Mathematical Special Relativity

Ofelia Norgren
Abstract

When Albert Einstein released his theory of special relativity in 1905 it cleared up much of the confusion concerning the properties of light which appeared to be incompatible with Newtonian Physics. Formulated in the language of kinematics, special relativity postulated that space and time had to be merged together in a 4-dimensional object, the so-called spacetime, with a clear distinction between space and time. In 1908, Hermann Minkowski, Einstein’s former mathematics professor, presented a geometric interpretation of this merging, a geometry that was subsequently named Minkowski spacetime. The study of this geometry is in fact what special relativity is, in its simplest form.

In this degree project we present the mathematical foundation of special relativity, focusing on both classical and modern aspects of the subject. In particular, we present a detailed study of the geometry of Minkowski spacetime and show how well-known relativistic effects such as the constancy of the speed of light and the twin paradox can be explained in terms of this geometry. In the final part of this project we discuss a question motivated by problems in modern cosmology: how to define a valid notion of a distance for Minkowski spacetime which does not carry a positive definite inner product?
## Contents

1 Introduction .......... 1

2 Minkowski Spacetime .......... 2
   2.1 The Inner Product .......... 2
   2.2 Pythagoras’s Theorem .......... 3
   2.3 Basic Causality Theory .......... 4
   2.4 Basic Theory of Curves in $\mathbb{R}^4_1$ .......... 6
   2.5 Lorentz Transformations .......... 7

3 Relativistic Effects .......... 11
   3.1 The Speed of Light .......... 11
   3.2 The Twin Paradox .......... 13
   3.3 Constant Acceleration .......... 14

4 Minkowski Spacetime as a Metric Space .......... 15
   4.1 Time Functions and Null Distances .......... 16
   4.2 Examples .......... 17

5 Conclusion and Discussion .......... 23

References .......... 24
1 Introduction

Light and its properties have long been somewhat of a mystery. The question to what light actually is gives a very confusing answer, an answer that is both intricate and easy, in other words a very paradoxical answer. Scientists had conjectured through history that the elements of light are either waves or particles but later it was shown that light behaves as both waves and particles at the same time. This wave-particle duality, and, more generally, the uncertainty of the right way to treat light has through history caused some dilemmas and contradictions [1].

In 1887, Albert A. Michelson and Edward Morley conducted an optical experiment that would change the view of the properties of light. The classical Michelson-Morley, or aether-drift, experiment was based on the idea that there existed the so-called luminiferous aether, a medium that filled the universe. If light was to be treated as a wave it should travel through this medium, according to our knowledge of wave motion. In fact, the goal of the Michelson-Morley experiment was to confirm the existence of the luminiferous aether, a goal that numerous experiments carried out during the 19th century had failed to achieve. The idea of Michelson and Morley was to detect the relative motion of the aether by measuring the speed of light in different directions. However, to everyone’s surprise, the experiment showed that the speed of light was constant in all directions which led to the conclusion that there was no relative motion between the aether and the earth [1], [2]. This finding was not given any clear explanation back in 1887 and it continued to puzzle physicists for a few years to come. It was not until Einstein worked out the details of the Michelson-Morley experiment using them to create his theory of special relativity that this mystery was cleared up [3].

In fact, all these uncertainties concerning light and its dual properties caused large problems for Newtonian Physics, reaching an all time high at the end of the 19th century. As it turned out later, the problem of the Newtonian approach to light was that it was treated as relative and that its speed was considered to be infinite. In other words, it was thought that there was no "speed limit" meaning that the speed of light could (in theory) increase infinitely. At the same time, the Michelson-Morley experiment seemed to indicate that the properties of light did not depend on the observer, in particular its speed always appeared to be the same. Albert Einstein was the first to realize that the light should be treated as an absolute quantity with the constant speed $c$ same for all observers. These ideas have subsequently
formed the very basis for the theory of special relativity that Einstein published in 1905 [4], [5].

To express these ideas mathematically Einstein treated space and time quite differently from Newtonian mechanics, combining them together into a single object called spacetime. After the mathematical properties of this object were described by Hermann Minkowski in 1908, the spacetime was fittingly named Minkowski spacetime [4], [5].

2 Minkowski Spacetime

Roughly speaking, Minkowski spacetime is a geometry needed to explain the physical universe from the perspective of special relativity. As we will see in this section, a few characteristics of special relativity that cannot be explained in terms of Newtonian physics follow rather effortlessly from the geometry of Minkowski spacetime. We will mostly follow [5] and [6].

2.1 The Inner Product

The coordinate change which Einstein employed to account for the properties of light (see Section 2.5 below), arises naturally when 3-dimensional space and 1-dimensional time are combined into a certain 4-dimensional object, the so-called Minkowski spacetime denoted by $\mathbb{R}^4_1$ [5]. Mathematically, $\mathbb{R}^4_1$ is $\mathbb{R}^4$ equipped with the so-called Lorentzian inner product, a symmetric bilinear form $\eta$ defined as follows. Let $e = \{e_1, e_2, e_3, e_4\}$ be an orthonormal basis of $\mathbb{R}^4$ equipped with the standard Euclidean inner product $\langle \cdot, \cdot \rangle$, and let $v = v^ae_a$ and $w = w^ae_a$ be the coordinate expressions of $v, w \in \mathbb{R}^4$ in this basis. Then

$$\eta(v, w) = v^1w^1 + v^2w^2 + v^3w^3 - v^4w^4.$$  \hspace{1cm} (1)

As we can see, $\eta$ differs from $\langle \cdot, \cdot \rangle$, which in this case would have the expression

$$\langle v, w \rangle = v^1w^1 + v^2w^2 + v^3w^3 + v^4w^4.$$  \hspace{1cm} (2)

This difference reflects the fact that the direction given by $e_4$ is thought of as time whereas the linear subspace spanned by $\{e_1, e_2, e_3\}$ is thought of as space. Note that the geometric structure of spacetime is coordinated with a choice of an orthonormal basis, $e$, which in this case is referred to as frame of reference [6].
2.2 Pythagoras’s Theorem

As a preliminary to our investigation of the geometry of Minkowski space-time, we take a closer look at the well-known Pythagorean theorem following [7]. One possible formulation of this theorem states that the length of a two-dimensional vector with components \((dx, dy)\) is given by

\[
ds^2 = dx^2 + dy^2
\]

as illustrated in the following picture:

This result generalizes to higher dimensions. For example, in dimension 3 we have

\[
ds^2 = dx^2 + dy^2 + dz^2,
\]

for the vector \((dx, dy, dz)\) as in the picture:
Similarly, in dimension 4 Pythagoras’s theorem takes the form

\[ ds^2 = dx^2 + dy^2 + dz^2 + dt^2. \] (5)

In the view of (2) as opposed to (1), the generalization of (5) to Minkowski spacetime is

\[ ds^2 = dx^2 + dy^2 + dz^2 - dt^2. \] (6)

This formula reflects the main difference between the geometry of 4-dimensional Euclidean space and the geometry of Minkowski spacetime: the roles of space and time are different in their unification in spacetime. In fact, one can think of Minkowski spacetime as of a modification of the Euclidean geometry, where one has ”flipped the sign” in the Pythagorean theorem. This ”sign flip” results in a new geometry that turns out to be different from Euclidean geometry in many respects. The study of this geometry is special relativity in its simplest form [7].

### 2.3 Basic Causality Theory

While Minkowski spacetime can be thought of as our universe ”for all times”, a point \((x, y, z, t)\) in Minkowski spacetime is used to model an event, an instant incidence with no inflation through space and with no time span.
As explained in Section 2.2, a 4-vector defined by its four components

\[ \vec{ds} = (dx, dy, dz, dt) \]  

(7)
can be used to measure the interval between two events by means of the formula

\[ ds^2 = dx^2 + dy^2 + dz^2 - dt^2. \]  

(8)

Depending on whether \( ds^2 \) is zero, positive or negative, vectors in Minkowski spacetime are divided into three types: lightlike, spacelike and timelike. This splitting, referred to as the causal structure of Minkowski spacetime, is illustrated in Figure 1 and is discussed below in detail.

![Figure 1: Illustration of time-, space and lightlike vectors in spacetime.](image)

A vector \( \vec{ds} \) is said to be lightlike if

\[ ds^2 = 0. \]  

(9)
This is equivalent to \( dt^2 = dx^2 + dy^2 + dz^2 \), therefore all lightlike vectors will form the boundary of a solid cone (the so-called light cone) in Minkowski spacetime \( \mathbb{R}_1^4 \) when their starting points are placed in the coordinate origin (see Figure 1). Physical significance of lightlike vectors in special relativity can be summarized as follows: if there is a lightlike vector between two events then these two events can be linked by a light ray.

A vector \( \vec{ds} \) is said to be spacelike if

\[ ds^2 > 0. \]  

(10)
Geometrically, this means that the time component of the vector is smaller than its space component. As a consequence, all spacelike vectors are located outside of the light cone (see Figure 1). If there is a spacelike vector between two events then there exists a reference frame where these two events happen at the same time but at different locations. This will be explained in detail in Section 2.5.

Finally, a vector $\vec{ds}$ is said to be timelike if
\[ ds^2 < 0. \] (11)

In this case the time component is larger than the space component, which implies that all timelike vectors are located inside the light cone (see Figure 1). As we will see in Section 2.5, in this case it is possible to find a reference frame where the respective events happen at the same location but at different times.

We take yet another look at Figure 1. We may think of a point at the tip of the light cone as the event being observed. Removing the tip, the cone splits into two components, and we may choose one of them to represent the future of the event and the other one to represent its past. Such a choice made consistently for all events is called the time orientation. A timelike vector lying in the future (respectively past) part of the light cone is called future (respectively past) pointing. Note that any event located inside the future light cone can be caused (or affected) by the event located at the origin. This and some other aspects of causality theory will be further discussed in Section 2.5.

### 2.4 Basic Theory of Curves in $\mathbb{R}^4_1$

In this section we review some basics concerning curves in $\mathbb{R}^4_1$ that will be needed for our exploration of relativistic effects in Section 3 and time functions and distances in Minkowski spacetime in Section 4.

A curve in $\mathbb{R}^4_1$ is a smooth map
\[ \alpha : I \to \mathbb{R} \quad s \mapsto (x(s), y(s), z(s), t(s)) \] (12)
where $I \subseteq \mathbb{R}$ is a connected nonempty interval. Its velocity vector is computed for $s \in I$ as
\[ \alpha'(s) = (x'(s), y'(s), z'(s), t'(s)). \] (13)
Just like Euclidean space, Minkowski spacetime can be viewed as a semi-Riemannian manifold $\mathcal{M} = (\mathbb{R}^4_1, \gamma)$ whose semi-Riemannian metric $\gamma$ is given at all points by the inner product $\eta$ as described in (1), see e.g. [5]. As a consequence, the Christoffel symbols of $\gamma$ vanish, and the (tangential) acceleration of the curve $\alpha$ is simply

$$\alpha''(s) = (x''(s), y''(s), z''(s), t''(s)).$$

Recall that $\alpha$ is called a geodesic if it has zero acceleration at all points. In the view of (14) we see that the geodesics of Minkowski spacetime are straight lines, just like the geodesics of Euclidean space. However, in contrast to Euclidean case, a straight line segment joining two points in $\mathbb{R}^4_1$ is not the shortest of all curves joining the two points. For example, a straight line segment joining $(0, -1, 0, 0)$ and $(0, 1, 0, 0)$ has length 2 whereas a broken line through $(0, -1, 0, 0)$, $(1, 0, 0, 0)$ and $(0, 1, 0, 0)$ has length 0, as one readily computes using (5). This will become important in Section 4.

Finally, we recall some standard terminology (see e.g. [8]) that will be needed in Section 4. A curve $\alpha = \alpha(s)$ in Minkowski spacetime is called causal if its tangent vector $\alpha'(s)$ given by (13) is either timelike or lightlike for every $s$. A causal curve is called future pointing if its tangent vector is future pointing for all points on the curve.

### 2.5 Lorentz Transformations

As discussed in Section 2.3, events in special relativity are modelled as points of Minkowski spacetime. For describing and localizing events the notion of observer is used, meaning roughly that a choice of a coordinate system has been made [6]. If two observers are in relative motion with constant velocity then the relation between their descriptions of the same event (or, equivalently, the relation between the coordinates of the point representing the event in the respective coordinate systems) is described by a Lorentz transformation. Lorenz transformations are linear transformations which preserve spacetime intervals between points (as defined in Section 2.2). As a consequence, we may represent a Lorentz transformation by the matrix

$$\Lambda = \begin{bmatrix}
\Lambda_1^1 & \Lambda_1^2 & \Lambda_1^3 & \Lambda_1^4 \\
\Lambda_2^1 & \Lambda_2^2 & \Lambda_2^3 & \Lambda_2^4 \\
\Lambda_3^1 & \Lambda_3^2 & \Lambda_3^3 & \Lambda_3^4 \\
\Lambda_4^1 & \Lambda_4^2 & \Lambda_4^3 & \Lambda_4^4
\end{bmatrix}$$

which is required to satisfy

$$\Lambda^T \eta \Lambda = \eta$$

(16)
where $\Lambda^T$ is the transpose of $\Lambda$ and $\eta$ is the matrix of the Minkowskian inner product given by

$$
\eta = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & -1
\end{bmatrix}
$$

(17)

A Lorentz transformation may include a rotation of space, however in special relativity one is mostly interested in ”space rotation free” Lorentz transformation that are called (Lorentz) boosts. Boosts can be thought of as Minkowskian analogues of rotations of Euclidean space where the rotation takes place in planes spanned by one spacelike and one timelike direction such as $xt$-plane. Boosts and their properties become especially important when it comes to describing well-known relativistic effects, see Section 3.

In order to understand boosts in more detail, we focus on the two dimensional case, i.e. we consider the Minkowski plane $\mathbb{R}^2_1$ with the coordinates $(x,t)$ and the inner product $\eta$ given by the matrix

$$
\eta = \begin{bmatrix}
1 & 0 \\
0 & -1
\end{bmatrix}
$$

(18)

(A detailed treatment of the four dimensional case is very similar in spirit and can be found in [6]).

**Theorem 2.1.** Let

$$
\Lambda = \begin{pmatrix}
\Lambda^1_1 & \Lambda^1_2 \\
\Lambda^2_1 & \Lambda^2_2
\end{pmatrix}
$$

be such that $\Lambda^T \eta \Lambda = \eta$ holds, then $\Lambda$ is given by

$$
\Lambda = \pm \left( \frac{1}{\sqrt{1-v^2}} \begin{bmatrix}
\frac{1}{\sqrt{1-v^2}} & \frac{v}{\sqrt{1-v^2}} \\
\frac{v}{\sqrt{1-v^2}} & \frac{1}{\sqrt{1-v^2}}
\end{bmatrix} \right)
$$

(20)

for a constant $v$ such that $|v| < 1$.

**Proof.** We have

$$
\Lambda^T \eta \Lambda = \begin{pmatrix}
\Lambda^1_1^2 - (\Lambda^2_1)^2 & \Lambda^1_2 \Lambda^1_1 - \Lambda^1_2 \Lambda^2_2 \\
\Lambda^2_1 \Lambda^1_2 - \Lambda^2_2 \Lambda^2_1 & (\Lambda^1_2)^2 - (\Lambda^2_2)^2
\end{pmatrix}
$$

(21)
As a consequence, the elements of \( \Lambda \) have to satisfy the relations

\[
\begin{align*}
(\Lambda_1^1)^2 - (\Lambda_1^2)^2 &= 1 \\
\Lambda_2^1 \Lambda_1^1 - \Lambda_1^2 \Lambda_2^2 &= 0 \\
\Lambda_1^1 \Lambda_2^1 - \Lambda_2^2 \Lambda_1^1 &= 0 \\
(\Lambda_2^2)^2 - (\Lambda_1^2)^2 &= -1
\end{align*}
\] (22)

Solving these equations for \( \Lambda_1^1, \Lambda_1^2, \Lambda_2^1 \) and \( \Lambda_2^2 \) we obtain

\[
\begin{align*}
\Lambda_1^1 &= \pm \frac{1}{\sqrt{1-v^2}} \\
\Lambda_2^1 &= \pm \frac{v}{\sqrt{1-v^2}} \\
\Lambda_2^2 &= \pm \frac{v}{\sqrt{1-v^2}} \\
\Lambda_1^2 &= \pm \frac{1}{\sqrt{1-v^2}} 
\end{align*}
\] (23)

for \(|v| < 1\). This gives us \( \Lambda \) as in (20).

The main physical implication of this theorem is the following. Suppose that we have an \( xt \)-coordinate system associated with a stationary observer and an \( x't' \)-coordinate system associated with an observer that moves in the \( x \)-direction with the constant velocity \( v \). If the first observer assigns coordinates \((x, t)\) to an event then in the coordinate system of the second observer the event will have the coordinates

\[
x' = \frac{1}{\sqrt{1-v^2}} (x - vt) \\
t' = \frac{1}{\sqrt{1-v^2}} (t - vx) \] (24)

The parameter \( v \) in this formula is usually referred to as the velocity of the boost.

In the rest of this section we state and prove several important results concerning boosts to be used in Section 3.

**Theorem 2.2.** The inverse transformation to (24) mapping \((x', t') \mapsto (x, t)\) is the boost of velocity \(-v\), given by the matrix

\[
\Lambda^{-1} = \begin{pmatrix}
\frac{1}{\sqrt{1-v^2}} & \frac{v}{\sqrt{1-v^2}} \\
\frac{v}{\sqrt{1-v^2}} & \frac{1}{\sqrt{1-v^2}}
\end{pmatrix}
\] (25)

One can also express the solutions using hyperbolic functions. In this case \( \Lambda \) looks similar to a rotation matrix with \( \sin \) and \( \cos \) replaced by \( \sinh \) and \( \cosh \), see [5] for details.
Proof. A computation.

As a consequence, we get the following expressions for \( x \) and \( t \):

\[
\begin{align*}
  x &= \frac{1}{\sqrt{1-v^2}} (x' + t'v) \\
  t &= \frac{1}{\sqrt{1-v^2}} (t' + x'v)
\end{align*}
\]  

(26)

**Theorem 2.3.** A boost preserves the interval between a point and the origin, i.e. \( x^2 - t^2 = (x')^2 - (t')^2 \). As a consequence, the causal character is preserved under a boost, i.e. spacelike vectors are mapped to spacelike vectors, timelike vectors to timelike vectors, and lightlike vectors to lightlike vectors.

Proof. In fact, the statement of the theorem is a consequence of the relation \( \Lambda^T \eta \Lambda = \eta \). One can also verify it directly, using (26):

\[
\begin{align*}
  x^2 - t^2 &= \frac{1}{1-v^2} (t'v + x')^2 - \frac{1}{1-v^2} (t' + x'v)^2 = \\
  &= \frac{1}{1-v^2} ((t'v + x')^2 - (t' + x'v)^2) = \\
  &= \frac{1}{1-v^2} ((t')^2 v^2 + (x')^2 - (t')^2 - (x')^2 v^2) = \\
  &= \frac{1}{1-v^2} (-v^2((x')^2 - (t')^2) + ((x')^2 - (t')^2)) = \\
  &= \frac{1}{1-v^2} ((1-v^2)((x')^2 - (t')^2)) = \\
  &= (x')^2 - (t')^2.
\end{align*}
\]  

(27)

**Theorem 2.4.** If \( t > |x| \) then \( t' > |x'| \) and if \( t < |x| \) then \( t' < |x'| \). As a consequence, a boost preserves the property of a timelike vector to be future (respectively past) pointing.

Proof. We prove the first statement of the theorem, the second one is proven similarly. Given that \( t > |x| \) we have \( t^2 - x^2 > 0 \), whereas by Theorem 2.3 we know that \( x^2 - t^2 = (x')^2 - (t')^2 \). This gives

\[
(t' - |x'|)(t' + |x'|) > 0.
\]  

(28)

This inequality implies that either \( t' > |x'| \) or \( t' < -|x'| \). To see that the second option is not possible, it suffices to show that \( t' > 0 \). For this we
recall that \( |v| < 1 \) hence

\[
v x \leq |v x| \leq |v| |x| < |x|.
\]

Consequently, we have \(-v x > -|x|\) which implies

\[
t' = \frac{t - v x}{\sqrt{1 - v^2}} > \frac{t - |x|}{\sqrt{1 - v^2}} > 0,
\]

and \( t' > |x'| \) follows.

\[\square\]

3 Relativistic Effects

Many phenomena in special relativity have natural interpretation in terms of the geometry of Minkowski spacetime. In this section we focus on discussing the constancy of the speed of light, the twin paradox, and the existence of the "speed limit" in special relativity following [5] and [7]. Some other relativistic effects are explained in [5].

3.1 The Speed of Light

As mentioned in Section 2.5, the causal structure of Minkowski spacetime is preserved under a boost transformation, in particular, lightlike vectors are mapped to lightlike vectors. We will use this fact to explain why the speed of light is the same for all observers. The example below is adapted from [7].

Consider the particles \( A, B, \) and \( C \) moving in the \( x \)-direction as shown in the picture:
To compute the relative velocities of the particles, we will view each of the particles $A$ and $B$ as an observer by assigning to it a coordinate system in which the particle is at $x = 0$ for all times. As we can see, the chosen coordinate system is the one associated with the particle $A$. From $A$’s perspective, the particle $B$ is moving with the velocity $v = \frac{5}{13}$ and the particle $C$ is moving with the velocity $v = 1$.

The coordinate system associated with the particle $B$ can be constructed by performing the boost of velocity $v = -\frac{5}{13}$ as shown in the picture:

Here we used the results of Section 2.5 to compute the coordinates of the particles $A$ and $C$ in this coordinate system. We find that from $B$’s perspective $A$ has the velocity $v = -\frac{5}{13}$, which is consistent with what was said earlier about the velocity of $B$ as measured by $A$. What is more surprising is that the velocity of $C$ has not changed: measured in the coordinate system associated with $B$ it is still $1$. This is due to the fact that the vector $(18, 18)$ is lightlike and will remain such after any boost: its components will change but their ratio will always be $1$. As a consequence, the speed of light is $c = 1$ for all observers, not only for $A$ and $B$.  

12
3.2 The Twin Paradox

The famous twin paradox is in fact not a paradox but yet another natural consequence of the geometry of Minkowski spacetime. Consider two identical twins, one of which stays on earth while the other travels 13 years away from earth with the velocity \( v = \frac{5}{13} \) and then returns back with the same speed. The time experienced by each twin in his own coordinate system is the length of his worldline, i.e. the path traversed through spacetime. This can be visualized as follows:

\[
2\sqrt{13^2 - \left(\frac{5}{13}\right)^2} = 2\sqrt{13^2 - \frac{5^2}{13^2}} = 24.
\]  

As a consequence, we see that the second twin, who is moving, experiences less time than the first twin, who is at rest, and should therefore technically be younger. Had the second twin moved even faster, with the speed close to the speed of light, he would have experienced almost no time at all. To see this, one computes the length of the black curve representing the light travelling 13 years away from earth and then returning back to obtain

\[
2\sqrt{13^2 - 13^2} = 0.
\]
3.3 Constant Acceleration

By analogy with Euclidean spheres (e.g. $x^2 + y^2 + z^2 + t^2 = 1$ defines the unit sphere in $\mathbb{R}^4$) one can define spheres in Minkowski spacetime. As the interval $x^2 + y^2 + z^2 - t^2$ between $(0,0,0,0)$ and $(x,y,z,t)$ can be either positive, or negative, or zero, we have three types of spheres to consider, as shown in the following picture:

The black cone is the sphere of radius zero, i.e. $x^2 + y^2 + z^2 - t^2 = 0$. The blue two-sheet hyperboloid represents the sphere $x^2 + y^2 + z^2 - t^2 = -1$, and the red one-sheet hyperboloid represents the sphere $x^2 + y^2 + z^2 - t^2 = 1$. The red hyperboloid appears then to be the closest analogue of the unit sphere in Euclidean space.

Now, consider the curve given by the intersection of the red hyperboloid with the plane $y = z = 0$. This curve is the hyperbola that can be parametrized by $\gamma : \tau \mapsto (\cosh h\tau, 0, 0, \sinh h\tau)$. Differentiating we see that $\eta(\gamma', \gamma') = -1$ meaning that the curve represents the motion of a material particle parametrized by its proper time $\tau$. Furthermore, $\eta(\gamma'', \gamma'') = 1$, meaning that the particle is moving with constant acceleration 1. Still, we see that its velocity never exceeds the speed limit set by the speed of light as the curve will never cross the black cone remaining asymptotic to it for large $\tau$. 

\[14\]
4 Minkowski Spacetime as a Metric Space

We begin this section by recalling the definition of a metric space.

**Definition 4.1.** [9] Let $X$ be an arbitrary set. A function $d : X \times X \to \mathbb{R} \cup \{\infty\}$ is a metric on $X$ if the following conditions are satisfied for all $x, y, z \in X$:

1. $d(x, y) > 0$ if $x \neq y$, and $d(x, x) = 0$.
2. $d(x, y) = d(y, x)$.
3. $d(x, z) \leq d(x, y) + d(y, z)$.

A set $X$ equipped with a metric $d$ is called a metric space.

It is well-known that Euclidean space is a natural metric space whose metric is the distance between two given points as defined by Pythagoras’s theorem. In contrast, Minkowski spacetime does not carry any naturally defined metric. In particular, Pythagoras’s theorem cannot be used for defining one as, in particular, the interval between any two points on the light cone is zero, violating (1) in the above definition. See Section 2.2 for more details.

It is important to have a suitable notion of a distance (metric) on Minkowski spacetime as well as on other spacetimes that are important in general relativity and cosmology. This is needed for purposes that require one to compare a given geometry (such as the geometry of our universe) with a model geometry (such as the FLRW model that is used in cosmology to model our universe). In this section we review the basics of the so-called null distance on Minkowski spacetime as defined by Sormani and Vega in [10]. Whether this null distance will ultimately turn out to be a metric or not depends crucially on one’s choice of a time function, as will be explained in Section 4.2 where two examples from [10] are discussed.
4.1 Time Functions and Null Distances

By definition (see e.g. [10]), a generalized time function on \( n \)-dimensional Minkowski spacetime is a function \( \tau : \mathbb{R}^n_1 \to \mathbb{R} \) that is strictly increasing along all future-pointing causal curves.

Given a time function, the notion of null distance can be defined as follows.

**Definition 4.2.** [10] Let \( \tau : \mathbb{R}^n_1 \to \mathbb{R} \) be a generalized time function.

1. Let \( \beta : [a, b] \to \mathbb{R}^n_1 \) be a piecewise causal curve with breaks \( a = s_0 < ... < s_k = b \). We set \( x_i = \beta(s_i) \) and define the null length of \( \beta \) by
   \[
   \hat{L}_\tau(\beta) := \sum_{i=1}^{k} |\tau(x_i) - \tau(x_i - 1)|
   \]

2. For any \( p, q \in M \), we define the null distance by
   \[
   \hat{d}_\tau(p, q) := \inf \{ \hat{L}_\tau(\beta) : \beta \text{ piecewise causal from } p \text{ to } q \}
   \]

It turns out that the null distance \( \hat{d}_\tau(p, q) \) satisfies all the axioms of metric (see Definition 4.1) except for possibly the definiteness axiom, i.e. \( \hat{d}_\tau(p, q) > 0 \) for \( p \neq q \). Furthermore, when \( q \) is in the future of \( p \) the null distance \( \hat{d}_\tau(p, q) \) has a very simple expression. This is summarized in the following result which was proven (in a more general setting) in [10, Lemma 3.10 and Lemma 3.11].

**Lemma 4.1.** For any generalized time function \( \tau \) on \( \mathbb{R}^n_1 \) we have
   \[
   \hat{d}_\tau(p, q) \geq |\tau(q) - \tau(p)| \quad \text{for any } p, q \in \mathbb{R}^n_1.
   \]

Consequently, definiteness can only fail for points in the same time slice:
   \[
   \hat{d}_\tau(p, q) = 0 \Rightarrow \tau(q) = \tau(p).
   \]

Furthermore, if \( q \) is in the causal future \( J^+(p) \) of \( p \) (meaning that the vector \( \vec{pq} \) is either future directed timelike or future directed lightlike) we have
   \[
   \hat{d}_\tau(p, q) = \tau(q) - \tau(p).
   \]
4.2 Examples

In this section we will investigate which properties make a time function "good" or on contrary "bad". A time function \( \tau \) is considered to be "good" if the null distance as defined in Definition 4.2 is a metric (see Definition 4.1) and if it encodes the causal structure of spacetime in a certain sense to be made precise below.

We will now discuss two examples of time functions on Minkowski spacetime from [10] and study the properties of the respective null distances in detail.

Example 1

Our first example is concerned with the "natural" time function \( \tau = t \). In this case the following has been shown in [10, Proposition 3.3].

Proposition 4.2. For the time function \( \tau = t \) on \( \mathbb{R}^n_1 \) we have

\[
\hat{d}_t \text{ is a definite, translation invariant metric on } \mathbb{R}^n_1, \text{ with } \\
\hat{d}_t(p, q) = \begin{cases} 
|t(q) - t(p)| & q \in J^+(p) \\
|q_s - p_s| & q \notin J^+(p)
\end{cases}
\]

where \( J^+(p) \) is the causal future of \( p \) as defined in Lemma 4.1 and \( J^-(p) \) is the causal past of \( p \) defined similarly.

\[
\hat{d}_t \text{ encodes the causality of } \mathbb{R}^n_1 \text{ via } \\
p \leq q \iff \hat{d}_t(p, q) = t(q) - t(p). \tag{34}
\]

\[
\text{The } \hat{d}_t\text{-sphere of radius } r \text{ centered at the origin is the coordinate cylinder given by } \max \{|t|, ||x||\} = r \text{ as shown in the picture:}
\]
Proof. For simplicity we will work in dimension $n = 2$, i.e. on $\mathbb{R}^2_1$. The case of $\mathbb{R}^n_1$ for $n > 2$ is treated similarly by replacing the absolute value $| \cdot |$ by the Euclidean norm $\| \cdot \|$ of in the arguments below.

Clearly, the null distance is translation invariant, so without loss of generality we may assume $p = (0, 0)$. As for $q = (x, t)$, we will assume that $x > 0$ and $t > 0$ as the remaining cases can be treated similarly.

Note that when $q \in J^\pm(p)$ the stated formula for $\hat{d}_t(p, q)$ is available by Lemma 4.1. Consequently, in order to prove (33) we only need to consider the case when $q \notin J^\pm(p)$. In this case the vector $\vec{pq}$ is spacelike and $q = (x, t)$ satisfies $t < x$. Consequently, we can construct a piecewise null curve $\eta$ with a break in $z$, as shown in the picture. Its null length is

$$\hat{L}_t(\eta) = |t(z) - t(p)| + |t(z) - t(q)| = a + b + b = a + 2b = x.$$  

(35)
It follows that $\hat{d}_i(p,q) < x$. Thus, in order to finish the proof of (33) we only have to show that $d_i(p,q) \geq x$ to finish this proof.

For this purpose we consider an arbitrary piecewise causal curve $\beta = \beta_1 \cdot \beta_2 \cdot \ldots \cdot \beta_k$ from $p$ to $q$, i.e. for each segment $\beta_i$, $i = 1,\ldots,k$, either $\beta_i$ or $-\beta_i$ is future directed causal, as shown in the picture:

In the case when $\beta_i$ is future directed we can parametrize it by $\beta_i(t) = (t, x_i(t))$ for $t \in [t_i, t_i + \delta]$ as shown in the picture:
In the case when $\beta_i$ is past directed, the same applies to $-\beta_i$.

Since $\beta_i$ is causal its tangent vector

$$\beta_i' = (1, x_i'(t))$$

is either timelike or null which implies $|x_i'(t)| \leq 1$. It follows that

$$|x_i(t_i + \delta_i) - x_i(t_i)| = \left| \int_{t_i}^{t_i + \delta_i} |x_i'(t)| \, dt \right| \leq \int_{t_i}^{t_i + \delta_i} |x_i'(t)| \, dt \leq \int_{t_i}^{t_i + \delta_i} 1 \, dt = \delta_i.$$  

Consequently, the null length of $\beta$ satisfies

$$\hat{L}_t(\beta) = \sum_{i=1}^{k} \delta_i \geq \sum_{i=1}^{k} (x_i(t_i + \delta_i) - x_i(t_i)) = x.$$  

Taking infimum over all possible piecewise causal $\beta$ from $p$ to $q$ we see that $\hat{d}_t(p, q) \geq x$ which concludes our proof of (33).

The second and the third statement of the proposition are direct consequences of (33) combined with Lemma 4.1. $\square$

**Example 2**

Our second example is taken from [10, Proposition 3.4] and is concerned with the time function $\tau = t^3$. As we will see, the properties of the respective null distance $\hat{d}_\tau$ are in stark contrast to the properties of $\hat{d}_t$ as described in Proposition 4.2.

**Proposition 4.3.** For the time function $\tau = t^3$ on $\mathbb{R}^n$ the following holds.
(1) \( \hat{d}_\tau \) fails to be definite. In particular, for any two points \( p \) and \( q \) in the \( t = 0 \) slice, we have \( \hat{d}_\tau(p, q) = 0 \).

(2) \( \hat{\tau} \) fails to encode the causal structure. In particular, for any two points \( p = (t_p, p_s) \) and \( q = (t_q, q_s) \), with \( t_p < 0 < t_q \), we have \( \hat{d}_\tau(p, q) = \tau(q) - \tau(p) \).

Proof. We assume \( n = 2 \) throughout the proof, the case \( n > 2 \) is very similar. To prove the first claim of the proposition, for any positive integer \( k \) we consider a piecewise null curve \( \eta_k \) through \( p = (T, p_s) \) and \( q = (T, q_s) \) as shown in the picture:

Here \( q_s - p_s = l \) and the curve has \( 2k \) segments. Clearly, we have

\[
\hat{d}_\tau(p, q) \leq \hat{L}_\tau(\eta_k) = \left( \left( T + \frac{l}{2k} \right)^3 - T^3 \right) \cdot 2k. \tag{39}
\]

Passing to the limit when \( k \to \infty \) we obtain

\[
\hat{d}_\tau(p, q) \leq \lim_{k \to \infty} \hat{L}_\tau(\eta_k) = \lim_{k \to \infty} \left( \left( T + \frac{l}{2k} \right)^3 - T^3 \right) \cdot 2k
\]

\[
= \lim_{k \to \infty} \frac{(T + \frac{l}{2k})^3 - T^3}{\frac{l}{2k} \cdot \frac{1}{l}}
\]

\[
= \lim_{h \to 0} \frac{(T + h)^3 - T^3}{h} \cdot l = \frac{d}{dx} \bigg|_{x=T} x^3 \cdot l
\]

\[
= 3T^2 \cdot l.
\]

For \( T = 0 \) this gives us \( \hat{d}_\tau(p, q) \leq 0 \). On the other hand, by the definition of null distance we have \( \hat{d}_\tau(p, q) \geq 0 \). It follows that \( \hat{d}_\tau(p, q) = 0 \) for \( p \) and \( q \) in
the $t = 0$ slice which completes the proof of the first claim.

We now proceed to the proof of the second claim. For $p = (t_p, p_s)$ and $q = (t_q, q_s)$ with $t_p < 0 < t_q$ we define $p_0 = (0, p_s)$ and $q_0 = (0, q_s)$, see the picture:

Note that by Lemma 4.1 we have

$$\tau(q) - \tau(p) \leq \hat{d}_r(p, q).$$

(41)

so it suffices to show the reverse inequality. For this we use the fact that $\hat{d}_r$ satisfies the triangle inequality, which is obvious from the definition. In particular, we have

$$\hat{d}_r(p, q) \leq \hat{d}_r(p, p_0) + \hat{d}_r(p_0, q_0) + \hat{d}_r(q_0, q).$$

(42)

By the first claim of the proposition, we have $\hat{d}_r(p_0, q_0) = 0$ since $p_0, q_0 \in \{t = 0\}$. Furthermore, it is obvious that $p_0 \in J_+(p)$ and $q \in J_+(q_0)$, hence

$$\hat{d}_r(p_0, q_0) = \tau(p_0) - \tau(q_0) = 0 - \tau(q) = -\tau(p)$$

$$\hat{d}_r(q_0, q) = \tau(q) - \tau(q_0) = \tau(q) - 0 = \tau(q)$$

(43)

by Lemma 4.1. All in all, we obtain

$$\hat{d}_r(p, q) = \tau(q) - \tau(p),$$

(44)

which completes the proof.  

\[\square\]
5 Conclusion and Discussion

The main purpose of this project was to understand the mathematical foundation of the theory of special relativity, focusing on both classical and modern aspects of this subject. In the first part of this project we presented a detailed study of the geometry of Minkowski spacetime. In the second part we showed how well-known relativistic effects such as the constancy of the speed of light and the twin paradox can be explained in terms of geometric properties of Minkowski spacetime. In the third and the last part of this project we studied the notion of null distance given by Sormani and Vega in [10] which, for an appropriate choice of a time function, can be used for equipping Minkowski spacetime with metric space structure, encoding also the causal structure.

It would be interesting to investigate which properties of a time function \( \tau \) make it "good" in the sense described above and on the contrary "bad". The findings of [10], partly presented here, indicate that "goodness" respectively "badness" is related to non-vanishing respectively vanishing of the Lorentzian norm of the gradient \( |\nabla \tau| \). To obtain a complete classification of time functions on Minkowski spacetime based on the properties of their respective null distances would be a first natural step towards understanding the metric properties of more general spacetimes.
References


