha_1, \ldots, \alpha_r, v_1, \ldots, v_s) = A(\omega_1, \ldots, \omega_r, X_1, \ldots, X_s)(p),$$

where $\omega_1, \ldots, \omega_r$ are *any* one-forms on M such that $\omega_{ip} = \alpha_i$ for all $i = 1, \ldots, r$ and $X_1, \ldots, X_s$ are *any* vector fields on M such that $X_{j_p} = v_j$ for all $j = 1, \ldots, s$.

It is easy to check that A_p is $\mathbb{R}$-multilinear and thus that A_p is an (r, s) tensor over the $\mathbb{R}$-module (i.e. vector space) T_pM. Hence, we can regard $A \in \mathfrak{T}^r_s(M)$ as a field smoothly assigning to each $p \in M$ the tensor A_p.

3.4 Semi-Riemannian manifolds

As we have already said, semi-Riemannian geometry is the generalisation to smooth manifolds of semi-Euclidean geometry. Its object of study are semi-Riemannian manifolds, which are smooth manifolds equipped with a metric tensor that plays the role of the inner product in semi-Euclidean geometry.

Definition 3.42. A *metric tensor* g on a smooth manifold M is a symmetric non-degenerate $(0, 2)$ tensor field on M of constant index ν.

This is to say that a metric tensor g smoothly assigns to every $p \in M$ an inner product

$$g_p : T_pM \times T_pM \to \mathbb{R}$$

on its tangent space, and that the index of g_p is the same for all $p \in M$. The smoothness of g means that for all vector fields $X, Y \in \mathcal{X}(M)$ the function $g(X, Y) : M \to \mathbb{R}$ defined by $g(X, Y)(p) = g_p(X_p, Y_p)$ is smooth.

Definition 3.43. An *n-dimensional semi-Riemannian manifold* is a pair (M, g) where M is an n-dimensional smooth manifold and g is a metric tensor on M. We say that (M, g) is a

1. *Riemannian manifold*, if $\nu = 0$. In this case, g is called a *Riemannian metric.*
2. *Lorentzian manifold*, if $\nu = 1$ and $n \geq 2$. In this case, g is called a *Lorentz metric.*

Remark 3.44. Let (M, g) be an n-dimensionnal semi-Riemannian manifold, with g having index ν. Then the metric tensor g makes of each tangent space T_pM an inner product space of dimension n and index ν.

The last statement is the semi-Riemannian generalisation of how differential geometry allows to assign to each point on a smooth manifold a vector space of the same dimension.

The condition $\nu = 0$ in Riemannian manifolds implies that g defines on every tangent space of M a positive-definite inner product. In particular, this allows to turn every Riemannian manifold into a metric space by the definition of a distance and implies that every submanifold N of M is itself Riemannian with the restriction $g|_N$. None of these two assertions holds for arbitrary semi-Riemannian manifolds. Another important feature of the Riemannian case regards the existence of such a structure. Indeed, every smooth manifold is known to admit a Riemannian metric but it may not admit metrics of different index. For Lorentz metrics, for example, one has the following result (Prop. 5.37 in [O'N83]).

Proposition 3.45. *Let M be a smooth manifold, then the following are equivalent:*

1. *M admits a Lorentz metric.*
2. *There is a non-vanishing vector field on M.*
3. *Either M is not compact or M is compact with Euler characteristic $\chi(M) = 0$.*

Recall from Remark 3.21 that for each $p \in \mathbb{R}^n$ there is a natural isomorphism from $T_p\mathbb{R}^n$ to $\mathbb{R}^n$ sending every $v_p = v^i \partial_i|_p \in T_p\mathbb{R}^n$ to $v = v^i e_i \in \mathbb{R}^n$, where $\{e_i\}_{i=1,\dots,n}$ is the canonical basis on $\mathbb{R}^n$. If h is an inner product of index ν on $\mathbb{R}^n$ given by

$$h(v, w) = -v^1 w^1 - \cdots - v^\nu w^\nu + v^{\nu+1} w^{\nu+1} + \cdots + v^n w^n,$$

then we can define a metric tensor g of index ν on $\mathbb{R}^n$ just by letting

$$g_p(v_p, w_p) = h(v, w),$$

for each $p \in \mathbb{R}^n$. We shall denote the resulting semi-Riemannian manifold $(\mathbb{R}^n, g)$ simply by $\mathbb{R}^n_\nu$.

Examples 3.46.

1. For $\nu = 0$, the inner product h is just the standard Euclidean inner product " $\cdot$ " on $\mathbb{R}^n$. The corresponding metric tensor, that we shall denote by δ, is then defined by
$$\delta_p(v_p, w_p) = v \cdot w.$$
Thus, $(\mathbb{R}^n, \delta)$ is a Riemannian manifold, which we will refer to simply as $\mathbb{R}^n$.
2. For $\nu = 1$ and $n = 4$, the inner product h is just the inner product η on Minkowski spacetime $\mathcal{M}$. For simplicity, let us denote the resulting metric tensor also by η. Thus, Minkowski spacetime can be regarded as the 4-dimensional Lorentzian manifold $(\mathbb{R}^4, \eta)$, or simply $\mathbb{R}^4_1$.

More generally, let E be any inner product space with inner product h. Then E is a smooth manifold. If we now let $\{e_i\}_{i=1,\dots,n}$ be an orthonormal basis on E, then the natural isomorphism sending each $v_p = v^i \partial_i|_p \in T_pE$ to $v = v^i e_i \in E$ allows us to define, as before, a metric tensor g on E by letting

$$g_p(v_p, w_p) = h(v, w),$$

for each $p \in E$. This argument is summarised in the following remark.

Remark 3.47. Any inner product space is a semi-Riemannian manifold. In particular, every Euclidean vector space is a Riemannian manifold and every Lorentzian vector space is a Lorentzian manifold.

Remarks 3.44 and 3.47 are of utmost importance, as they show how semi-Riemannian geometry somehow generalises semi-Euclidean geometry from vector spaces to smooth manifolds.

It is useful for calculations to express the metric tensor in terms of its components with respect to some coordinate system. If (U, φ) is a chart on M with coordinate functions $x^1, \dots, x^n$, then the components of the metric tensor are the functions $g_{ij} : M \to \mathbb{R}$ given by

$$g_{ij} = g(\partial_i, \partial_j), \qquad i, j = 1, \dots, n.$$

Hence, for vector fields $X = X^i \partial_i$ and $Y = Y^j \partial_j$ we can write

$$g(X, Y) = g_{ij} X^i Y^j.$$

Since g is non-degenerate, the components g_{ij} form a regular matrix. We will then denote the components of its inverse matrix by g^{ij}. We shall not go into details since it will not be necessary for our purposes, but it is worth noting that these components naturally define what is called the *inverse metric tensor field* g^{-1}.

Example 3.48. The components of the metric tensor g of $\mathbb{R}^n_\nu$ are given, in terms of the Kronecker delta δ_{ij}, by

$$g_{ij} = \epsilon_j \delta_{ij}, \qquad \text{where} \qquad \epsilon_j = \begin{cases} -1, & \text{for } j = 1, \dots, \nu. \\ +1, & \text{for } j = \nu + 1, \dots, n. \end{cases}$$

In particular the metric tensor δ previously introduced for $\mathbb{R}^n$ has components δ_{ij}, hence the notation. Similarly, the metric tensor η of Minkowski spacetime $\mathbb{R}^4_1$ has components

$$\eta_{ij} = \begin{cases} -1, & \text{if } i = j = 1. \\ \delta_{ij}, & \text{otherwise.} \end{cases}$$

We now want to introduce a special type of map between semi-Riemannian manifolds: isometries. Isometries preserve metric tensors and allow to define a notion of equivalence in semi-Riemannian geometry, just as diffeomorphisms preserve smooth structures and allow to define a notion of equivalence in differential geometry. Thus, semi-Riemannian geometry can be thought of as the study of isometric invariants, in the same way that differential geometry can be regarded as the study of diffeomorphic invariants.

Definition 3.49. Let (N, g) be a semi-Riemannian manifold and M a smooth manifold and consider a map $\phi : M \to N$. We define the *pullback* $\phi^*(g)$ *of* g *by* ϕ as the map

$$\phi^* g : \mathcal{X}(M) \times \mathcal{X}(M) \to \mathcal{X}(M)$$

defined by $(\phi^* g)_p(v, w) = g_p(d\phi_p(v), d\phi_p(w))$, for all $p \in M$ and $v, w \in T_p M$.

It is easy to check that $\phi^* g$ is a $(0, 2)$ tensor field on M. However, if the index of g is different from zero, $\phi^* g$ may not be a metric tensor on M.

Definition 3.50. Let (M, g_M) and (N, g_N) be two semi-Riemannian manifolds. A map $\phi : M \to N$ is an *isometry* if it is a diffeomorphism and it preserves metric tensors, i.e., if $\phi^*(g_N) = g_M$.

Equivalently, an isometry ϕ between semi-Riemannian manifolds is a diffeomorphism for which $d\phi_p : T_p M \to T_{\phi(p)} N$ is a linear isometry for every $p \in M$.

Let (E, g) and (F, h) be two inner product spaces and let us denote by $(\tilde{E}, \tilde{g})$ and $(\tilde{F}, \tilde{h})$ the corresponding semi-Riemannian manifolds. The following result further shows how semi-Riemannian geometry generalises semi-Euclidean geometry.

Lemma 3.51. *If $\phi : E \to F$ is a linear isometry, then $\tilde{\phi} : \tilde{E} \to \tilde{F}$ is an isometry.*

Proof. Since linear maps are smooth and ϕ is a linear isomorphism, we have that $\tilde{\phi}$ is a diffeomorphism. In addition, if $v_p \in T_p E$, then by Example 3.25 we have $d\phi(v_p) = \phi(v)_{\phi(p)}$. Thus

$$\tilde{h}_p(d\phi(v_p), d\phi(w_p)) = \tilde{h}_p(\phi(v)_{\phi(p)}, \phi(w)_{\phi(p)}) = h(\phi(v), \phi(w)) = g(v, w) = \tilde{g}_p(v_p, w_p). \qquad \square$$

The next corollary then follows immediately.

Corollary 3.52. *Every inner product space of dimension n is isometric to $\mathbb{R}^n_\nu$, for some ν.*

Remark 3.53. This results shows that at each point p of a semi-Riemannian manifold M its tangent space T_pM is isometric to $\mathbb{R}^n_\nu$. In particular, by Example 3.48, this means that in terms of an orthonormal basis $\{e_i\}_{i=1,\dots,n}$ of T_pM, the inner product inherited from the metric tensor will take the simple form

$$g_p(e_i, e_j) = \epsilon_j \delta_{ij}.$$

However, let us stress that this need not be the case when considering the usual basis $\{\partial_i|_p\}_{i=1,\dots,n}$ for T_pM, since it may not be orthonormal. We shall later introduce a special coordinate system around p having this interesting property.

In analogy with differential geometry, it is interesting to know under which conditions a subset of a semi-Riemannian manifold inherits its metric structure. This idea is formalised in the following definition.

Definition 3.54. Let M be a smooth submanifold of a semi-Riemannian manifold (N, g). If the pullback $\phi^*(g)$ is a metric tensor on M, then we say that $(M, \phi^*(g))$ is a *semi-Riemannian submanifold* of (N, g).

Example 3.55. Let (M, g) be a semi-Riemannian manifold with index ν and $U \subset M$ open. Then, $(U, g|_U)$ is a semi-Riemannian submanifold of M with index ν.

3.5 Geodesic curves

Let X and Y be two vector fields on a semi-Riemannian manifold M. We are now interested in defining a new vector field such that its value at each point $p \in M$ is the vector rate of change of Y in the direction given by X_p.

Definition 3.56. A *connection* on a smooth manifold M is a map

$$D : \mathcal{X}(M) \times \mathcal{X}(M) \longrightarrow \mathcal{X}(M) \quad ; \quad (X, Y) \longmapsto D_X Y$$

such that

(i) D is $\mathcal{F}(M)$-linear in X,

(ii) D is $\mathbb{R}$-linear in Y,

(iii) $D_X(f \cdot Y) = X(f) \cdot Y + f \cdot D_X Y$, for all $f \in \mathcal{F}(M)$ and $X, Y \in \mathcal{X}(M)$.

The vector field $D_X Y$ is then called the *covariant derivative of Y with respect to X.*

Note how conditions (ii) and (iii) imply that D is a derivation in Y, hence the name. In turn, condition (i) is to say that D is tensorial in X. This means that fixing $Y \in \mathcal{X}(M)$ yields an $\mathcal{F}(M)$-linear map

$$DY : \mathcal{X}(M) \longrightarrow \mathcal{X}(M) \quad ; \quad X \longmapsto D_X Y$$

that defines a family of $\mathbb{R}$-linear maps

$$DY_p : T_pM \longrightarrow T_pM \quad ; \quad v \longmapsto D_v Y$$

by letting $D_v Y = (D_X Y)_p$, where X is any vector field such that $X_p = v$. Therefore, the notion of covariant derivative of Y can be considered with respect to tangent vectors, and not only vector fields.

Example 3.57. Let $u^1, \ldots, u^n$ be the natural coordinates on $\mathbb{R}^n_\nu$. For every X, Y vector fields on $\mathbb{R}^n_\nu$, the map sending (X, Y) to the vector field

$$D_X Y = X(Y^i)\partial_i$$

is a connection called *flat connection on* $\mathbb{R}^n_\nu$.

Definition 3.58. A connection D on semi-Riemannian manifold (M, g) is said to be

1. *symmetric*, if $D_X Y - D_Y X = [X, Y]$.
2. *compatible with the metric tensor*, if $X(g(Y, Z)) = g(D_X Y, Z) + g(Y, D_X Z)$,

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

The existence and uniqueness of such a connection is guaranteed by the following result, usually known as the fundamental theorem of semi-Riemannian geometry. We shall not include its proof here, but we refer the reader to, for instance, Theorem 3.11 in [O'N83].

Theorem 3.59. *On a semi-Riemannian manifold there exists a unique connection that is symmetric and compatible with the metric tensor.*

Definition 3.60. We define the *Levi-Civita connection* on a semi-Riemannian manifold (M, g) as the unique connection on M that is both symmetric and compatible with g.

Remark 3.61. Straightforward computations using properties (i) and (ii) show that the Levi-Civita connection D satisfies the *Koszul formula*:

$$2g(D_X Y, Z) = X(g(Y, Z)) + Y(g(Z, X)) - Z(g(X, Y)) - g(X, [Y, Z]) + g(Y, [Z, X]) + g(Z, [X, Y]).$$

In fact, one can define the Levi-Civita connection via the Koszul formula and then show that properties (i) and (ii) hold.

From now on, D will denote the Levi-Civita connection on M, unless otherwise specified. The following definition introduces the functions that locally characterise the Levi-Civita connection.

Definition 3.62. Let (U, φ) be a chart on a semi-Riemannian manifold M with coordinate functions $x^1, \ldots, x^n$. We define the *Christoffel symbols* for (U, φ) as the functions $\Gamma^k_{ij} : U \to \mathbb{R}$ such that

$$D_{\partial_i}\partial_j = \Gamma^k_{ij}\partial_k, \qquad i, j = 1, \ldots, n.$$

Remark 3.63. Since $[\partial_i, \partial_j] = 0$, it follows from the symmetry property of D that the Christoffel symbols are symmetric in the lower indices, namely $\Gamma^k_{ij} = \Gamma^k_{ji}$.

Proposition 3.64. *Let (U, φ) be a chart on a semi-Riemannian manifold (M, g), with coordinate functions $x^1, \ldots, x^n$. Then, for every vector field $Y \in \mathcal{X}(M)$ and all $i, j, k = 1, \ldots, n$,*

1. $D_{\partial_i} Y = \left(\partial_i Y^k + \Gamma^k_{ij} Y^j\right) \partial_k$.
2. $\Gamma^k_{ij} = \frac{1}{2} g^{kl} \left(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}\right)$,

where $1 \leq l \leq n$.

Proof. Let $Y = Y^k\partial_k$, then (1) is obtained by direct application of property (iii) in Definition 3.56 together with the definition of the Christoffel symbols. To prove (2), we apply the Koszul formula for $X = \partial_i$, $Y = \partial_j$ and $Z = \partial_l$. Since $[\partial_i, \partial_j] = 0$ for all $i, j = 1\dots, n$, we get

$$2g(D_{\partial_i}\partial_j, \partial_l) = 2g(\Gamma^m_{ij}\partial_m, \partial_l) = 2\Gamma^m_{ij}g_{ml} = \partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}.$$

The final result can then be obtained by multiplying the last equality by g^{kl}:

$$2\Gamma^m_{ij}g_{ml}g^{kl} = 2\Gamma^m_{ij}\delta^k_m = 2\Gamma^k_{ij} = g^{kl}(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}). \qquad \square$$

Remark 3.65. It can be seen that the flat connection on $\mathbb{R}^n_\nu$ is symmetric and compatible with the metric tensor and hence is the Levi-Civita connection on $\mathbb{R}^n_\nu$. As shown in Example 3.48, the components of the metric tensor of $\mathbb{R}^n_\nu$ are constant and therefore by the previous result the Christoffel symbols vanish everywhere:

$$\Gamma^k_{ij} = 0, \qquad \text{for all } i, j, k = 1, \dots, n.$$

Our next goal is to generalise the notion of straight line in semi-Euclidean geometry via the introduction of geodesic curves on a semi-Riemannian manifold. First, however, we shall see how to properly describe objects such as vector fields or covariant derivatives when only considered along the trajectory of a curve.

Definition 3.66. Let $\gamma : I \to M$ be a curve on M. A *vector field on* γ is a smooth map $V : I \to TM$ such that $\pi \circ V = \gamma$.

A vector field V on γ is then given by $V(t) = (\gamma(t), V_{\gamma(t)})$ and therefore it smoothly assigns to each $t \in I$ a vector tangent to M at $\gamma(t)$.

Examples 3.67.

1. The map sending each $t \in I$ to $(\gamma(t), \gamma'(t))$ is a vector field on γ, called its *velocity vector field*. We will also denote it by γ' whenever it is understood from the context whether we refer to the velocity vector field or the velocity vector.
2. The restriction to $\gamma(I)$ of any vector field X on M naturally defines a vector field X_γ on γ by letting $X_\gamma(t) = (\gamma(t), X_{\gamma(t)})$.

We denote by $\mathcal{X}(\gamma)$ the set of all vector fields on γ, which is a module over the ring $\mathcal{F}(I)$. For every $V \in \mathcal{X}(\gamma)$, the following result provides a natural way to define its vector rate of change.

Proposition 3.68. *Let $\gamma : I \to M$ be a curve on M and $V \in \mathcal{X}(\gamma)$. Then, there is a unique map*

$$\frac{D}{Dt} : \mathcal{X}(\gamma) \to \mathcal{X}(\gamma) \quad ; \quad V \mapsto \frac{D}{Dt}V$$

such that

1. $\frac{D}{Dt}(aV + bW) = a\frac{D}{Dt}V + b\frac{D}{Dt}W$,
2. $\frac{D}{Dt}(fV) = \frac{df}{dt}V + f\frac{D}{Dt}V$,
3. $\frac{D}{Dt}(X_\gamma) = D_{\gamma'}X$,

for all $a, b \in \mathbb{R}$, $V, W \in \mathcal{X}(\gamma)$ and $f \in \mathcal{F}(I)$.

Proof. Let us begin by proving uniqueness assuming existence. We can assume without loss of generality that $\gamma(I)$ lies entirely in the domain of a single chart (U, φ) with coordinate functions $x^1, \ldots, x^n$. Then, in terms of its coordinate functions $V^i : I \to \mathbb{R}$ defined by $V^i(t) = V_{\gamma(t)}(x^i)$, every $V \in \mathcal{X}(\gamma)$ can be expressed as $V = V^i \partial_{i\gamma}$. Let us drop the index γ in $\partial_{i\gamma}$ for clarity. Using the properties above, we have

$$\frac{D}{Dt}V = \frac{D}{Dt}(V^i \partial_i) = \frac{dV^i}{dt}\partial_i + V^i \frac{D}{Dt}\partial_i = \frac{dV^i}{dt}\partial_i + V^i D_{\gamma'}\partial_i.$$

Therefore, $\frac{D}{Dt}$ is determined by the previous coordinate expression, and its uniqueness follows from the uniqueness of D.

Regarding the existence, consider any subinterval $J \subset I$ such that $\gamma(J)$ is entirely contained in the domain of some chart on M. Then, it suffices to define $\frac{D}{Dt}$ by the formula above. Straightforward computations show that it fulfills the three required properties. By the uniqueness, these local definitions constitute a single vector field in $\mathcal{X}(\gamma)$. $\square$

Definition 3.69. The map $\frac{D}{Dt}$ in Proposition 3.68 is called the *induced covariant derivative on* γ.

Remark 3.70. We may rewrite the above expression for $\frac{D}{Dt}V$ by introducing the Christoffel symbols:

$$\frac{D}{Dt}V = \left\{ \frac{dV^k}{dt} + \Gamma^k_{ij} \frac{d(x^i \circ \gamma)}{dt} V^j \right\} \partial_k.$$

Definition 3.71. The *acceleration* of a curve $\gamma : I \to M$ is the vector field γ'' on γ defined by

$$\gamma'' = \frac{D}{Dt}\gamma'.$$

Definition 3.72. A *geodesic* on M is a curve $\gamma : I \to M$ such that $\gamma'' = 0$.

Proposition 3.73. *Let (U, φ) be a chart on M with coordinate functions $x^1, \ldots, x^n$. Then, a curve $\gamma : I \to U$ is a geodesic on M if and only if its coordinate functions $x^k \circ \gamma$ satisfy the system of differential equations*

$$\frac{d^2(x^k \circ \gamma)}{dt^2} + \Gamma^k_{ij} \frac{d(x^i \circ \gamma)}{dt} \frac{d(x^j \circ \gamma)}{dt} = 0.$$

Proof. It follows from the definition of geodesic by using Remark 3.70 in the particular case where $V = \gamma'$. $\square$

Example 3.74. Let $u^1, \ldots, u^n$ be the natural coordinates on $\mathbb{R}^n_\nu$. Then, using the previous result, the geodesics on $\mathbb{R}^n_\nu$ satisfy

$$\frac{d^2(u^i \circ \gamma)}{dt^2} = 0, \qquad i = 1, \ldots, n,$$

where we have used the vanishing of the Christoffel symbols for $\mathbb{R}^n_\nu$ shown in Remark 3.65. Solving the system of differential equations yields

$$\gamma(t) = p + tv,$$

for some $p, v \in \mathbb{R}^n_\nu$. Therefore, the geodesics of $\mathbb{R}^n_\nu$ are straight lines.

Remark 3.75. It can be shown by direct substitution into the geodesic equations that a linear reparametrisation of a geodesic is again a geodesic. Furthermore, one can show that these are the only reparametrisations that preserve the geodesical character. Most of the following results involving geodesics will also hold for their linear reparametrisations, although we may not explicitly specify it.

The following result is then a consequence of the local existence and uniqueness theorem for ordinary differential equations.

Corollary 3.76. *For every $p \in M$ and every $v \in T_pM$ there is an interval $I \subset \mathbb{R}$ with $0 \in I$ and a unique geodesic $\gamma : I \to M$ such that $\gamma(0) = p$ and $\gamma'(0) = v$.*

In this case, we say that γ is a geodesic *starting at p with initial velocity v*.

Definition 3.77. A geodesic $\gamma : I \to M$ starting at p with initial velocity v is said to be *maximal* or *geodesically inextendible* if for every geodesic $\alpha : J \to M$ starting at p with initial velocity v, then $J \subset I$ and $\alpha = \gamma|_J$.

The following result is an application of the existence and uniqueness theorem of maximal solutions for ordinary differential equations.

Proposition 3.78. *For every $p \in M$ and every $v \in T_pM$ there is a unique maximal geodesic $\gamma_{p,v} : I_{p,v} \to M$ starting at p with initial velocity v.*

Note that the condition $\gamma'(0) = v$ for $v \in T_pM$ already implies that $\gamma(0) = p$. Therefore, whenever it is understood to which tangent space T_pM the tangent vector v belongs to, we shall drop the subindex p and simply write γ_v and I_v to refer to the maximal geodesic starting at p with initial velocity v and its domain of definition.

Definition 3.79. A semi-Riemannian manifold for which every maximal geodesic is defined on the entire real line is said to be *(geodesically) complete*.

Example 3.80. Since its geodesics are lines, $\mathbb{R}^n_\nu$ is geodesically complete.

Let us denote by $\mathcal{D}_p$ the set of all vectors $v \in T_pM$ for which γ_v is defined at least in $[0, 1]$, that is, $[0, 1] \subset I_v$. Note that $\mathcal{D}_p \neq \emptyset$ because $0 \in \mathcal{D}_p$. Indeed, the constant map $\gamma_0(t) = p$ defined for all $t \in \mathbb{R}$ is a geodesic starting at p with initial velocity 0.

Definition 3.81. We define the *exponential map at $p \in M$* by

$$\begin{aligned} exp_p : T_pM \supset \mathcal{D}_p &\longrightarrow M \\ v &\mapsto exp_p(v) = \gamma_v(1). \end{aligned}$$

Of course $\mathcal{D}_p$ is the largest subset of T_pM on which exp_p can be defined. Note also that if M is geodesically complete, then for every $p \in M$, we have $\mathcal{D}_p = T_pM$ and so exp_p is defined globally.

Example 3.82. According to Example 3.74, for every $p \in \mathbb{R}^n_\nu$ and $v_p \in T_p\mathbb{R}^n_\nu$, we have $\gamma_{v_p}(t) = p + tv$. Therefore,

$$exp_p(v_p) = p + v.$$

It follows that exp_p is a diffeomorphism since it is the composition of the natural isomorphism $T_p\mathbb{R}^n_\nu \cong \mathbb{R}^n_\nu$ and the translation $x \mapsto p + x$. Moreover, if $T_p\mathbb{R}^n_\nu$ is given its usual metric tensor, then exp_p is an isometry.

Remark 3.83. For a fixed $v \in T_pM$, then for all $\lambda \in \mathbb{R}$, since a linear parametrisation of a geodesic is a geodesic, the map $t \mapsto \gamma_v(\lambda t)$ is a geodesic starting at p with initial velocity $w = \lambda\gamma_v'(0) = \lambda v$. Hence,

$$\gamma_{\lambda v}(t) = \gamma_v(\lambda t), \qquad \text{for all } \lambda \in \mathbb{R} \text{ and } t \in I_{\lambda v}.$$

It follows that $exp_p(\lambda v) = \gamma_{\lambda v}(1) = \gamma_v(\lambda)$ and therefore exp_p carries lines through the origin in T_pM to geodesics on M. Moreover,

- If $v \in \mathcal{D}_p$, then for all $0 \leq \lambda \leq 1$, $\lambda v \in \mathcal{D}_p$.
- If $v \notin \mathcal{D}_p$, then there exists some $\epsilon > 0$ such that $\epsilon v \in \mathcal{D}_p$.

Therefore, $\mathcal{D}_p$ contains a disk in T_pM centered at the origin and in particular it contains an open neighborhood V of T_pM around 0.

Remark 3.84. Note also that the smooth dependence of solutions to ordinary differential equations with respect to initial conditions applied to the system of differential equations defining the geodesics shows that the exponential map exp_p is smooth, in the usual sense, for every $p \in M$.

We now want to show that the exponential map exp_p is a local diffeomorphism between the tangent space T_pM and M. To do so, let us regard T_pM as a smooth manifold and consider its tangent space $T_0(T_pM)$ at the origin. There is of course a natural identification $T_0(T_pM) \cong T_pM$ given by $v_0 = v_0^i\partial_i|_0 \mapsto v = v^i\partial_i|_p$.

Lemma 3.85. *Let M be geodesically complete and $p \in M$. If we identify $T_0(T_pM)$ with T_pM, then*

$$d(exp_p)_0 = Id|_{T_pM}.$$

Proof. Define a curve $\alpha : I \to T_pM \cong T_0(T_pM)$ by $\alpha(t) = tv$, and hence such that $\alpha(0) = 0$ and $\alpha'(0) = v_0 \sim v$. Then, $exp_p \circ \alpha : I \to M$, $t \mapsto exp_p(tv)$, is a curve on M with $(exp_p \circ \alpha)(0) = exp_p(0) = p$. However, as noted in Remark 3.83, $exp_p(tv) = \gamma_v(t)$ and therefore

$$d(exp_p)_0(v) = d(exp_p)_0(\alpha'(0)) = (exp_p \circ \alpha)'(0) = \gamma_v'(0) = v. \qquad \square$$

Requiring M to be complete is only necessary in order for exp_p to be defined in all $T_0(T_pM) \cong T_pM$. However, we could have relaxed this hypothesis and get a local version of the previous result for the open neighbourhood V around 0 in T_pM given in Remark 3.83. Locally, then, we have that $d(exp_p)_0|_V = Id|_V$. Note that this is true only because V is an open submanifold of T_pM. In particular, $d(exp_p)_0|_V$ is an isomorphism and so the next result follows immediately from the inverse function theorem (Theorem 3.29).

Corollary 3.86. *For every $p \in M$ there is a neighbourhood V of $0 \in T_pM$ and a neighbourhood U of $p \in M$ such that $exp_p : V \to U$ is a diffeomorphism.*

Definition 3.87. A non-empty subset S of a vector space E is called *starshaped* if for every $v \in S$ then also $\lambda v \in S$, for all $0 \leq \lambda \leq 1$.

Example 3.88. The set $\mathcal{D}_p \subset T_pM$ in which exp_p is defined is starshaped by Remark 3.83.

Definition 3.89. A *normal neighbourhood* U of a point $p \in M$ is a subset of M such that there is a starshaped neighbourhood V of the origin in T_pM with exp_p acting as a diffeomorphism between V and U.

Example 3.90. Let $p \in M$. In Remark 3.83 one can always choose the neighbourhood $V \subset \mathcal{D}_p$ to be starshaped, taking for example $V = B_\epsilon(0)$, for some $\epsilon > 0$. Then by Corollary 3.86, $exp_p : V \to exp_p(V)$ is a diffeomorphism. Hence, $U = exp_p(V)$ is a normal neighbourhood of $p \in M$.

The following result somehow generalises the notion of starshapedness from vector spaces to semi-Riemannian manifolds.

Proposition 3.91. *If $U = exp_p(V)$ is a normal neighbourhood of $p \in M$, then for every $q \in U$ there is a unique geodesic $\alpha : [0,1] \to U$ from p to q lying entirely in U. Furthermore, $\alpha'(0) = exp_p^{-1}(p) \in V$.*

Proof. By definition V is starshaped around $0 \in T_pM$ and $exp_p : V \to U$ is a diffeomorphism. For every $q \in U$ consider $v = exp_p^{-1}(q) \in V$. Since V is starshaped, the segment $\rho(t) = tv$ $(0 \leq t \leq 1)$ lies in V. Therefore, the geodesic segment $\alpha = exp \circ \rho$ lies entirely in U and goes from p to q, thus proving the existence.

Now, since $\rho'(0) = v_0$ we have

$$\alpha'(0) = (exp_p \circ \rho)'(0)) = d(exp_p)_0(\rho'(0) = d(exp_p)_0(v_0) = v.$$

Suppose $\beta : [0,1] \to U$ is an arbitrary geodesic in U from p to q. If $w = \beta'(0)$, then the geodesic $t \mapsto exp_p(tw)$ and β both start at p with the same initial velocity, hence are equal. Now, the segment $r(t) = tw$ $(0 \leq t \leq 1)$ does not leave V, for if it did there would be some $0 < t_0 < 1$ such that $t_0 w \in V$ but $exp_p(t_0 w) \in U \setminus \beta([0,1])$. Thus $w \in V$. But $exp_p(w) = \beta(1) = q = exp_p(v)$ and exp_p is injective, hence $w = v$. Finally, by the uniqueness of geodesics, $\beta = \alpha$. $\square$

Definition 3.92. For any $p, q \in U$, the geodesic given in Proposition 3.91 is called the *radial geodesic from p to q.*

Normal neighbourhoods allow to define a special coordinate system with very interesting and useful properties. Take $p \in M$ and fix an orthonormal basis $\{e_i\}_{i=1,\dots,n}$ for T_pM. If U is a normal neighbourhood of p, and denoting by ϕ the natural isomorphism $\phi : \mathbb{R}^n \xrightarrow{\sim} T_pM$, then we have

$$\mathbb{R}^n \overset{\phi}{\cong} T_pM \supset V \xrightarrow{exp_p} U \subset M,$$

showing that (U, φ), where $\varphi = \psi^{-1} \circ exp_p^{-1}$, is a local chart on M around p. Its coordinate functions $x^1, \dots, x^n$ thus define a coordinate system $\{U; x^1, \dots x^n\}$.

Definition 3.93. A coordinate system $\{U; x^1, \dots x^n\}$ as defined above is called a *normal coordinate system at p*. Every point $q \in U$ is then said to have *normal coordinates* $(x^1(q), \dots, x^n(q))$.

The normal coordinate system determined by $\{e_i\}_{i=1,\dots,n}$ establishes via the exponential map a correspondence between points $q \in U$ having normal coordinates $(x^1(q), \dots, x^n(q))$ and vectors in V having linear coordinates $(x^1(q), \dots, x^n(q))$ relative to $\{e_i\}_{i=1,\dots,n}$. That is,

$$exp_p^{-1}(q) = x^i(q) e_i.$$

This fact already shows the adequateness of such coordinates systems and how they may allow to simplify calculations on manifolds. For instance, let $v = v^i e_i \in T_pM$ and consider the geodesic $\gamma_v(t) = exp_p(tv)$. Then, the point $\gamma_v(t)$ for each t such that γ_v remains in U has normal coordinates

$$\gamma_v(t) = (tv^1, \dots, tv^n).$$

Now, note that the basis $\{e_i\}_{i=1,\dots,n}$ for T_pM being orthonormal implies that $g_p(e_i, e_j) = \epsilon_j \delta_{ij}$. As we anticipated in Remark 3.53, this is in general not true for $g_p(\partial_i|_p, \partial_j|_p)$. However, as the next result shows, it is true for normal coordinates.

Proposition 3.94. *Let $\{U; x^1, \dots, x^n\}$ be a normal coordinate system at $p \in M$. Then, for all $i, j, k = 1, \dots, n$:*

$$g_{ij}(p) = \epsilon_j \delta_{ij} \qquad \textit{and} \qquad \Gamma^k_{ij}(p) = 0.$$

Proof. Let $v = v^i e_i \in T_pM$ and consider $\gamma_v(t) = exp_p(tv)$. We have seen that

$$x^i(\gamma_v(t)) = tv^i, \quad \text{for all } i = 1, \dots, n.$$

But then $v = \gamma'(0) = v^i \partial_i|_p$ and comparing the two different expression for v gives $e_i = \partial_i|_p$, for all $i = 1, \dots, n$. Thus

$$g_{ij}(p) = g(\partial_i, \partial_j)(p) = g_p(\partial_i|_p, \partial_j|_p) = g_p(e_i, e_j) = \epsilon_j \delta_{ij}.$$

Now, plugging the previous expression for $x^i(\gamma_v(t))$ in the geodesic equation in Proposition 3.73 gives

$$\Gamma^k_{ij}(\gamma_v(t))v^i v^j = 0 \quad \Rightarrow \quad \Gamma^k_{ij}(p) v^i v^j = 0,$$

for all $k = 1, \dots, n$, after evaluating at $t = 0$. For a fixed k, this must hold for all $v = (v^1, \dots, v^n) \in \mathbb{R}^n$. It follows that all the eigenvalues of the symmetric matrix $\left(\Gamma^k_{ij}(p)\right)_{1 \le i,j \le n}$ are zero, and hence $\Gamma^k_{ij}(p) = 0$. $\square$

To end this section, we introduce the notion of convexity in a semi-Riemannian manifold.

Definition 3.95. An open subset C in a semi-Riemannian manifold M is *convex* if it is a normal neighbourhood of each of its points.

In particular, by Proposition 3.91, for any two points $p, q \in C$ there is a unique geodesic segment from p to q lying entirely in C. It is worth noting that, in contrast to the usual notion of convexity in $\mathbb{R}^n$, there might as well be other geodesics from p to q that do not remain entirely in C.

Convex subsets will be useful for our subsequent discussion, in particular we will make use of the following result (see for instance Proposition 5.7 in [O'N83] for a proof).

Proposition 3.96. *Every point $p \in M$ has a convex neighborhood.*

4 General spacetimes

The aim of this section is to introduce the notion of spacetime from a mathematical perspective. As we will see, the description of general spacetimes will rely heavily on that of Minkowski spacetime via some of the results of differential and semi-Riemannian geometry that have been previously introduced.

In Section 2 we saw how Minkowski spacetime, and more generally the theory of SR, offers an adequate framework to study the laws of Physics in absence of gravity. The ambition to include gravitational phenomena in the description of spacetime led Einstein to formulate the theory of GR, in which gravity is represented by the metric tensor of the semi-Riemannian manifold that is spacetime.

The necessity to account for causality, which motivated the choice of a Lorentzian vector space structure for Minkowski spacetime, imposes that the semi-Riemannian manifold representing spacetime be Lorentzian. Again, we shall rely on human experience to fix the dimension of the manifold to 4, although the majority of the results that will be introduced generalise to arbitrary dimension. The discussions carried out in Section 3.4 now give us further mathematical insight into GR. For instance, we saw how the metric tensor of any Lorentzian manifold makes of each of its tangent spaces a Lorentzian vector space. Furthermore, we showed the local existence at every point of a normal coordinate system, with respect to which the description of the Lorentzian tangent space is exactly that of Minkowski spacetime. In this sense, one could say that the mathematical meaning behind the Principle of Equivalence is encoded in Proposition 3.94.

Let us discuss some other motivations that will lead to our definition of a "mathematical spacetime". As for Minkowski spacetime, we shall rely on human experience to fix the dimension of the Lorentzian manifold to 4, although the majority of the results that will be introduced generalise to arbitrary dimension. Since we still think of spacetimes as models of the history (or some part of the history) of the universe (or some portion of it), we shall only consider connected manifolds, as there would be no way for us to ever know of the existence of a disconnected component. The smoothness assumption also corresponds to our intuitive notions of space and time and is probably the most reasonable one for a mathematician. However, let us stress that at the same time it is perhaps the most unclear one from a physicist's point of view. Indeed, understanding how spacetime behaves at extremely small scales by means of a quantum theory of gravity is one of the biggest challenges that Physics faces nowadays.

A mathematical spacetime following all these motivations, basically a 4-dimensional connected Lorentzian manifold, may still fail to account for causality. In the subsequent discussion we shall address this issue by introducing the notion of *time-orientability* of Lorentzian manifolds, which we shall add as a last requirement in our definition of spacetime. As we will see, even then some non-physical behaviours such as causal paradoxes might be possible. This issue will be dealt in Section 4.3 by introducing further hypothesis on spacetimes rather than by restricting the definition.

The general references for this section are [O'N83], [SW77] and [HE73].

4.1 Lorentzian manifolds and spacetimes

Let M be an n-dimensional Lorentzian manifold. The fact that every tangent space T_pM of a M is a Lorentzian vector space, hence isometric to $\mathbb{R}^n_1$, implies that all vectors tangent to M are naturally asigned a causal character. More specifically:

Definition 4.1. Let (M, g) be a Lorentzian manifold and $p \in M$. A vector $v \in T_pM$ is said to be

1. *timelike* if $g_p(v, v) < 0$,
2. *null* or *lightlike* if $g_p(v, v) = 0$ and $v \neq 0$,
3. *spacelike* if $g_p(v, v) > 0$ or $v = 0$.

A vector v is called *causal* if it is not spacelike.

In particular, curves on a manifold may also have a causal character depending on the causal character of its velocity vector.

Definition 4.2. A curve $\gamma : I \to M$ is *timelike, null (or lightlike)* or *spacelike* if $\gamma'(t)$ has that causal character for all $t \in I$. A curve $\gamma : I \to M$ is called *causal* if $\gamma'(t)$ is non-spacelike for all $t \in I$.

Remark 4.3. Note that a curve need not have a causal character. Note also that a curve γ is causal if $g_{\gamma(t)}(\gamma'(t), \gamma'(t)) \leq 0$ for all $t \in I$ and therefore timelike and null curves are included in this definition, but also are curves without a causal character whose velocity vector may change from timelike to null.

A causal character can also be assigned to some smooth submanifolds of M as follows. If $S \subset M$ is a submanifold of M such that the subspace T_pS has the same causal character in T_pM (see Definition 2.11) for all $p \in S$, then that causal character is assigned to the S itself. It follows that an arbitrary submanifold need not have a causal character and that semi-Riemannian submanifolds of a Lorentzian manifold are either timelike or spacelike. The set of null vectors in Minkowski spacetime $\mathbb{R}^4_1$ is an example of null submanifold of $\mathbb{R}^4_1$.

In Section 2.3 we saw how the existence of a causal character for elements of $\mathcal{M}$ led to the emergence of its causal structure due to the possibility to separate timelike vectors into two connected components. Now, we are interested in determining under which conditions a causal structure may arise on Lorentzian manifolds. To that purpose, we shall make extensive use of the results in Section 2.3, which as we stressed can be straighforwardly generalised from dimension 4 to n.

Let us begin by considering, for each $p \in M$, the set $\mathcal{T}_0$ of timelike vectors in T_pM. Since $\mathcal{T}_0$ is open in T_pM, it is an open submanifold of T_pM of dimension n. Furthermore, by Corollary 2.26, $\mathcal{T}_0$ has two connected components and we know that an arbitrary labelling of the two determines a time-orientation for T_pM. A fundamental question then arises: is it possible to time-orient every tangent space T_pM in a suitably consistent way?

The first step in order to answer the previous question is to introduce a notion of causal character in the whole tangent bundle TM. This is done in the following natural way: the causal character of $(p, v) \in TM$ will simply be that of $v \in T_pM$. By doing so, we will be able to somehow deal with all the tangents spaces at the same time, rather than dealing with each tangent space separately. It then makes sense to consider, for instance, the set $\mathcal{T} \subset TM$ of timelike elements of TM. Now, recall from Section 3 that we can identify each tangent space T_pM with the fibre $M_p = \pi^{-1}(\{p\})$ of the tangent bundle. In particular, this means that for each $p \in M$ there is also a natural identification of the set $\mathcal{T}_0$ of timelike vectors in T_pM with the set $\mathcal{T}_p := M_p \cap \mathcal{T} \subset TM$. Since $\mathcal{T}_0$ has two connected components, also does $\mathcal{T}_p$. Given the natural identification between $\mathcal{T}_0$ and $\mathcal{T}_p$, from now on we shall use $\mathcal{T}_p$ in both cases, whether it is a subset of T_pM or of $M_p \subset TM$ being understood from the context.

The following is somehow a generalisation of both Proposition 2.25 and Corollary 2.26.

Proposition 4.4. *Let M be a connected Lorentzian manifold. Then, $\mathcal{T}$ is an open submanifold of TM having either one or two connected components.*

Proof. Define $h : TM \to \mathbb{R}$ by $h(p,v) = g(v,v)$. Since h is $\mathcal{C}^\infty$, we have that $\mathcal{T} = h^{-1}(-\infty, 0)$ is open in TM, hence $\mathcal{T}$ is an open submanifold. Let A be a connected component of $\mathcal{T}$. Denote by $\psi : \mathcal{T} \to \mathcal{T}$ the homeomorphism defined by $\psi(p,v) = (p,-v)$, then $\psi(A)$ is also a connected component of $\mathcal{T}$. We want to show that $\mathcal{T} = A \cup \psi(A)$. Let $B = A \cup \psi(A)$ and $C = \mathcal{T} \setminus B$. Since $\mathcal{T}$ is a manifold, its connected components are both open and closed, thus B is open and closed in $\mathcal{T}$. But then also C is open and closed in $\mathcal{T}$. It follows that B and C are open in TM.

We claim $\pi(B) \cap \pi(C) = \emptyset$. Suppose otherwise, i.e., that there exist $(p,w) \in B$ and $(p,u) \in C$ for some $p \in M$. Let $D \subset M_p$ be that one of the two components of $\mathcal{T}_p$ in which (p,u) lies. Then $D \cap C \neq \emptyset$. Since C is a union of connected components of $\mathcal{T}$, this implies $D \subset C$. Now, either (p,w) or $(p,-w)$ is in D (see proof of Proposition 2.25), while both are in B by definition of B. Thus $B \cap D \neq \emptyset$, and therefore also $D \subset B$ since B is a union of connected components. It follows that $B \cap C \neq \emptyset$, which is a contradiction.

We therefore have $\pi(B) \cap \pi(C) = \emptyset$. Since $\pi(B) \cup \pi(C) = M$ and M is connected, this means that $\pi(C) = \emptyset \Rightarrow C = \emptyset \Rightarrow \mathcal{T} = A \cup \psi(A)$. If $A \cap \psi(A) = \emptyset$, then $\mathcal{T}$ has two connected components. Otherwise, $A = \psi(A)$ and $\mathcal{T}$ has only one connected component. □

Definition 4.5. A connected Lorentzian manifold M is said to be *time-orientable* if and only if $\mathcal{T}$ has two connected components. A *time orientation* for M is a labeling of the two components of $\mathcal{T}$ as $\mathcal{T}^+$ (called the *future*) and $\mathcal{T}^-$ (called the *past*). In this case, we say that M is *time-oriented.*

Remark 4.6. Our approach to time-orientability of Lorentzian manifolds has been that of [SW77], which is somehow in the line of how we addressed this same issue for Minkowski spacetime. Many references ([O'N83], [HE73]), however, offer another characterisation of time-orientability, based on the existence of an everywhere timelike vector field on M.

A time-orientation on M determines a consistent time-orientation on each of its tangent spaces. Indeed, $\mathcal{T}_p^+ := \mathcal{T}^+ \cap M_p$ and $\mathcal{T}_p^- := \mathcal{T}^- \cap M_p$ are the two connected components of each $\mathcal{T}_p$ and are to be labeled as the future and the past, respectively, of each tangent space. At this point we can generalise the notions of future and past to null tangent vectors as well, in absolute analogy with what was done for Minkowski spacetime. Concretely, by Corollary 2.32, the set $\mathcal{N}_p$ of null vectors of T_pM has two connected components: $\mathcal{N}_p^+ := \{w \in \mathcal{N}_p |\, g_p(w,v) < 0,\ \forall v \in \mathcal{T}_p^+\}$ and $\mathcal{N}_p^- := \{w \in \mathcal{N}_p |\, g_p(w,v) > 0,\ \forall v \in \mathcal{T}_p^-\}$. Recall also from Remark 2.33 that in the case $n = 2$, $\mathcal{N}_p$ splits in 4 connected components that can still be grouped pairwise in $\mathcal{N}_p^+$ and $\mathcal{N}_p^-$.

Remark 4.7. Each of the two components of $\mathcal{T}_p$ is diffeomorphic to $\mathbb{R}^n$ and for $n \geq 3$ each of the two components of $\mathcal{N}_p$ is diffeomorphic to $\mathbb{R} \times S^{n-2}$.

We are now able to classify causal tangent vectors and curves on a time-oriented Lorentzian manifold M according to their future or past directions.

Definition 4.8. A causal vector $v \in T_pM$ is said to be *future-directed*, (resp. *past-directed*) if $v \in \mathcal{T}_p^+ \cup \mathcal{N}_p^+$ (resp. $v \in \mathcal{T}_p^- \cup \mathcal{N}_p^-$). A smooth causal curve γ is said to be *future-directed*, (resp. *past-directed*) if $\gamma'(t)$ is everywhere future-directed (resp. past-directed).

The time-orientability or not of a Lorentzian manifold is independent of its orientability as a smooth manifold and involves not only the underlying smooth structure, but also the Lorentzian structure. Thus, a smooth manifold M may admit two different Lorentzian metrics g_1 and g_2 in such a way that (M, g_1) is time-orientable and (M, g_2) is not. One can convince oneself of this by looking at the following figure:

To summarise, we have seen that time-orientability is not given for an arbitrary connected Lorentzian manifold although of course it is a necessary condition for the causal structure

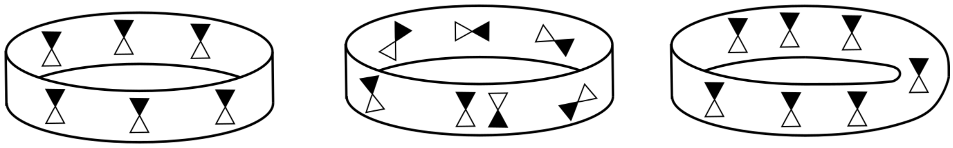

Figure 3: Form left to right: a time-orientable Lorentz metric on the orientable band $S^1 \times \mathbb{R}$, a non time-orientable Lorentz metric on the orientable band $S^1 \times \mathbb{R}$ and a time-orientable Lorentz metric on the non-orientable Möbius band.

to emerge. Therefore, when aiming to describe spacetime we shall restrict our attention to time-orientable (more specifically, time-oriented) Lorentzian manifolds. We can finally give the following definition that formalises the notion of mathematical spacetime.

Definition 4.9. A *spacetime* is a connected 4-dimensional time-oriented Lorentzian manifold (M, g). A point $p \in M$ is called an *event.*

Two spacetimes (M, g) and (M', g') are physically equivalent (in the sense that they define the same gravitational field) if they are isometric. Thus, strictly speaking, a spacetime is a whole equivalence class of isometric pairs (M, g) rather than just one of its representatives.

Examples 4.10.

1. Minkowski spacetime $\mathbb{R}^4_1$ is a spacetime.
2. More generally, for every $n \geq 2$, $\mathbb{R}^n_1$ is a natural generalisation of Minkowski spacetime that we shall call *n-dimensional Minkowski spacetime.*
3. Defining on $\mathbb{R}^4$ the metric tensor g with components given by the matrix

$$(g_{ij}) = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & (a \circ u^1)^2 & 0 & 0 \\ 0 & 0 & (a \circ u^1)^2 & 0 \\ 0 & 0 & 0 & (a \circ u^1)^2 \end{pmatrix}$$

   where $a : \mathbb{R} \to (0, +\infty)$ is a smooth function yields a spacetime known as flat Robertson-Walker spacetime. It is of utmost importance in Physics as it is one of the main models describing the isotropic and homogeneous universe of the standard cosmology. The function a describes the relative expansion of the universe and is known as the *cosmic scale factor.*

In complete analogy with what we saw for Minkowski spacetime, curves on a manifold are used to describe the worldlines of particles. In this sense, a (future-directed) timelike curve represents the wordline of a material particle moving, at every point, at a speed lower that the speed of light while a (future-directed) null curve corresponds to motion at the speed of light. Again, spacelike curves correspond to motion at speeds higher that the speed of light, which is physically forbidden. Furthermore, in the same way that free moving particles described (future-directed) timelike lines in Minkowski spacetime, free moving material particles in general spacetimes such as a satellite in orbit around the Earth or a planetary orbit around the Sun follow (future-directed) timelike geodesics. Finally, in the same way that the worldlines of photons in Minkowski spacetime are light rays (null lines), in the general case we have that the worldlines of photons are represented by future-directed null geodesics.

4.2 Causality relations

Let M be a spacetime. In the previous discussion we have established in a consistent way a future and a past direction for every event $p \in M$. Having done so, we can now study for every $p \in M$ which events can causally affect p and which can be causally affected by p, thus determining the causal structure of M. To do so, we shall introduce the so called *causality relations*, that are nothing but a mathematical formalisation of our usual notions of causality.

Up until now we have always assumed curves to be smooth. In the interest of what follows, however, we shall relax the smoothness assumption to piecewise smoothness. This will prove to be technically advantadgeous since in many cases it is easier to construct a piecewise smooth curve with certain properties than a smooth one. This consideration, however, has no further repercussion in our discussion since any piecewise smooth curve can be approximated by a sequence of smooth curves. One can convince oneself of this by inspection, but still we refer the reader to Lemma 4.6.1 in [Kri99] for a detailed proof. In particular, the interpretation of smooth timelike and null curves as the worldlines of physical particles still holds for piecewise smooth timelike and null curves, although the assignment of a causal character in this case will deserve some comment.First, however, let us recall what we mean by a piecewise smooth curve.

Definition 4.11. Let $I = [a, b]$ be an interval on the real line. A map $\gamma : I \to M$ is a *piecewise smooth curve on* M if there is a finite partition $a = t_0 < \cdots < t_k = b$ of I such that $\gamma|_{[t_{i-1},t_i]}$ is a smooth curve for all $i = 1, \ldots, k$.

In order to define the notion of causal piecewise smooth curve we shall require that its tangent vector be causal wherever it is well defined but also that time-orientation not be reversed from one break to another. More precisely, consider a piecewise smooth curve $\gamma : I \to M$ and for every $i = 1, \ldots, k-1$ denote by $\gamma'(t_i^-)$ the tangent vector obtained from $\gamma|_{[t_{i-1},t_i]}$ and by $\gamma'(t_i^+)$ the tangent vector obtained from $\gamma|_{[t_i,t_{i+1}]}$. Then our requirement may be formalised as follows.

Definition 4.12. A piecewise smooth curve is *timelike* (resp. *null*) if $\gamma'(t)$ is timelike (resp. null) for every $t \in I \setminus \{t_0, \ldots, t_k\}$ and

$$g_{\gamma(t_i)}(\gamma'(t_i^-), \gamma'(t_i^+)) \leq 0, \text{ for every } i = 1, \ldots, k-1.$$

Note how the condition in the definition imposes that $(\gamma'(t_i^+)$ and $(\gamma'(t_i^-)$ belong to the same connected component of $\mathcal{T}_{\gamma(t_i)}$ (in the timelike case) or $\mathcal{N}_{\gamma(t_i)}$ (in the null case). We can now generalise this notion to again include curves whose causal character may vary from timelike to null.

Definition 4.13. A piecewise smooth curve $\gamma : I \to M$ is *causal* if $\gamma'(t)$ is non-spacelike for every $t \in I \setminus \{t_0, \ldots, t_k\}$ and

$$g_{\gamma(t_i)}(\gamma'(t_i^-), \gamma'(t_i^+)) \leq 0,$$

for every $i = 1, \ldots, k-1$. A causal curve is *future-directed* (resp. *past-directed*) if $\gamma'(t)$ is everywhere future-directed (resp. past-directed).

From now on, whenever we say "a curve" it will be implied that we are referring to "a piecewise smooth curve".

Definition 4.14. If a future-directed causal curve $\gamma : I \to M$ satisfies $\lim_{t\to b} \gamma(t) = q$ (resp. $\lim_{t\to a} \gamma(t) = a$), where a, b $(-\infty < a < b < +\infty)$ are the extremes of the interval I, the event q (resp. p) is called the *future* (resp. *past*) *endpoint* of γ. If the same holds for γ past-directed, then q (resp. p) is called the *past* (resp. *future*) *endpoint* of γ.

As observed in [HKM76], if a causal curve $\gamma : I \to M$ has a future or a past endpoint $q \notin \gamma(I)$ one can always find a new causal curve $\gamma' : I \cup \{t_1\} \to M$ such that $\gamma'|_I = \gamma$ and $\gamma'(t_1) = q$, where t_1 is either the upper or lower bound of I. Thus, we shall assume without loss of generality that all causal curves contain both their future and past endpoints, if they have them.

Let us now recall two familiar notions regarding curves on a manifold. Let $p, q, r \in M$ and consider two piecewise smooth curves $\alpha, \beta : [0, 1] \to M$ such that α goes from p to q and β goes from q to r, then the composition $\alpha * \beta$ is defined as the curve obtained by first traversing α and then β:

$$(\alpha * \beta)(t) = \begin{cases} \alpha(2t), & \text{for } 0 \leq t \leq \frac{1}{2}, \\ \beta(2t-1), & \text{for } \frac{1}{2} \leq t \leq 1. \end{cases}$$

Similarly, one can define the inverse of α as the curve $\alpha^{-1} : [0, 1] \to M$ obtained by traversing α in the opposite sense:

$$\alpha^{-1}(t) = \alpha(1-t).$$

In both cases, the result is again a piecewise smooth curve. Regarding the causal character of $\alpha * \beta$, it can be seen using the chain rule that $\alpha * \beta$ has the causal character of α for $0 \leq t < \frac{1}{2}$ and that of β for $\frac{1}{2} \leq t \leq 1$. If in addition α and β agree in their time-orientation, then $\alpha * \beta$ inherits this time-orientation. For example, if α and β are timelike (resp. causal) and future-directed, then $\alpha * \beta$ is timelike (resp. causal) and future-directed. Finally, again by the chain rule, it can be seen that the inverse path preserves the causal character but reverses time orientation. That is, if α is timelike (resp. causal) and future-directed, then α^{-1} is timelike (resp. causal) and past-directed.

We shall now introduce the *chronological* and *causal relations*. Let $p, q \in M$, we say that *p chronologically precedes q* and write $p \ll q$ if there is a future-directed (piecewise smooth) timelike curve connecting p to q. In a similar way, we say that *p causally precedes q* and write $p < q$ if there is a future-directed (piecewise smooth) causal curve connecting p to q. As usual, $p \leqslant q$ will denote that either $p < q$ or $p = q$. The next result follows from the fact that the composition of piecewise smooth timelike (resp. causal) curves is a piecewise smooth timelike (resp. causal) curve with the same time-orientation.

Proposition 4.15. *The relations $\ll$ and $<$ are transitive.*

Remark 4.16. In fact, one can see that if $p \ll q$ then there are infinitely as many $r \in M$ such that $p \ll r \ll q$. The same holds for $<$.

The chronological and causal relations have a natural description in terms of the following sets.

Definition 4.17. For each $p \in M$, we define the *chronological future*, the *causal future* and the *future horismos* of p, respectively, as

1. $I^+(p) := \{q \in M |\, p \ll q\}$,
2. $J^+(p) := \{q \in M |\, p \leqslant q\}$,
3. $E^+(p) := J^+(p) \setminus I^+(p)$,

Remark 4.18. These definitions have duals, in which *future* is replaced by *past*, $+$ is replaced by $-$ and the positions of p and q are reversed in the inequalities. In general, past definitions and results follow from their future versions (and viceversa) just by reversing time-orientation, and are often regarded as self-evident.

Therefore, the chronological future of p consists of all events in M that can be reached from p by the worldine of some material particle. The causal future of p is then the set of all events that can be causally affected by p whereas the causal past of p is the set of all events that can causally affect p.

Example 4.19. For Minkowski spacetime $\mathbb{R}^4_1$, the chronological future of any event $p \in \mathbb{R}^4_1$ is simply the time cone at p. Similarly, the causal future of p is the union of the time cone at p and the null cone at p. Then, the future horismos of p is the null cone at p. That is:

$$I^+(p) = \mathcal{C}_T(p), \qquad J^+(p) = \mathcal{C}_T(p) \cup \mathcal{C}_N(p), \qquad E^+(p) = \mathcal{C}_N(p).$$

The next result follows from the properties of inverse curves discussed above.

Proposition 4.20. *Let $p, q \in M$, then $q \in I^+(p)$ if and only if $p \in I^-(q)$. The same result holds for J and E.*

More generally, one can define the chronological and causal futures and the future horismos of any subset $A \subset M$.

Definition 4.21. We define the *chronological future*, the *causal future* and the *future horismos* of $A \subset M$, respectively, as

1. $I^+(A) := \bigcup_{p \in A} I^+(p)$,
2. $J^+(A) := \bigcup_{p \in A} J^+(p)$,
3. $E^+(A) := J^+(A) \setminus I^+(A)$,

Remark 4.22. Note that the definition implies $I^+(A) \cup A \subset J^+(A)$.

The following result is relevant in order to establish whether two points in a Lorentzian manifold may be connected by a timelike curve, and therefore is relevant in the description of the causal structure. Roughly speaking, it states that it is possible to deform a segment of a causal curve which is not a null geodesic to obtain a timelike curve with the same endpoints. Its proof involves the notion of variation of a smooth curve $\gamma : [a, b] \to M$, namely a map $\theta : [a, b] \times [-\delta, \delta] \to M$ such that $\theta(t, 0) = \gamma(t)$ for all $t \in [a, b]$. Fixing $s_0 \in [-\delta, \delta]$ yields a smooth curve $\theta(\cdot, s_0) : [a, b] \to M$ and therefore one can think of θ as a one-parameter family of curves on M, parametrised by $s \in [-\delta, \delta]$. In the particular case in which the curves $\theta(\cdot, s_0)$ are geodesics for all $s_0 \in [-\delta, \delta]$, this consideration leads to the notion of Jacobi field. Informally, a Jacobi field is a vector field defined on a geodesic curve that describes its deviation with respect to neighbouring geodesics. We shall not carry out this discussion any further as it lies somewhat out of the scope of this work. However, we refer the interested reader to chapters 8 and 10 in [O'N83], where the topic is widely discussed. In particular, to Proposition 10.46 for a detailed proof of the next result.

Lemma 4.23. *Let M be a a Lorentzian manifold and $p, q \in M$. If γ is a causal curve from p to q that is not a null geodesic, then there is a timelike curve from p to q arbitrarily close to γ.*

Remark 4.24. There is actually a stronger version of the previous lemma stating that the result still holds for null geodesics provided there is some $r \in \gamma(I)$ such that r and q are *conjugate*, i.e., such that there exists a non-zero Jacobi field on γ that vanishes at r and q.

The following result is a fundamental consequence of Lemma 4.23.

Corollary 4.25. *For every $p, q, r \in M$,*

$$\left.\begin{array}{r} p \ll q,\ q \leqslant r \\ p \leqslant q,\ q \ll r \end{array}\right\} \Longrightarrow p \ll r.$$

Proof. If $p \ll q$ and $q \leqslant r$ then there is a future-directed timelike curve α from p to q and a future-directed causal curve β from q to r. Therefore, the composition $\alpha * \beta$ is a future-directed causal curve from p to r which is not a null geodesic (even if β is) and so by Lemma 4.23 there exists a (future-directed) timelike curve from p to r. The other case follows analogously. □

Remark 4.26. The previous result expressed in terms of the chronological and causal sets of a subset $A \subset M$ together with Remark 4.16 show that

$$I^+(A) = I^+(I^+(A)) = I^+(J^+(A)) = J^+(I^+(A)) \subset J^+(J^+(A)) = J^+(A).$$

Let U be an open set of a time-oriented spacetime M, then U is a 4-dimensional Lorentzian manifold of its own. Of course, U is also connected and time-oriented and thus U may be regarded itself as a spacetime. If $A \subset U$, it then makes sense to consider the chronological and causal futures of A, thought of as a subset of the spacetime U. We will denote such sets by $I^+(A, U)$ and $J^+(A, U)$.

A particularly interesting case is that of considering an open convex subset $C \subset M$. Then C is a normal neighbourhood of each of its points, and therefore for every $p \in C$ there exists an open starshaped neighbourhood $V \subset T_pM$ with $exp_p : V \to C$ acting as a local diffeomorphism. Since $V \subset T_pM \cong \mathbb{R}^4_1$ and by Example 4.19, the description in terms of normal coordinates of the chronological and casual futures (and pasts) of a point $p \in C$ is essentially inherited from the coordinate description of its time and causal cones in the tangent space. More precisely,

Proposition 4.27. *Let $\{C; x^1, \dots, x^n\}$ be a normal coordinate system at $p \in C$. Then*

$$I^+(p, C) = \{q \in C \mid -(x^1(q))^2 + (x^2(q))^2 + (x^3(q))^2 + (x^4(q))^2 < 0,\ x^1(q) > 0\}.$$

$$J^+(p, C) = \{q \in C \mid -(x^1(q))^2 + (x^2(q))^2 + (x^3(q))^2 + (x^4(q))^2 \leq 0,\ x^1(q) \geq 0\}.$$

The same holds for $I^-(p, C)$ and $J^-(p, C)$ just by inverting the inequalities on $x^1(q)$.

Note that the inequalities on $x^1(q)$ make sense assuming that the tangent Minkowski space has an admissible basis, in the sense of Definition 2.29. So, basically, we then have that the causal structure of the (local) spacetime C is exactly that of Minkowski spacetime. In particular:

Corollary 4.28. *Let C be an open convex subset of a spacetime M. For every $p, q \in C$, $p \neq q$, let $\gamma_{p,q} : [0, 1] \to C$ be the only geodesic on C from p to q. Then,*

1. *$q \in I^+(p, C)$ if and only if $\gamma'_{p,q}(0) \in T_pM$ is timelike.*
2. *$q \in J^+(p, C)$ if and only if $\gamma'_{p,q}(0) \in T_pM$ is causal.*
3. *$I^+(p, C)$ is open in C (and hence in M).*
4. *$J^+(p, C)$ is the closure in C of $I^+(p, C)$.*

Only the third of the previous statements holds for arbitrary spacetimes. In fact, a stronger result holds:

Lemma 4.29. *The chronological relation $\ll$ is open, i.e., for every $p, q \in M$ such that $p \ll q$ there are open neighbourhoods U of p and V of q such that $p' \ll q'$ for every $p' \in U$ and $q' \in V$.*

Proof. Let γ be a future-directed tiemlike curve from p to q. Let C_p and C_q be convex neighbourhoods of p and q, respectively. Let $p^+ \in C_p$ be a point laying on γ after p and let $q^- \in C_q$ be a point laying on γ after p^+ and before q. Then, the sets $U = I^-(p^+, C)$ and $V = I^+(q^-, C)$ are open by Corollary 4.28 and satisfy the required condition. Indeed, if $p' \in I^-(p^+, C)$ and $q' \in I^+(q^-, C)$ then there are future-directed timelike curves α and β from p' to p^+ and from q^- to q', respectively. The composition $\alpha * \gamma * \beta$ is then a future-directed timelike curve from p' to q'. $\square$

The previous result has the next fundamental corollary.

Corollary 4.30. *For every $p \in M$, the set $I^+(p)$ is open in M.*

Note how this results links the causal structure of M to its topology. Taking into account that $I^+(A) = \bigcup\limits_{p \in A} I^+(p)$ we obtain a more general result.

Corollary 4.31. *$I^+(A)$ is open for every $A \subset M$.*

Remark 4.32. Note that, in general, $J^+(p)$ is not necessarily closed. To see this, consider the spacetime $M = \mathbb{R}^4_1 \setminus \{q\}$, that is, Minkowski spacetime with a point removed. As shown in the figure, then the dashed line is part of the boundary of $J^+(p)$ but is not contained in $J^+(p)$.

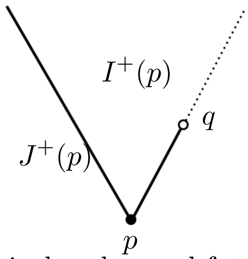

Figure 4: Chronological and causal futures of $p \in \mathbb{R}^4_1 \setminus \{q\}$.

Finally, the next result further shows the underlying connection between causality and topology in M. Recall that, given a topological space X, we denote its interior by $\mathrm{Int}(X)$ and its boundary by ∂X.

Proposition 4.33. *For any subset $A \subset M$,*

1. $\mathrm{Int}(J^+(A)) = I^+(A)$.
2. $J^+(A) \subset \overline{I^+(A)}$
3. $\overline{J^+(A)} = \overline{I^+(A)}$
4. $\partial J^+(A) = \partial I^+(A)$.

Proof. (1) Since $I^+(A)$ is open and $I^+(A) \subset J^+(A)$, we have that $I^+(A) \subset J^+(A)$. For the other inclusion, if $q \in \mathrm{Int}(J^+(A))$, then for a convex neighbourhood C of q we have that $I^-(q, C)$ contains some point in $J^+(A)$. Therefore, $q \in I^+(J^+(A)) \subset I^+(A)$, using Remark 4.26.

(2) It is enough to prove the result for a single point p. Let $q \in J^+(p)$ and note that since $p \in \overline{I^+(p)}$ we can assume $q > p$. Then there is a future-directed timelike curve γ from p to q. Let C be a convex neighbourhood of q and take $q^- \in J^-(q, C)$ a point on γ. Now, by Corollary 4.28, we have $J^+(q^-, C) \subset \overline{I^+(q^-, C)}$. Using Remark 4.26, we have $I^+(q^-, C) \subset I^+(J^+(p)) \subset I^+(p)$, and so we obtain that $q \in \overline{I^+(p)}$.

(3) The inclusion $\supset$ follows from $I^+(A) \subset J^+(A)$. The other one is obtained by using (2) and the fact that $\overline{I^+(A)}$ is closed and the closure of $J^+(A)$ is the smallest closed subject on M containing $J^+(A)$.

(4) The last assertion follows directly from (1) and (3) by using that $I^+(A) = \mathrm{Int}(I^+(A))$ since it is open.

$\square$

The motivation to further investigate on the connection between the casual and the topological structures of a spacetime has led to the definition of new topologies on spacetimes, that we shall briefly comment in Section 5.

4.3 Causality conditions

As we suggested earlier, the sole requirement of time-orientability for a 4-dimensional Lorentzian manifold M is not enough to exclude pathological causal behaviours. For instance, even if M is time-oriented, nothing prevents the existence of closed future-directed timelike curves on it. If this were the case, then the physical realisation of such a spacetime would include the possibility of time-traveling to the past under certain conditions. Of course, this could in turn lead to all sorts of logical paradoxes (the "grandfather paradox", for instance) with strong philosophical consequences.

In this section, we study the different conditions regarding the causal features of spacetimes that one may require in order to prevent non-physical behaviours. These are known as the causality conditions and play an important role in the study of the global properties of spacetimes. They are essential, for example, in the formulation of the so-called Singularity Theorems ([HE73], Chapter 8), that determine the conditions under which spacetime singularities may arise. We shall introduce some of this causality conditions, from less to more restrictive, and see how they naturally establish a causal hierarchy that somehow measures how "physical" a spacetime is. A very thorough review of this topic is given in [MS06], which is the main reference for this section, together with [HE73].

Definition 4.34. A spacetime M is called *non-totally vicious* if $p \not\ll p$ for some $p \in M$.

Note how in spacetimes not satisfying this condition (which we call *totally vicious spacetimes*) the chronological relation $\ll$ is reflexive, since we have $p \ll p$ for all $p \in M$. One could think that totally vicious spacetimes are only of geometrical interest and from a pedagogical point of view. However, one finds relativistic examples of totally vicious spacetimes (relativistic here meaning "that are a solution to the Einstein field equations"). The most paradigmatic one is the so-called *Gödel spacetime*, an exact solution to the Einstein field equations proposed by K. Gödel ([G4̈9]).

Definition 4.35. The *chronology* (resp. *causality*) *condition* is said to hold on M if it has no closed timelike (resp. causal) curves. In this case, M is called *chronological* (resp. *causal*).

Note that in a chronological (resp. causal) spacetime M the chronological (resp. causal) relation is anti-reflexive, i.e, for every $p, q \in M$ one has

$$p \ll q \Rightarrow p \neq q \quad (\text{resp. } p < q \Rightarrow p \neq q),$$

and in fact one can use this as an alternative definition.

Physically, the chronology condition prevents the possibility that under certain conditions, an observer (future-directed timelike worldline) could time-travel to the past, but it does not rule out the possibility to communicate with the past by sending light signals (future-directed null geodesics). To exclude this equally pathological case one further requires the causal condition. However, observe that this does not mean that in the hypothetical physical realisation of a non-chronological spacetime one could decide to time-travel to *any* past event *instantaneously.* Indeed, this time travel would still be subject to the physical requirement that $v < c$ and could only take place between the set of events in M that are connected through closed timelike curves.

Definition 4.36. The *chronology* (resp. *causality*) *violating set* of a spacetime M is the set of points in M that lie in the image of some closed timelike (resp. causal) curve on M.

The following result allows to characterise the chronology violating set.

Proposition 4.37. *The chronology violating set of M is the disjoint union of sets of the form $I^+(p) \cap I^-(p)$, for $p \in M$. In particular, the chronology violating set is open in M.*

Proof. If $q \in M$ is in the chronology violating set of M, then there is a closed timelike curve with past and future endpoints at q. Therefore $q \in I^+(q) \cap I^-(q)$.

If $q \in I^+(p) \cap I^-(p)$ for some $p \in M$, then there is a future-directed curve α from p to q and a past-directed curve β from p to q. Then, the composition $\alpha * \beta^{-1}$ is a closed timelike curve passing through q and hence q is in the chronology violating set of M. Finally, if there is some $r \in M$ such that

$$q \in \left(I^+(p) \cap I^-(p)\right) \cap \left(I^+(r) \cap I^-(r)\right),$$

then p, q and r can all be joined by a closed timelike curve and we have

$$I^+(p) \cap I^-(p) = I^+(r) \cap I^-(r).$$

The fact that the chronology violating set is open follows from the fact that $I^+(p)$ is open. □

This result tells us more about the hypothetical time travel that could take place in a non-chronological spacetime M. Imagine that at some point an observer's worldline met a closed timelike curve γ and that he or she entered this "causal loop". Then, the fact that the chronology violating set is open implies that such an observer would be free to deviate from the closed timelike curve's trajectory, at least in its immediate surroundings.

Proposition 4.38. *If M is compact, then the chronology violating set of M is non-empty.*

Proof. Let $q \in M$ and consider an open convex neighbourhood C of q. Then it is clear that there exists some $p \in C$ such that $q \in I^+(p, C)$. Therefore, $q \in I^+(p)$ and we have that the collection $\{I^+(p)\}_{p \in M}$ is an open cover of M. Since M is compact, it admits a finite subcover $\{I^+(p_1), \dots, I^+(p_k)\}$. We can assume without loss of generality that $I^+(p_1)$ is not contained in any other $I^+(p_j)$ for $j = 2, \dots, k$ (otherwise discard $I^+(p_1)$). But this means that $p_1 \notin I^+(p_j)$, otherwise we would have $I^+(p_1) \subset I^+(p_j)$. Therefore $p_1 \in I^+(p_1)$, which means that there is a closed timelike curve passing through p whose points are in the chronology violating set of M. □

This result suggests that the physical spacetime is not compact. The next result is a characterisation of the causality violating set, completely analogous to Proposition 4.37.

Proposition 4.39. *The causality violating set of M is the disjoint union of the form $J^+(p) \cap J^-(p)$, for $p \in M$.*

Corollary 4.40. *If M is chronological but not causal, then it admits a closed null geodesic.*

Proof. Let $p \in M$ such that the causal condition is violated in M and consider a closed causal curve γ through p. If γ were not a null geodesic, then by Lemma 4.23 one would obtain $p \ll p$, against the chronological condition assumption. $\square$

We have seen that the chronological and causal conditions rule out the possibility to have closed causal curves. At this point, it would seem reasonable to require as well that no causal curve returned arbitrarily close to its point of origin. Or than no causal curve passed arbitrarily close to some other causal curve that then passed arbitrarily close to the origin of the first one. One already sees that this restriction can be pushed to an arbitrary degree of contact resulting in different conditions. We shall only introduce the two first cases.

Definition 4.41. The *future* (resp. *past*) *distinguishing condition* is said to hold at $p \in M$ if every neighbourhood U of p contains a neighbourhood $V \subset U$ of p which no future-directed (resp. past-directed) causal curve starting at p intersects more than once. A spacetime M is *future* (resp. *past*) *distinguishing* if the *future* (resp. *past*) *distinguishing condition* holds for every $p \in M$. Finally, a spacetime M that is both future and past distinguishing is said to be *distinguishing.*

The first definition equivalently states that for every future-directed (resp. past-directed) causal curve $\gamma : [a, b] \to M$ with $\gamma(a) = p$ and $\gamma(b) \in V$, then γ is entirely contained in V. The future (resp. past) distinguishing conditions for a spacetime M have a natural characterisation in terms of the chronological future (resp. past) of its points, namely:

$$\begin{aligned} M \text{ is past distinguishing} \quad &\iff \quad I^-(p) = I^-(q) \Rightarrow p = q, \text{ for all } p, q \in M. \\ M \text{ is future distinguishing} \quad &\iff \quad I^+(p) = I^+(q) \Rightarrow p = q, \text{ for all } p, q \in M. \end{aligned}$$

Definition 4.42. A *causality neighbourhood* D of a point $p \in M$ is a neighbourhood of p such that for every causal curve $\gamma : I \to M$, the preimage $\gamma^{-1}(D)$ is connected.

It is worth noting that this is a stronger condition than requiring the connectedness of $\gamma(I) \cap D$, as this last case would include closed causal curves.

Definition 4.43. The *strong causality condition* is said to hold at $p \in M$ if it admits a neighbourhood basis of causality neighbourhoods. A spacetime M is *strongly causal* it the strong causality condition holds at every $p \in M$.

Even strong causality still admits some non desirable causal behaviours. For instance, in order to have more physically realistic situations, one would aim to have a spacetime for which causality conditions were preserved under small perturbations of the metric tensor. For example, one would like to exclude the possibility of having strongly causal spacetimes for which a slight variation of the metric could alter the initial causal structure as to introduce a closed causal curve. This property is known as the stable causality condition. Its formal definition is intuitive, but technically complicated as it involves the definition of a topology on the set of all Lorentz metrics on a given manifold, called the C^0 *open topology.* Once this is done (see [HE73], Chapter 6), the following definition makes sense.

Definition 4.44. The *stable causality condition* holds on (M, g) if g has an open neighbourhood U in the C^0 open topology such that for every $g' \in U$, the spacetime (M, g') is causal. A spacetime M is *stably causal* if the stable causal condition holds on M.

The last causality condition that we want to comment on is that of global hyperbolicity. The motivations for this definition are beyond the scope of this work, but we would like to include it anyway as it has a very simple form in terms of elements that have been just discussed.

Definition 4.45. The *global hyperbolicity condition* is said to hold in M if M is strongly causal and for every $p, q \in M$ the set $J^+(p) \cap J^-(q)$ is compact. A spacetime M is globally hyperbolic if the global hyperbolicity condition holds on M.

As we anticipated, the causality relations have been presented from the least to the most restrictive. Therefore, the following chain of implication holds and is usually referred to as the *causal ladder*.

$$\begin{gathered}
\text{Globally hyperbolic} \\
\Downarrow \\
\text{Stably causal} \\
\Downarrow \\
\text{Strongly causal} \\
\Downarrow \\
\text{Distinguishing} \\
\Downarrow \\
\text{Causal} \\
\Downarrow \\
\text{Chronological} \\
\Downarrow \\
\text{Non-totally vicious}
\end{gathered}$$

5 Topologies on spacetimes

In our previous discussions we have always assumed spacetimes to have the topology defining its smooth structure, which we shall from now on refer to as the *manifold topology* $\mathcal{T}$. Of course, there was no reason to assume otherwise. However, there does not seem to be any physical motivation for the consideration of such a topology, and indeed it is basically lacked of any physical meaning.

This realisation motivated the investigation of alternative topologies on spacetimes. The possibility to define a topology using the chronological future sets was first pointed out by A. D. Alexandrov in [Ale59]. This idea was further developed by E. H. Kronheimer and R. Penrose in [KP67] and led to the notion of *Alexandrov topology*. However, it was E. C. Zeeman the first to offer a complete description of a new topology having very appealing physical features: in [Zee66], he defined the *Fine topology* for Minkowski spacetime. In his paper, he already suggested that this topology could have a very natural generalisation for arbitrary spacetimes, which was then provided by R. Göbel ([Gö76]) in what he called the *Zeeman topologies*. These topologies, although already of profound physical meaning, were rather complex to deal with. This led S. W. Hawking, A. R. King and P. J McCarthy to the definition of the so-called *Path topology* ([HKM76]), which is found to be much more manageable from a mathematical point of view and offers other important improvements. Then, some years later D. T. Fullwood combined the ideas of Hawking, King and McCarthy with those of Alexandrov and proposed ([Ful92]) a new topology that is physically appealing and quite simple (like the Path topology) and that can be obtained from the causal structure only (like the Alexandrov topology).

In this last section, we would like to give a general overview of some of these topologies, briefly commenting on their main properties and how they are related. Incidentally, this will provide an original example in which the causality relations and conditions play an important role. Surprisingly enough, there is no much literature offering a review on this topic ([Guc11] is the only one that we are aware of). It is also our goal to contribute to remedy this fact.

5.1 The Fine topology

The manifold topology $\mathcal{T}$ in Minkowski spacetime is simply the 4-dimensional Euclidean topology, and so we shall denote it by $\mathcal{E}$. Recall that, in general, the n-dimensional Euclidean topology on $\mathbb{R}^n$ is the topology generated by the Euclidean balls or ϵ-neighbourhoods

$$B_\epsilon(x) = \{y \in \mathbb{R}^n \,|\, d(x,y) < \epsilon\},$$

for some $\epsilon > 0$, where d is the usual n-dimensional Euclidean metric defined by

$$d(x,y) = \sqrt{(x^1 - y^1)^2 + \cdots + (x^n - y^n)^2}.$$

However, the choice of this particular topology, although very natural, seems to lack any physical meaning. For instance, the topology $\mathcal{E}$ is locally homogeneous, whereas $\mathcal{M}$ is not, thus ignoring any difference between space and time and ultimately preventing the possibility to deduce the causal structure from $\mathcal{E}$. Furthermore, the group of all homeomorphisms of $\mathcal{E}$ is vast and of no physical significance.

Definition 5.1. The *Fine topology* $\mathcal{F}$ on $\mathcal{M}$ is the finest topology on $\mathcal{M}$ to induce the 1-dimensional Euclidean topology on every timelike line and the 3-dimensional Euclidean topology on every spacelike hyperplane.

In order to avoid confusion, we shall denote by $\mathcal{M}^{\mathcal{E}}$ Minkowski spacetime endowed with the Euclidean topology and by $\mathcal{M}^{\mathcal{F}}$ Minkowski spacetime endowed with the fine topology. The following result is just an equivalent formulation of the definition of $\mathcal{F}$.

Proposition 5.2. *A subset $U \subset \mathcal{M}$ is $\mathcal{F}$-open if and only if $U \cap \tau$ is $\mathcal{E}_1$-open and $U \cap \Sigma$ is $\mathcal{E}_3$-open for every timelike line τ and spacelike hyperplane Σ.*

It is obvious that $\mathcal{E}$ satisfies the condition of the proposition, thus showing that $\mathcal{F}$ is finer than $\mathcal{E}$. Moreover:

Proposition 5.3. *The topology $\mathcal{F}$ is strictly finer than $\mathcal{E}$.*

Proof. Let $x \in \mathcal{M}$ and $\epsilon > 0$ and consider the set

$$B_\epsilon^{\mathcal{F}}(x) := (B_\epsilon(x) \setminus \mathcal{C}_N(x)) \cup \{x\}.$$

If we denote by A any timelike line or spacelike hyperplane, we have that

$$B_\epsilon^{\mathcal{F}}(x) \cap A = \begin{cases} B_\epsilon(x) \cap A, & \text{if } x \in A, \\ (B_\epsilon(x) \setminus \mathcal{C}_T(x)) \cap A, & \text{if } x \notin A. \end{cases}$$

Now, since $B_\epsilon(x)$ and $Cc_T(x)$ are $\mathcal{E}$-open, both the right-hand sides are either $\mathcal{E}_1$-open or $\mathcal{E}_3$-open depending on whether A is a timelike line or a spacelike hyperplane. Therefore, $B_\epsilon^{\mathcal{F}}(x)$ is open. But $x \in B_\epsilon^{\mathcal{F}}(x)$ does not admit any Euclidean neighbourhood in $B_\epsilon^{\mathcal{F}}(x)$, thus showing that $B_\epsilon^{\mathcal{F}}(x)$ is not $\mathcal{E}$-open. □

The sets $B_\epsilon^{\mathcal{F}}(x)$ defined in the proof are called the *Fine ϵ-neighbourhoods*. Now, it is well known that the Euclidean ϵ-neighbourhoods $B_\epsilon(x)$ form a local basis of neighbourhoods at every point $x \in \mathcal{M}^{\mathcal{E}}$, from which one can obtain a countable basis $\{B_{1/n}(x)\}_{n\geq 1}$ of neighbourhoods for every $x \in \mathcal{M}^{\mathcal{E}}$, showing that $\mathcal{E}$ is first-countable. On the other hand, Zeeman showed that this is not the case for the Fine ϵ-neighbourhoods $B_\epsilon^{\mathcal{F}}(x)$ and that $\mathcal{F}$ is not first-countable. The following is another property of the Fine topology:

Proposition 5.4. *The fine topology induces the discrete topology on every light ray.*

Proof. Consider a light ray λ. For every point $x \in \lambda$, the set $B_\epsilon^{\mathcal{F}}(x) \cap \lambda = \{x\}$ is open in λ. □

The following proposition summarises the main topological properties of $\mathcal{F}$.

Proposition 5.5. *The topological space $\mathcal{M}^{\mathcal{F}}$ is Hausdorff, 2^{nd}-countable, connected and locally connected, but it is not 1^{st}-countable, normal nor locally compact.*

One can already see that such a topology is technically complicated. However, it has important physical advantages. For example, it restricts the notion of continuity only to curves that are physically meaningful. Moreover, the group of homeomorphisms of $\mathcal{M}^{\mathcal{F}}$ is generated by the Lorentz group together with translations and homothecies. This allows to deduce the causal cones from the topology, thus recovering the causal structure of Minkowski spacetime.

5.2 The Path topology

As said earlier, the Fine topology was generalised by Göbel to the spacetimes of GR essentially by replacing "timelike line" by "timelike geodesic" and "spacelike hyperplane" by "spacelike hypersurface". Although physically appealing, these topologies still presented some disadvantages. For example, the group of $\mathcal{F}$-homeomorphisms incorporates homothecies, which are not physically significant. Moreover, there seems to be no physical motivation as to consider spacelike entities (which are non-physical in nature), in the definition of the topologies. All this was pointed out by Hawking, King and McCarthy in [HKM76]. Their approach was then to focus on arbitrary timelike curves (and not only lines or geodesics) and forget about spacelike hypersurfaces.

Definition 5.6. The *Path topology* $\mathcal{P}$ on M is defined to be the finest topology to coincide with the topology induced by $\mathcal{T}$ on every timelike curve.

In particular, the Path topology is finer than the manifold topology and we have the following characterisation.

Proposition 5.7. *A subset $U \subset M$ is $\mathcal{P}$-open if and only if for every timelike curve γ on M there is a $\mathcal{T}$-open set V such that*

$$\gamma \cap U = \gamma \cap V.$$

To illustrate further properties of the Path topology, take $p \in M$, consider an open convex neighbourhood U of p and let us introduce the following sets:

$$C(p,U) := I^+(p,U) \cup I^-(p,U) \quad ; \quad K(p,U) := C(p,U) \cup \{p\}.$$

Then, define also

$$L_U(p,\epsilon) := B_\epsilon(p) \cap K(p,U).$$

Sets of this type form a basis for the topology $\mathcal{P}$ (Theorem 1 in [HKM76]). It can also be shown that sets of the form $K(p,U)$ and $L_U(p,\epsilon)$ are both $\mathcal{P}$-open. Since none of them is $\mathcal{T}$-open (in both cases p does not have any $\mathcal{T}$-neighbourhood), we have:

Proposition 5.8. *The topology $\mathcal{P}$ is strictly finer than $\mathcal{T}$.*

The following result summarises the main topological properties of $\mathcal{P}$.

Proposition 5.9. *The topological space $M^{\mathcal{P}}$ is Hausdorff, connected, locally connected and 1^{st}-countable, but it is not normal nor locally compact.*

The main difference with respect to the Fine and the Zeeman topologies is first-countability. Indeed, this makes the Path topology much easier to deal with than the previous ones. Regarding its physical meaning, the set of $\mathcal{P}$-continuous curves incorporates all timelike paths and hence all possible observers, accelerated or not. Finally, the group of $\mathcal{P}$-homeomorphisms is exactly the group of conformal diffeomorphisms (angle-preseving diffeomorphisms) of (M,g). This means that $\mathcal{P}$ incorporates the causal, differential and smooth (conformal) structure.

5.3 The Alexandrov and Fullwood topologies

The previous topologies all rely on the underlying manifold topology in their definition. The possibility to recover the causal structure from the topology is indeed a very interesting feature. However, one could think the other way round and investigate whether it is possible to define a topology on a spacetime *from* its causal structure. The last two topologies that we would like to mention follow this approach.

Definition 5.10. The *Alexandrov topology* $\mathcal{A}$ is defined to be the coarsest topology for which the sets $I^+(p) \cap I^-(q)$ are open for all $p, q \in M$.

Since the chronological futures are open in the manifold topology, it follows that the Alexandrov topology is in general coarser. However, under certain conditions, the two topologies may coincide. The following result is proven in [Pen72], Theoremm 4.24.

Proposition 5.11. *For a spacetime M, we have:*

$$M \textit{ is strongly causal} \iff M \textit{ is } \mathcal{A}\textit{-Hausdorff} \iff \mathcal{A} = \mathcal{T}.$$

The Alexandrov topology correctly fulfills the motivation of defining a topology on a spacetime solely from its causal structure without relying on the manifold topology. However, it still presents the same problems that motivated the definition of the Fine topology in the first place. The simultaneous consideration of these two motivations and the combination of the ideas of Alexandrov and of Hawking, King and McCarthy resulted in the definition of a new topology, by D. T. Fullwood. In order to define it, let us put

$$I(p, q) := I^+(p) \cap I^-(q).$$

Then, the Fullwood topology can be defined as follows.

Definition 5.12. The *Fullwood topology* $\tilde{\mathcal{P}}$ on a spacetime M is the topology generated by the sets

$$I(p, q) \cup \{q\} \cup I(q, r), \qquad \text{for all } p, q, r \in M.$$

This topology incorporates all the physical significance of the Path topology and has the further appealing feature of being defined only in terms of the spacetime's causal structure. In [Ful92], a proof is presented for the next result, that implies that in distinguishing spacetimes $\tilde{\mathcal{P}}$ relates to $\mathcal{P}$ as in strongly causal spacetimes $\mathcal{A}$ relates to $\mathcal{T}$.

Proposition 5.13. *A spacetime M is distinguishing if and only if $\tilde{\mathcal{P}} = \mathcal{P}$.*

References

[Ale59] A. D. Alexandrov, *The philosophical content and meaning of relativity*, Voprosy Filosofii (The Problems of Philosophy) (1959), no. 1, 67–81.

[Ebe09] F. Eberhardt, *Introduction to the epistemology of causation*, Philosophy Compass **4** (2009), no. 6, 913–925.

[Ful92] D. T. Fullwood, *A new topology on space-time*, Journal of Mathematical Physics **33** (1992), no. 6, 2232–2241.

[G4̈9] K. Gödel, *An example of a new type of cosmological solutions of Einstein's field equations of gravitation*, Rev. Mod. Phys. **21** (1949), 447–450.

[Gol67] A. I. Goldman, *A causal theory of knowing*, Journal of Philosophy **64** (1967), no. 12, 357–372.

[GPS05] A. García-Parrado and J. M. M. Senovilla, *Causal structures and causal boundaries*, Classical and Quantum Gravity **22** (2005), no. 9, R1–R84.

[Guc11] G. Guccione, *Shaping a spacetime from causal structure*, Master thesis (2011).

[Gö76] R. Göbel, *Zeeman topologies on space-times of general relativity theory*, Comm. Math. Phys. **46** (1976), no. 3, 289–307.

[HE73] S. W. Hawking and G. F. R. Ellis, *The large scale structure of space-time*, Cambridge University Press, London-New York, 1973, Cambridge Monographs on Mathematical Physics, No. 1.

[HKM76] S. W. Hawking, A. R. King, and P. J. McCarthy, *A new topology for curved space-time which incorporates the causal, differential, and conformal structures*, J. Mathematical Phys. **17** (1976), no. 2, 174–181.

[KP67] E. H. Kronheimer and R. Penrose, *On the structure of causal spaces*, Mathematical Proceedings of the Cambridge Philosophical Society **63** (1967), no. 2, 481–501.

[Kri99] M. Kriele, *Spacetime*, Lecture Notes in Physics. New Series m: Monographs, vol. 59, Springer-Verlag, Berlin, 1999, Foundations of general relativity and differential geometry.

[MS06] E. Minguzzi and M. Sanchez, *The Causal hierarchy of spacetimes*, 9 2006, pp. 299–358.

[Nab12] G. L. Naber, *The geometry of Minkowski spacetime*, Applied Mathematical Sciences, vol. 92, Springer, New York, 2012, An introduction to the mathematics of the special theory of relativity.

[Nak90] M. Nakahara, *Geometry, topology and physics*, Graduate Student Series in Physics, Adam Hilger, Ltd., Bristol, 1990.

[O'N83] B. O'Neill, *Semi-Riemannian geometry*, Pure and Applied Mathematics, vol. 103, Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York, 1983, With applications to relativity.

[Pen72] R. Penrose, *Techniques of differential topology in relativity*, Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1972.

[SJ14] R V Saraykar and Sujatha Janardhan, *Causal and topological aspects in special and general theory of relativity*, 2014.

[SW77] R. K. Sachs and H. H. Wu, *General relativity for mathematicians*, Springer-Verlag, New York-Heidelberg, 1977, Graduate Texts in Mathematics, Vol. 48.

[Zee66] E. C. Zeeman, *The topology of Minkowski space*, Topology **6** (1966), 161–170.