Alma Mater Studiorum · Università di Bologna

Dipartimento di Fisica e Astronomia "Augusto Righi"
Corso di Laurea in Fisica

Differential Geometry of Spacetime: from Curves and Surfaces to Special Relativity

Relatore:
Prof. Alessia Cattabriga

Presentata da:
Sofia Santolini

Anno Accademico 2024/2025

Sommario

Questa tesi esplora il profondo legame tra geometria e fisica, mettendo in relazione la descrizione puramente matematica di curve e superfici con le leggi che governano l'universo nel regime delle alte velocità. Partendo dalla relatività ristretta di Einstein, si evidenzia come lo spazio e il tempo siano unificati in un continuo quadridimensionale e come questa nuova concezione di spaziotempo influenzi concetti fondamentali come gli intervalli spazio-temporali, le leggi della dinamica o la simultaneità degli eventi. Successivamente, la trattazione si estende alla geometria intrinseca delle superfici, introducendo concetti quali metrica, curvatura e geodetiche, strumenti essenziali per la comprensione dello spaziotempo curvo della relatività generale. Infine, viene analizzato l'universo di De Sitter, un esempio concreto di universo curvo contrapposto alla geometria piatta dello spaziotempo di Minkowski. Tutto ciò mostra come l'intuizione geometrica possa fornire una profonda comprensione dei fenomeni fisici, dalla curvatura locale delle superfici alla struttura globale del cosmo.

Abstract

This thesis investigates the profound connection between geometry and physics, establishing a direct correlation between the mathematical description of curves and surfaces and the laws that govern the universe at high velocities. Starting from Einstein's special relativity, we draw attention to how space and time merge in a 4-dimensional continuum and how this new understanding of spacetime reshapes fundamental concepts such as space-time intervals, laws of dynamics, or simultaneity. The discussion then extends to the intrinsic geometry of surfaces, introducing key notions like metric, curvature, and geodesics, essential tools to comprehend the curved spacetime of general relativity. Finally, we analyse the De Sitter universe, a concrete example of a curved universe as opposed to the flat geometry of Minkowski spacetime. These insights show how geometric intuition provides a deep understanding of physical phenomena, from the local curvature of surfaces to the global structure of the cosmos.

Contents

Introduction

This thesis explores the profound connection between geometry and physics, relating the abstract mathematical description of curved surfaces to the laws governing the universe at high velocities. By the late 1800s, classical physics found itself at a deadlock. Maxwell's electromagnetic equations did not exhibit invariance when subjected to Galilean transformations, suggesting that either the Galilean transformations were unsuitable for describing electromagnetism or that Maxwell's equations were valid only in a privileged reference frame. However, the Michelson-Morley experiment (1887) evidently failed to detect the Earth's motion through the hypothetical "luminiferous aether", thereby proving that such privileged reference frame (the one at rest with the aether) is nonexistent. It was in this tense atmosphere that Lorentz, Poincaré, and other scientists developed the mathematical transformations that we know today as Lorentz transformations. However, these were originally conceived as simple mathematical tools to preserve the form of classical electromagnetism and to explain the (apparent) validity of Galileo's principle of relativity, despite the failure to detect motion relative to the aether. The year 1905 marked a turning point with the formulation of Einstein's theory of special relativity. Despite reaching conclusions similar to those of Lorentz, Einstein's approach was radically different. By elevating the constancy of the speed of light and the principle of relativity to fundamental postulates, he achieved the same results but with a deeper and more coherent interpretation. In fact, he revealed that the Lorentz transformations embodied profound implications about the nature of space and time, and that they were not merely *ad hoc* mathematical adjustments, but the beginning of one of the most radical paradigm shifts of all time. The geometric implications of this new theory became fully evident through the work of Minkowski, Einstein's former mathematics professor. He showed that special relativity found its most natural expression in a four-dimensional continuum known as Minkowski spacetime, where space and time merge into a single entity. This geometric reformulation elevated time from a simple variable to a coordinate on an equal basis as the spatial ones. The central theme of this thesis is that geometry is not merely a tool for calculating physical quantities: it is the language in which the laws of physics are most naturally expressed.

The first chapter of this thesis is dedicated to special relativity. We begin by enunciating Einstein's postulates and deducing directly from them the relativistic transformations

between inertial frames (the so called Lorentz transformations). This new approach unifies space and time in a four-dimensional framework called Minkowski spacetime. Its geometry leads to new concepts such as space-time intervals, light cones, and worldlines. We also discuss the geometric formulation of special relativity introducing the notions of metric tensor, norm, and scalar product in Minkowski spacetime. To follow, we study some of the most relevant physical consequences of special relativity such as the phenomena of time dilation and length contraction, the new law for addition of velocities, the Doppler effect. Finally, we see some fundamental physical properties such as velocity, momentum, and force expressed in their covariant form as four-vectors, which we will use to build the covariant formulation of dynamics.
The second chapter aims to define the notion of a regular surface in $\mathbb{R}^3$. We start by introducing some preliminary notions on curves (like curvature, parametrization, tangent vector...) that will be useful later on in this chapter to fully understand surfaces and their properties. Then we define regular surfaces, providing the reader with some criteria that should help when trying to decide whether a given subset of $\mathbb{R}^3$ is a regular surface or not. To follow, we begin to study the intrinsic geometry of the surface through the introduction of the metric tensor (or first fundamental form), a natural instrument to treat metric aspects like lengths, angles or areas. After this, we extend the concept of curvature to surfaces, followed by some relevant definitions (the Gauss map, principal curvatures and directions, Gaussian curvature, mean curvature). Then we start the study of intrinsic geometry, that is, the study of those features which can be deduced directly from the metric without reference to the external embedding. A pivotal result of this section is Gauss's Theorema Egregium, which shows that the Gaussian curvature is actually an intrinsic property of surfaces. We conclude the chapter with the study of geodesics, which can be interpreted as the generalization of "straight lines" on a curved surface. This geometric framework, in particular the concepts of intrinsic geometry, geodesics, and curvature, provides the basic mathematical tools for transitioning from the flat spacetime of special relativity to the curved one of general relativity. An example of curved spacetime is De Sitter spacetime, which we analyse in Chapter 3.
The third and final chapter presents an example of curved spacetime: the De Sitter universe. This is one of the simplest and yet most fundamental examples of a curved universe. While special relativity describes the physical phenomena on the flat Minkowski spacetime, this model extends this description to a curved universe. We will start by describing it as a (1+2)-dimensional spacelike hyperboloid of one sheet embedded in the (1+4)-dimensional Minkowski space, and then describe it through a proper parametrization.
Through this progression, this work aims to show how geometric intuition provides the deepest understanding of physics, from the local curvature of a surface to the global structure of the cosmos.
The leading sources for the material presented in this thesis are [Bar04], [dC76], [Cal00].

Chapter 1

Special Relativity

This chapter is dedicated to special relativity, the theory that Albert Einstein formulated in 1905 that completely subverted the traditional concepts of space, time, and motion in a way no-one had ever done before. Newtonian mechanics provide an excellent description of motion in the regime of low velocities, but this approach fails for objects moving at speeds comparable to the one of light. Special relativity resolves the conflicts by introducing a new universal constant, the speed of light in a vacuum c, and by stating that the laws of physics must be the same for all inertial observers. To sum up, this chapter will equip us with the tools to understand physics in the regime of high velocities where the classical intuition and the Newtonian laws must be abandoned.

In the first section (Section 1.1), we present the foundation of special relativity enunciating the two postulates and deducing directly from them the relativistic transformations between inertial frames (the so called Lorentz transformations). This new approach unifies space and time in a four-dimensional framework called Minkowski spacetime. Its geometry leads to new concepts such as space-time intervals, light cones, and worldlines that we explore at the end of the section. In Section 1.2, we begin to discuss of the geometric formulation of special relativity. This section also draws a parallel between rotations in Euclidean geometry and Lorentz transformations, which can also be interpreted as hyperbolic rotations, in Minkowski spacetime. This parallellism is identified to extend the notions of metric tensor, norm, and scalar product to the Minkowski spacetime. To follow, Section 1.3 aims to discuss some of the most relevant physical consequences of special relativity and to see the direct applications of the Lorentz transformations. Among them we have the phenomena of time dilation and length contraction, the new law for addition of velocities, the Doppler effect. Finally, in Section 1.4, we will see some fundamental physical properties such as velocity, momentum and force expressed in their covariant form as four-vectors, which we will use to build the covariant formulation of dynamics.

The leading source for the material presented in this chapter is [Bar04].

1.1 Einstein's solution

Special relativity can be entirely built upon these two postulates that Einstein formulated in 1905.

- The speed of light in a vacuum has the same value $c = 299\,792\,458$ m/s in all inertial frames of reference, independent of the motion of the source or of the observer.
- The laws of physics take the same mathematical form in all *inertial* reference frames.

Inertial frames of reference are connected by particular transformations, and the point is now to find these transformations that preserve the form of any physic's law. Let

$$B_v : K' \to K$$

be the wanted transformation from two reference frames K' and K. How should B_v be defined?

- **Linearity**. Suppose a body A moves at a constant speed in K', then, since K and K' are in uniform motion with respect to one another, A moves at a constant speed in K as well. Hence, A must have a straight worldline both in K and in K', that is, we want the transformation to map straight lines into straight lines. Because we implicitly assume that B_v is invertible and sends the origin to itself, it follows that B_v is a linear transformation.
- **The transformation must have the same form for all pairs of observers**. Since there is no real distinction between uniformly moving observers, two such references K and K' should use the same transformation when they convert from their own spacetime to the other's. When K''s velocity with respect to K is v, K's velocity with respect to K' must be $-v$. Therefore, since $B_v : K \to K'$, we must have $B_{-v} : K \to K'$. But the map $K' \to K$ has to be the inverse of the map $K \to K'$, hence $B_v^{-1} = B_{-v}$.

We will call these transformations **Lorentz transformations**.

1.1.1 Events

In physics, an *event* is a set of four real numbers (x, y, z, t), three of which are spatial coordinates that indicate *where* the event took place, and the fourth is the instant of time that specifies *when* the event occurred.

Spacetime Interval

Let's now introduce the notion of *spacetime interval* between two events. Let $A = (x_A, y_A, z_A, t_A)$ and $B = (x_B, y_B, z_B, t_B)$ be two events. Then the squared **interval** between A and B is defined as

$$\Delta s^2 = c^2(t_B - t_A)^2 - (x_B - x_A)^2 - (y_B - y_A)^2 - (z_B - z_A)^2,$$

or, more briefly

$$\Delta s^2 = c^2\Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2.$$

Let us also introduce the infinitesimal interval ds between two events of type (t, x, y, z) and $(t + dt, x + dx, y + dy, z + dz)$ defined as

$$ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2.$$

The set of all events equipped with the above structure is called **Minkowski spacetime**.

Starting from the principles of relativity, we will show an extremely important conclusion: the interval between two events must be the same for all inertial observers, that is, the interval is an invariant quantity.

Null Interval

Let's consider a reference frame K. Suppose that event $A = (x_A, y_A, z_A, t_A)$ corresponds to the emission of a light signal, and event $B = (x_B, y_B, z_B, t_B)$ corresponds to the receiving of that same light signal. The interval between them is

$$\Delta s^2 = c^2(t_B - t_A)^2 - (x_B - x_A)^2 - (y_B - y_A)^2 - (z_B - z_A)^2.$$

According to the first postulate, light signals propagate isotropically at speed $v = c$. After a time t, the emitted signals occupy a spherical wavefront of radius $r = ct$, where $r = \sqrt{x^2 + y^2 + z^2}$. Squaring both sides, we obtain the equation of the wavefront,

$$c^2(t_B - t_A)^2 = (x_B - x_A)^2 + (y_B - y_A)^2 + (z_B - z_A)^2.$$

This implies that the previous interval is equal to zero: $\Delta s^2 = 0$.

Now let's see what happens in another inertial reference frame K'. The events will be identified by the coordinates $A' = (x'_A, y'_A, z'_A, t'_A)$ and $B' = (x'_B, y'_B, z'_B, t'_B)$. They are separated by the interval

$$\Delta s'^2 = c^2(t'_B - t'_A)^2 - (x'_B - x'_A)^2 - (y'_B - y'_A)^2 - (z'_B - z'_A)^2.$$

According again to the first postulate, the speed of light c is constant for any inertial observer, so the equation of the wavefront in K' is

$$c^2(t'_B - t'_A)^2 = (x'_B - x'_A)^2 + (y'_B - y'_A)^2 + (z'_B - z'_A)^2,$$

hence, the interval between A' and B' is null in K' too: $\Delta s'^2 = 0$.
We therefore conclude that

$$\Delta s^2 = 0 \implies \Delta s'^2 = 0.$$

General Invariance of the Interval

We now wish to show that this holds for any interval, not only for null ones. Assuming that the coordinate transformation between K and K' is linear, then $\Delta s^2 = 0 \implies \Delta s'^2 = 0$ implies that

$$\Delta s^2 = \lambda(v)\Delta s'^2,$$

where, by space-time isotropy, λ depends only on v and is an even function of v, that is, $\lambda(v) = \lambda(-v)$. So we have

$$\Delta s'^2 = \frac{1}{\lambda(v)}\Delta s^2.$$

On the other hand, according to the second postulate, all inertial frames are equivalent to each other, so the transformation from K to K' must be the same as the one from K' to K, except that v changes sign.[1]

$$\Delta s'^2 = \lambda(-v)\Delta s^2.$$

Thus, we get to

$$\lambda(-v) = \frac{1}{\lambda(v)} \implies \lambda^2(v) = 1.$$

The solution $\lambda(v) = -1$ is not acceptable because, for $v = 0$, we must have the identity $\Delta s'^2 = \Delta s^2$, that is, $\lambda(0) = 1$. We conclude that $\lambda(v) = +1$, hence $\Delta s^2 = \Delta s'^2$.
To sum up, the interval between two events is invariant under Lorentz transformations, and any transformation between two inertial reference frames must preserve this property.

1.1.2 The Lorentz Transformations

As shown in the previous section, transformation between two systems K and K' must preserve the interval between two events to be consistent with Einstein's postulates:

$$\Delta s^2 = \Delta s'^2.$$

We now derive the explicit form of the Lorentz transformations from this condition.

[1] We are using the fact that $B_v^{-1} = B_{-v}$.

General Linear Transformation

Since Δs^2 must be conserved, we expect the transformations to involve a mixing between space and time coordinates. For simplicity, consider two frames of reference K and K' in motion with respect to each other at a constant speed v parallel to the x-axis, so that the y-axis and the z-axis are parallel, like in Figure 1.1. Choose the origin of times in such a way that at $t = t' = 0$, the origins O and O' coincide.

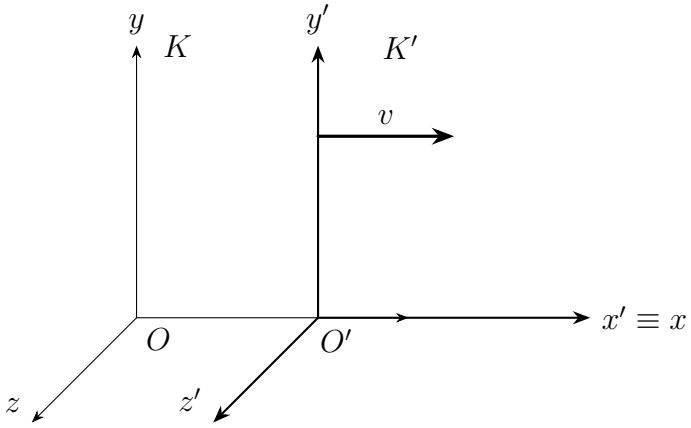

Figure 1.1: Reference frames K and K' in uniform relative motion along the x-axis.

The most general linear transformation between such reference frames is

$$\begin{pmatrix} x' \\ y' \\ z' \\ t' \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ t \end{pmatrix},$$

which becomes the following linear system

$$\begin{cases} x' = a_{11}x + a_{12}y + a_{13}z + a_{14}t, \\ y' = a_{21}x + a_{22}y + a_{23}z + a_{24}t, \\ z' = a_{31}x + a_{32}y + a_{33}z + a_{34}t, \\ t' = a_{41}x + a_{42}y + a_{43}z + a_{44}t. \end{cases}$$

Under the previous hypotheses $x' \equiv x$, and $v \parallel x$, with y' and z' parallel to y and z respectively, the space coordinates do not get mixed up (K's axes are not rotated with respect to K''s), and the transverse coordinates y' and z' do not depend on t since the relative motion is along the x-axis. To sum up, x' can only depend on x and t, while y' and z' can only depend on y and z respectively; we also expect t' to depend only on x and t, otherwise, K' would have a cinematic component depending on the transverse axes,

but the only physically relevant direction is the direction of relative motion between K and K'. These considerations translate into:

$$\begin{cases} x' = a_{11}x + a_{14}t, \\ y' = a_{22}y, \\ z' = a_{33}z, \\ t' = a_{41}x + a_{42}y + a_{43}z + a_{44}t, \end{cases}$$

where the coefficients a_{ij} will be, in general, a function of v. By space isotropy we will have

$$a_{22} = a_{33},$$

hence, the system becomes

$$\begin{cases} x' = a_{11}x + a_{14}t, \\ y' = a_{22}y, \\ z' = a_{22}z, \\ t' = a_{41}x + a_{44}t. \end{cases}$$

By renominating $a_{11} = A(v)$, $a_{14} = B(v)$, $a_{44} = C(v)$, $a_{41} = D(v)$, and $a_{22} = E(v)$, the equations become

$$\begin{cases} x' = A(v)x + B(v)t, \\ y' = E(v)y, \\ z' = E(v)z, \\ t' = C(v)t + D(v)x. \end{cases}$$

Application of the Physical Conditions

K''s origin moves according to the equation of motion $x = vt$, hence for $x' = 0$,

$$0 = Ax + Bt \implies x = -BA^{-1}t \implies v = -BA^{-1} \implies B = -Av.$$

The equation assumes the form

$$\begin{cases} x' = A(v)(x - vt), \\ y' = E(v)y, \\ z' = E(v)z, \\ t' = C(v)t + D(v)x. \end{cases}$$

If $v = 0$, the transformations must reduce to the identity transformation

$$x' = x, \quad y' = y, \quad z' = z, \quad t' = t.$$

This implies

$$A(0) = 1, \quad C(0) = 1, \quad D(0) = 1, \quad E(0) = 1.$$

Now we can impose the condition

$$c^2t'^2 - x'^2 - y'^2 - z'^2 = c^2t^2 - x^2 - y^2 - z^2,$$

which becomes

$$\begin{aligned} c^2t'^2 - x'^2 - y'^2 - z'^2 &= c^2(Ct + Dx)^2 - (A(x - vt))^2 - (Ey)^2 - (Ez)^2 \\ &= (c^2C^2 - v^2A^2)t^2 - (A^2 - c^2D^2)x^2 - E^2y^2 - E^2z^2 + 2(vA^2 + c^2CD)xt \\ &= c^2t^2 - x^2 - y^2 - z^2. \end{aligned}$$

By equating the corresponding coefficients, we obtain,

$$\begin{aligned} &E^2 = 1 \\ &c^2C^2 - v^2A^2 = c^2, \\ &A^2 - c^2D^2 = 1, \\ &vA^2 + c^2CD = 0. \end{aligned}$$

Since $E(0) = 1$, we immediately obtain $E(v) = 1$. The second equation gives

$$A^2 = \frac{c^2}{v^2}(C^2 - 1).$$

If substituted in the last two equations, it becomes

$$\begin{aligned} vA^2 + c^2CD = 0 \quad \Longrightarrow \quad & v\frac{c^2}{v^2}(C^2 - 1) + c^2CD = 0 \\ & C^2 - 1 + vCD = 0 \\ & D = \frac{1 - C^2}{vC}, \end{aligned}$$

and

$$A^2 - c^2D^2 = 1 \quad \Longrightarrow \quad \frac{c^2}{v^2}(C^2 - 1) - c^2D^2 = 1.$$

And by substituting D, we have

$$\begin{aligned} \frac{c^2}{v^2}(C^2 - 1) - c^2D^2 = 1 \quad \Longrightarrow \quad & \frac{c^2}{v^2}(C^2 - 1) - c^2\left(\frac{1 - C^2}{vC}\right)^2 = 1 \\ & C^2(1 - C^2) + (1 - C^2)^2 + \frac{v^2}{c^2}C^2 = 0 \\ & (1 - C^2)(1 - C^2 + C^2) + \frac{v^2}{c^2}C^2 = 0 \\ & C^2 = \frac{1}{1 - \frac{v^2}{c^2}}. \end{aligned}$$

Finally, we find

$$A^2 = \frac{c^2}{v^2}(C^2 - 1) = \frac{c^2}{v^2}\left(\frac{1}{1 - v^2/c^2} - 1\right) = \frac{1}{1 - \dfrac{v^2}{c^2}} = C^2.$$

The acceptable solutions for A and C are the positive ones due to the conditions $A(0) = C(0) = 1$, hence

$$\begin{aligned} A(v) = C(v) &= \frac{1}{\sqrt{1 - \dfrac{v^2}{c^2}}}, \\ D(v) &= -\frac{v}{c^2\sqrt{1 - \dfrac{v^2}{c^2}}}, \\ E(v) &= 1. \end{aligned}$$

Final Form of the Lorentz Transformations

Now that we have found the coefficients, we can write the Lorentz transformations in their final form

$$\begin{cases} x' = \dfrac{x - vt}{\sqrt{1 - \dfrac{v^2}{c^2}}}, \\ y' = y, \\ z' = z, \\ t' = \dfrac{t - \dfrac{vx}{c^2}}{\sqrt{1 - \dfrac{v^2}{c^2}}}. \end{cases}$$

The inverse transformations (from K' to K) are obtained by switching (x', y', z', t') with (x, y, z, t) and by inverting the sign of the velocity. These are the **Lorentz transformations**. The reason they are called "Lorentz transformations" is that they were first deduced by Lorentz [Lor04]. However, he created them *ad hoc* to be consistent with the facts. On the contrary, Einstein deduced them from the two postulates with which we opened the chapter.
By introducing

$$\gamma = \frac{1}{\sqrt{1 - \dfrac{v^2}{c^2}}}, \quad \beta = \frac{v}{c},$$

the expression simplifies to

$$\begin{cases} x' = \gamma(x - \beta ct), \\ y' = y, \\ z' = z, \\ ct' = \gamma\left(ct - \beta x\right), \end{cases}$$

in this form, the Lorentz transformations exhibit a symmetry between x and ct that can be better appreciated when written as matrices

$$\begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} = B_x(v) \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}$$

where

$$B_{v,x} = \begin{pmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The matrix $B_{v,x}$ is the Lorentz matrix for the transformation from $K \to K'$ in case of relative velocity along the x-axis. In this case, we say that the transformation performs a *boost* along x. The other cases are given by

$$B_{v,y} = \begin{pmatrix} \gamma & 0 & -\beta\gamma & 0 \\ 0 & 1 & 0 & 0 \\ -\beta\gamma & 0 & \gamma & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \text{boost along the } y\text{-axis;}$$

$$B_{v,z} = \begin{pmatrix} \gamma & 0 & 0 & -\beta\gamma \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\beta\gamma & 0 & 0 & \gamma \end{pmatrix}, \quad \text{boost along the } z\text{-axis.}$$

The Lorentz matrix corresponding to a boost in a generic direction is

$$B_v = \begin{pmatrix} \gamma & -\beta_x\gamma & -\beta_y\gamma & -\beta_z\gamma \\ -\beta_x\gamma & 1 + \left(\dfrac{\gamma - 2}{|\beta|^2}\right)\beta_x^2 & \dfrac{\gamma - 1}{|\beta|^2}\beta_x\beta_y & \dfrac{\gamma - 1}{|\beta|^2}\beta_x\beta_z \\ -\beta_y\gamma & \dfrac{\gamma - 1}{|\beta|^2}\beta_y\beta_x & 1 + \left(\dfrac{\gamma - 2}{|\beta|^2}\right)\beta_y^2 & \dfrac{\gamma - 1}{|\beta|^2}\beta_y\beta_z \\ -\beta_z\gamma & \dfrac{\gamma - 1}{|\beta|^2}\beta_z\beta_x & \dfrac{\gamma - 1}{|\beta|^2}\beta_x z\beta_y & 1 + \left(\dfrac{\gamma - 2}{|\beta|^2}\right)\beta_z^2 \end{pmatrix},$$

where, $\beta = (\beta_x, \beta_y, \beta_z)$ and $|\beta|^2 = \beta_x^2 + \beta_y^2 + \beta_z^2$.

Geometric Interpretation

It is easy to show that

$$\gamma^2 - (\gamma\beta)^2 = 1,$$

hence, we can interpret $\gamma = \cosh u$ and $\gamma\beta = \sinh u$, where $u = \tanh^{-1}(v/c)$. In this perspective, we can rewrite the Lorentz matrix in terms of the hyperbolic functions, leading to a new interpretation of these transformations: hyperbolic rotations. A boost along the x-axis will be described by

$$H_{u,x} = \begin{pmatrix} \cosh u & -\sinh u & 0 & 0 \\ -\sinh u & \cosh u & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

which represents a hyperbolic rotation by the hyperbolic angle u around the x-axis, and analogously for boosts in the other directions.
The difference between the two matrices B_v and H_u is simply the approach they use to describe Lorentz transformations: while H_u gives a geometric description based on the hyperbolic angle u, B_v provides a physical description in terms of the velocity parameter v. In other words, we can think of B_v as the **boost** by velocity v, and of H_u as the **hyperbolic rotation** by angle u.

1.1.3 Minkowski Diagrams

Minkowski Diagrams in (1+1) Dimensions

As discussed in the beginning of the chapter, in relativity, time is not simply a parameter, it is a variable just like the three space coordinates (x, y, z): these four numbers together can be used to identify any event in this new $(1 + 3)$-dimensional vector space called Minkowski spacetime, which we will denote $\mathcal{M}$. We can represent the set of all events by adding to the three spatial axes a fourth one that represents time. To see this more clearly, let's consider a $(1 + 1)$-dimensional case where events are identified by the pair (ct, x) according to an inertial frame of reference K. This can be graphically represented by a *Minkowski diagram* like the one in Figure 1.2.

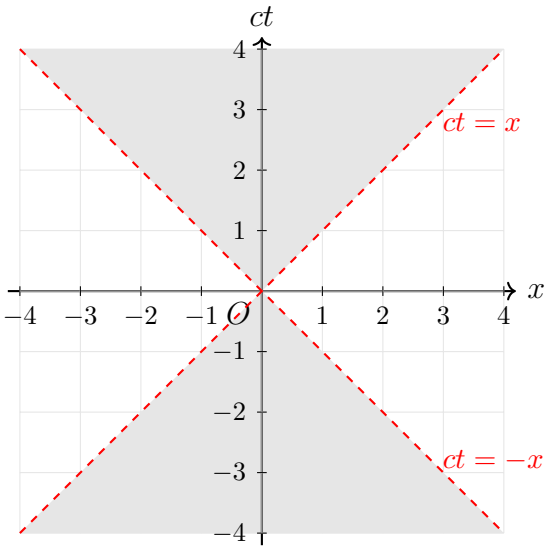

Figure 1.2: Minkowski space-time diagram in the $(1+1)$-dimensional case. On the x-axis the x coordinate, on the y-axis the time coordinate ct. The red dashed lines represent the worldlines of two photons traveling at speed $\pm c$.

Worldlines

Minkowski diagrams are pictures of space-time because they also have a coordinate axis for time. The motion of a particle in spacetime is then represented in these diagrams simply as a line with a slope equal to the velocity in units of c. For instance, a particle moving at uniform velocity v is represented by a straight line of slope v/c. We call this line **history** or **worldline** of the particle. If the particle is a photon, then its speed is constant and equal to $\pm c$, so the worldline is a straight line with slope ± 1. Since the maximum speed of any object is the speed of light, it follows that the worldlines of photons are the least steep trajectories in the Minkowski diagram.

The Light Cone

If we now consider all the possible worldlines of photons emitted from the origin, they form a boundary that no massive particle can cross. This boundary is called the **light cone**, and its equation $\Delta s^2 = 0$ corresponds to the invariant quantity that we introduced in the first section. In the 2-dimensional case, the light cone reduces to the two gray-colored areas in Figure 1.2, but in the full 4-dimensional case, the photons spread out in all directions and occupy the spherical shell $x^2 + y^2 + z^2 = c^2t^2$ of radius ct after t seconds have passed. (Figure 1.3). No trajectories can exist outside the light cone; all the permitted trajectories lie inside it.

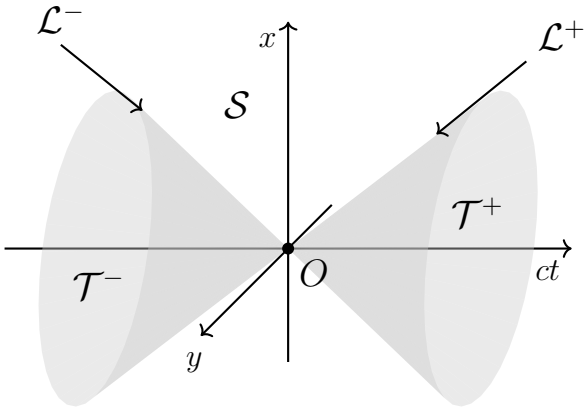

Figure 1.3: Relativistic light cone in the $(1+2)$-dimensional case of coordinates (ct, x, y). $\mathcal{L}^+$ and $\mathcal{L}^-$ stand for "future lightlike" and past "past lightlike" respectively and they refer to the future and past vectors lying on the light cone. $\mathcal{T}^+$ and $\mathcal{T}^-$ stand for "future timelike" and "past lightlike" respectively and they refer to the future and past vectors inside the light cone. $\mathcal{S}$ stands for "spacelike" and it refers to the vectors outside the light cone. This partition of spacetime will be analysed in the following section.

As we saw in the previous sections, the Lorentz transformations preserve the quadratic form $c^2t^2-x^2-y^2-z^2$, which is the equation of the light cone, hence these transformations preserve the light cone.

1.2 Minkowski Geometry

Despite having the same dimensions, $\mathcal{M}$ and $\mathbb{R}^4$ are not the same vector space, and in order to appreciate the differences and the analogies between them, we want to introduce the notion of *norm* in these spaces and see how it differs in the two cases. In the following paragraphs, we will be using Einstein's **summation convention**: whenever an index appears twice, once as a subscript, and once as a superscript, it must be summed over. For example,

$$x^i y_i = \sum_i x^i y_i,$$

1.2.1 The Euclidean norm

Cartesian Coordinates and Inner Product

Before extending these ideas to Minkowski spacetime, let's recall some basic concepts of Euclidean geometry in $\mathbb{R}^n$. Each point in $\mathbb{R}^n$ is identified by a set of n coordinates $(x^1, x^2, \ldots, x^n)$. The tool to measure length and angles in Cartesian geometry is the

standard **Euclidean inner product** defined, for two vectors $\mathbf{u} = (u^1, u^2, \ldots, u^n)$ and $\mathbf{v} = (v^1, v^2, \ldots, v^n)$, as

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^t \mathbb{I} \mathbf{v} = (v^1, v^2, \ldots, u^n) \begin{pmatrix} v^1 \\ v^2 \\ \vdots \\ v^n \end{pmatrix} = u^1 v^1 + u^2 v^2 + \cdots + u^n v^n.$$

with $\mathbb{I} = \delta_{ij}$ the identity matrix. Since in Euclidean geometry, the metric tensor is canonically associated with the identity matrix, we may sometimes simply refer to it as the identity matrix.

Given this, the Euclidean inner product becomes

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{ij} u^i v^j = \delta_{ij} u^i v^j.$$

Euclidean Norm, Distance and Angle

From the inner product, we get the length, or **Euclidean norm**, of a vector:

$$|\mathbf{x}| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{(x^1)^2 + (x^2)^2 + \cdots + (x^n)^2},$$

which can also be written as

$$|\mathbf{x}|^2 = \delta_{ij} x^i x^j.$$

We can now make basic measurements on the plane, like the distance from $\mathbf{u}$ to $\mathbf{v}$ can be expressed as

$$|\mathbf{u} - \mathbf{v}| = \sqrt{(u^1 - v^1)^2 + (u^2 - v^2)^2 + \cdots + (u^n - v^n)^2},$$

which, in terms of the metric tensor becomes

$$|\mathbf{u} - \mathbf{v}| = \sqrt{\delta_{ij} \Delta x^i \Delta x^j} = \Delta \mathbf{x}^2,$$

where $\Delta x^1 \equiv u^1 - v^1$, $\Delta x^2 = u^2 - v^2$, $\ldots \Delta x^n = u^n - v^n$.
We can also measure the angle between two non-null vectors $\mathbf{u}$ and $\mathbf{v}$:

$$\alpha = \arccos\left(\frac{\langle \mathbf{u}, \mathbf{v} \rangle}{|\mathbf{u}||\mathbf{v}|} \right).$$

The linear transformations that leave the Euclidean norm unchanged are called *orthogonal transformations*, which form a group called $O(n)$. The determinant of orthogonal matrices R satisfies $\det(R) = \pm 1$. The set of orthogonal matrices with determinant $\det = +1$ is still a group, and it is called the *special orthogonal group* $SO(n)$. Its elements are called *rotations*.

1.2.2 The Minkowski norm

In the previous section, we introduced the Euclidean inner product and norm in $\mathbb{R}^n$, which provide a tool for measuring lengths and angles in ordinary space. We now want to extend these notions to the $(1+3)$-dimensional Minkowski space $\mathcal{M}$.

Index Notation

In $\mathcal{M}$, events are identified by four coordinates x, y, z, t. Let's start by renaming them

$$x^0 \equiv ct, \quad x^1 \equiv x, \quad x^2 \equiv y, \quad x^3 \equiv z.$$

These new coordinates will be collectively indicated through a greek index, which will assume the following values

$$x^\mu = (ct, x, y, z) = (ct, \mathbf{x}), \qquad \mu = 0, 1, 2, 3.$$

From now on, when working on $\mathcal{M}$, the greek indices $(\alpha, \beta, \gamma, ...)$ will assume the values $0, 1, 2, 3$; whereas the latin indices $(i, j, k, ...)$ will assume the values $1, 2, 3$. The position of such indices is never arbitrary: we will see that moving one index from a lower to an upper position (or vice versa) can produce sign changes. Let's now introduce the inner product in $\mathcal{M}$.

The Minkowski Metric

Given two events $A = x^\mu_A = (x^0_A, x^1_A, x^2_A, x^3_A)$ and $B = x^\mu_B = (x^0_B, x^1_B, x^2_B, x^3_B)$, the (squared) distance between them in Euclidean geometry is given by

$$\Delta\mathbf{x}^2 = \delta_{ij}\Delta x^i \Delta x^j = \Delta x^2 + \Delta y^2 + \Delta z^2.$$

Whereas, in Minkowski spacetime, this generalises to

$$\Delta s^2 = \sum_{\mu=0}^{3}\sum_{\nu=0}^{3} g_{\mu\nu}\Delta x^\mu \Delta x^\nu,$$

where

$$\begin{aligned}\Delta x^\mu &= (\Delta x^0, \Delta\mathbf{x}) = (\Delta x^0, \Delta x^1, \Delta x^2, \Delta x^3) \\ &= ((x^0_B - x^0_A), (x^1_B - x^1_A), (x^2_B - x^2_A), (x^3_B - x^3_A))\end{aligned}$$

is the separation between the two events, and $g_{\mu\nu}$ is the so called *metric tensor*, whose components are

$$g = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

By expanding the product, we obtain the familiar squared *spacetime interval* that we introduced in Section 1.1

$$\Delta s^2 = (\Delta x^0)^2 - (\Delta x^1)^2 - (\Delta x^2)^2 - (\Delta x^3)^2 = c^2\Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2,$$

which differs from the euclidean distance by the presence of the time coordinate and the sign of the spatial part. Just as δ_{ij} defines the geometry of Euclidean space, $g_{\mu\nu}$ defines the geometry of Minkowski spacetime.

Covariant and Contravariant Components

The coordinates characterized by upper indices, x^μ, are called *contravariant*. Let us now introduce the *covariant* coordinates x_μ defined as

$$x_\mu = g_{\mu\nu}x^\nu.$$

More explicitly,

$$x_0 = x^0, \quad x_1 = -x^1, \quad x_2 = -x^2, \quad x_3 = -x^3.$$

Hence, lowering an index flips the sign of the spatial components but not of the temporal one.

We now want to introduce the contravariant metric tensor $g^{\mu\nu}$ which corresponds to the inverse matrix g^{-1}, that is

$$g_{\mu\nu}g^{\nu\rho} = \delta_\mu{}^\rho.$$

The inverse metric tensor performs the opposite operation: it raises indices, hence, it allows us to go from covariant to contravariant coordinates.

$$x^\mu = g^{\mu\nu}x_\nu$$

Given this, the invariant quantity Δs^2 can be written as

$$\Delta s^2 = \Delta x^\mu \Delta x_\mu.$$

Minkowski Inner Product and Norm

In analogy with the Euclidean inner product, which is invariant under orthogonal transformations, we can define the **Minkowski inner product** of two vectors $\mathbf{u} = u^\mu$ and $\mathbf{v} = v^\nu$ as the quantity that is preserved under Lorentz transformations, that is

$$\langle \mathbf{u}, \mathbf{v} \rangle = g_{\mu\nu}u^\mu v^\nu.$$

In matrix form

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T g \mathbf{v} = c^2 t_1 t_2 - x_1 x_2 - y_1 y_2 - z_1 z_2.$$

This inner product is the Minkowskian analogue of the Euclidean one: it provides a natural way to measure generalized angles and lengths $\mathcal{M}$.

The **Minkowski** (squared) **norm** of a 4-vector $\mathbf{w} = w^\mu$ is also an invariant quantity, and it is simply the Minkowski inner product of a vector with itself.

$$|\mathbf{w}|^2 = \langle \mathbf{w}, \mathbf{w} \rangle = c^2t^2 - x^2 - y^2 - z^2 = g_{\mu\nu} w^\mu w^\nu = w_\nu w^\nu.$$

From this point of view, the squared interval Δs^2 introduced above is nothing but the Minkowski squared norm of the displacement vector between the two events A and B

$$\Delta s^2 = \langle x_B^\mu - x_A^\mu, x_B^\nu - x_A^\nu \rangle = \langle \Delta x^\mu, \Delta x^\nu \rangle = g_{\mu\nu} \Delta x^\mu \Delta x^\nu.$$

Lorentz transformations are those that preserve the Minkowski inner product and squared norm, that is, they satisfy $|B_v(\mathbf{w})| = |\mathbf{w}|$. This condition implies that Lorentz matrices are pseudo-orthogonal[2]. In fact,

$$|B_v \mathbf{w}|^2 = \langle B_v(\mathbf{w}), B_v(\mathbf{w}) \rangle = (B_v(\mathbf{w}))^T g(B_v(\mathbf{w})) = \mathbf{w}^T B_v^T g B_v \mathbf{w}.$$

On the other hand,

$$|\mathbf{w}|^2 = \langle \mathbf{w}, \mathbf{w} \rangle = \mathbf{w}^T g \mathbf{w},$$

and the two must be equal, so

$$\mathbf{w}^T B_v^T g B_v \mathbf{w} = \mathbf{w}^T g \mathbf{w}.$$

This means that $B_v^T g B_v = g$, hence Lorentz transformations are performed by pseudo-orthogonal matrices having g as metric.

The Lorentz transformation B_v is designed to preserve the light cone, that is, the set $c^2t^2 - x^2 - y^2 - z^2 = 0$. Actually, B_v preserves each of the hyperbolas $c^2t^2 - x^2 - y^2 - z^2 = k$, $\forall k$, whose common asymptote is the light cone ($k = 0$).

We can see a direct correspondence between rotations in $\mathbb{R}^4$ performed by the $SO(4)$ group, and Lorentz transformations on $\mathcal{M}$ performed by the boost B_v. These matrices form a group called the Lorentz group $\mathcal{L}$ which is isomorphic to $SO(3,1)$, the group of 4×4 pseudo-orthogonal matrices with determinant $\det = +1$.

Classification of Events in Spacetime

The arising problem is that the Minkowski quadratic form defined above is not positive definite, so it may assume negative values. We can classify events according to the sign

[2] A matrix Λ is pseudo-orthogonal if there exists a symmetric non-degenerate matrix η, called the *metric*, such that $\Lambda^T \eta \Lambda = \eta$.

of their Minkowski norm: as we can see in Figure 1.3, the set $\langle \mathbf{w}, \mathbf{w} \rangle = 0$ is a cone that separates spacetime into two regions where $\langle \mathbf{w}, \mathbf{w} \rangle$ is either positive or negative (and null on the cone).

We have $\langle \mathbf{w}, \mathbf{w} \rangle > 0$ for all the points *inside* the cone, and all the events in this region are called **timelike**. *Outside* the cone $\langle \mathbf{w}, \mathbf{w} \rangle < 0$ and the events in this region are said to be **spacelike**. Finally, the events lying *on* the cone itself are **lightlike**, and they are such that $\langle \mathbf{w}, \mathbf{w} \rangle = 0$.

Since the physical distinction between past and future is important, we now want to refine our partition of spacetime.
Spacetime consists of the following six mutually exclusive sets of events or vectors $\mathbf{w} = (ct, x, y, z)$(Figure 1.3):

- $\mathcal{T}_+$: the future timelike set $\langle \mathbf{w}, \mathbf{w} \rangle > 0$, $t > 0$;
- $\mathcal{T}_-$: the past timelike set $\langle \mathbf{w}, \mathbf{w} \rangle > 0$, $t < 0$;
- $\mathcal{S}$: the spacelike set $\langle \mathbf{w}, \mathbf{w} \rangle < 0$;
- $\mathcal{L}_+$: the future lightlike set $\langle \mathbf{w}, \mathbf{w} \rangle = 0$, $t > 0$;
- $\mathcal{L}_-$: the past lightlike set $\langle \mathbf{w}, \mathbf{w} \rangle = 0$, $t < 0$;
- $\mathcal{O}$: the origin.

The coordinates of the events $\mathbf{w} = (ct, x, y, z)$ change depending on the reference frame; however, their "type" is invariant.
Moreover, each region of spacetime is mapped onto itself by Lorentz transformations. Indeed, these regions are classified according to the sign of the Minkowski norm of their vectors. Since the norm is invariant, the regions themselves remain unchanged under any Lorentz transformation.

1.3 Physical Consequences

Let us now examine the physical consequences of the previous statements. For simplicity, we will refer to inertial reference frames K and K' moving at a constant relative velocity v along the x-axis. In this configuration (see Figure 1.1), the y and z coordinates remain unchanged under any Lorentz transformation, allowing us to reduce the analysis to the $(1 + 1)$-dimensional case, a simplification that is sufficient to capture all the essential physical consequences of the previous discussion.

1.3.1 Length and Time

One of the first consequences of special relativity is that *length* and *time* are no longer independent. This is due to the introduction of a new physical constant: $c = 2.99792458 \times 10^8$ m/s, the speed of light. Setting $c = 1$ (geometric units), time and length can be expressed in the same unit.

1.3.2 Simultaneity

The principle of relativity implies that the notion of simultaneity is not physically meaningful. Two events that are simultaneous in a reference frame K, need not be likewise for another inertial observer K'.
Assume that two events A and B are simultaneous in K. We will have then,

$$A = (t, x_A),$$
$$B = (t, x_B).$$

Hence $\Delta t = t_B - t_A = 0$.
In another inertial frame K' in relative motion with respect to K with velocity v along the x-axis, the time coordinates will be

$$t'_A = \frac{t - \frac{vx_A}{c^2}}{\sqrt{1-\beta^2}},$$
$$t'_B = \frac{t - \frac{vx_B}{c^2}}{\sqrt{1-\beta^2}}.$$

Hence, in K', the events are separated by the time interval

$$\Delta t' = t'_B - t'_A = \gamma \frac{v}{c^2}(x_B - x_A),$$

which, in general, is nonzero. Two simultaneous events in K, are not simultaneous in K', unless they *coincide*. If two events coincide in one frame of reference ($x_A = x_B, t_A = t_B$), then they coincide in *any* inertial frame of reference.

1.3.3 Composition of Velocities

One of the main consequences of Einstein's postulates on special relativity is that the correct composition of velocities is not the Galilean one. To find the real composition law, let us consider two inertial frames of reference K and K' in uniform relative motion along the x-axis, having parallel y- and z-axes. Let v be the relative speed between the

two systems. The equation of motion of a particle will be $\mathbf{x}(t) = (x(t), y(t), z(t))$ in K and $\mathbf{x}'(t) = (x'(t), y'(t), z'(t))$ in K'. Its velocity will be

$$\mathbf{u} = \frac{d\mathbf{x}}{dt} \quad \text{in } K,$$
$$\mathbf{u}' = \frac{d\mathbf{x}'}{dt'} \quad \text{in } K'.$$

By differentiating the Lorentz transformations, one finds

$$dx' = \frac{u_x - v}{\sqrt{1 - \frac{v^2}{c^2}}} dt,$$
$$dy' = u_y dt,$$
$$dz' = u_z dt,$$
$$dt' = \frac{1 - \frac{u_x v}{c^2}}{\sqrt{1 - \frac{v^2}{c^2}}}.$$

Dividing the first three equations by the fourth one, we obtain

$$u'_x = \frac{u_x - v}{1 - \frac{u_x v}{c^2}},$$
$$u'_y = \frac{u_y \sqrt{1 - \frac{v^2}{c^2}}}{1 - \frac{u_x v}{c^2}},$$
$$u'_z = \frac{u_z \sqrt{1 - \frac{v^2}{c^2}}}{1 - \frac{u_x v}{c^2}}.$$

These are the relativistic composition laws of velocity. If the particle moves in K along the x-axis in the same direction as the relative motion of K', in K' that very same particle will move with velocity

$$u' = \frac{u - v}{1 - \frac{uv}{c^2}}.$$

This addition operation is consistent with the fact that c is the highest possible speed. If we consider a light signal with velocity $u = c$ in K', its velocity c' in K' will be

$$c' = \frac{c - v}{1 - \frac{cv}{c^2}} = c.$$

1.3.4 Causality

The **causal future** of an event A is the set of all events that A can influence. The **causal past** of A is the set of all the events that can influence A.
If the event A causes or influences the event B, then A must happen before B. Since we want this to be a physical law, A must happen before B for *all* inertial observers. As we saw in the previous section, each event P divides spacetime into three regions: besides the causal future and past of P, there is a region consisting of events that can neither influence nor be influenced by P which consists of the events Q for which the separation $Q - P$ is spacelike (see "Classification of Events in Spacetime" in Section 1.2.2). The partition is drawn in Figure 1.4.

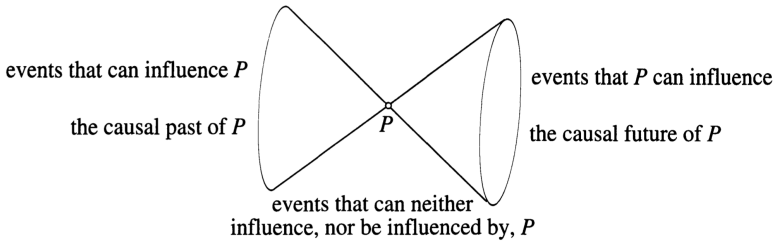

Figure 1.4: Light cone of event P. From [Cal00, page 77].

For the laws of physics to remain consistent, causality must be preserved. This condition is guaranteed by the theory of relativity. Consider two events A and B of coordinates $A = (t_A, x_A)$, $B = (t_B, x_B)$ corresponding, for instance, to the emission and receiving of a signal. They are clearly causally connected, let's say that A influences B, hence B happens after A: $\Delta t \equiv t_B - t_A > 0$. If the speed u of the signal is constant, we have

$$x_B - x_A = u(t_B - t_A).$$

Now analyse the same situation in an inertial reference frame K' moving at constant speed v with respect to the previous system K:

$$\begin{aligned}\Delta t' = t'_B - t'_A &= \gamma\left(t_B - \frac{vx_B}{c^2}\right) - \gamma\left(t_A - \frac{vx_A}{c^2}\right) \\ &= \gamma\left((t_B - t_a) - \frac{v}{c^2}(x_B - x_A)\right) = \gamma\left((t_B - t_A) - \frac{v}{c^2}u(t_B - t_A)\right) \\ &= \gamma\Delta t\left(1 - \frac{vu}{c^2}\right).\end{aligned}$$

If $u, v < c$, the quantity in brackets is positive, hence, $\Delta t'$ and Δt have the same sign. If v or u were greater than c, then it would be possible for the two intervals to have opposite signs. If this happened, causality would not be preserved because there would exist some frame of reference in which B, which is influenced by A, happens before A itself. The fact that c is constant and is the highest possible speed preserves causality.

1.3.5 Time Dilation

Suppose that frame K' is carrying a clock and is moving with velocity v relative to frame K. Then the clock marks K''s **proper time** τ, that is, the time measured in the reference frame at rest with respect to the clock. How does K measure K''s proper time? Consider two inertial frames K' and K in uniform motion with respect to each other with velocity v along the x-axis. A clock at rest in K' measures the proper time $\Delta\tau = t'_B - t'_A$ between two events $A = (t'_A, x')$ and $B = (t'_B, x')$ happening in the same position. How do these coordinates transform in K? According to Lorentz transformations we have:

$$t_A = \frac{t'_A + \dfrac{vx'}{c^2}}{\sqrt{1-\beta^2}},$$
$$t_B = \frac{t'_B + \dfrac{vx'}{c^2}}{\sqrt{1-\beta^2}}.$$

The time interval measured in K is:

$$\Delta t = t_B - t_A = \frac{1}{\sqrt{1-\beta^2}}\left(t'_B + \frac{vx'}{c^2} - t'_A - \frac{vx'}{c^2}\right) = \gamma(t'_B - t'_A) = \gamma\Delta\tau.$$

Hence,

$$\Delta t = \frac{\Delta\tau}{\sqrt{1-\beta^2}} > \Delta\tau.$$

In other words, K says that the clock runs slower than from K''s point of view by the factor $\sqrt{1-v^2}$. A moving clock's frequency is slower. This phenomenon is known as time dilation.

1.3.6 Length Contraction

Consider two reference frames K and K' in relative motion at constant speed v along the x-axis. Suppose K' carries a ruler lying along the x'-axis. The length of the ruler in K' is

$$\lambda_0 = x'_2 - x'_1,$$

and it is called the **proper length** of the ruler, that is, the length measured in the reference frame at rest with the ruler itself. To measure the length of the ruler in K where it is moving at constant speed v, it is necessary to determine the positions of its extremities at the very same moment t.

According to Lorentz transformations, the coordinates of the extremities are,

$$x_1' = \frac{x_1 - vt}{\sqrt{1-\beta^2}},$$
$$x_2' = \frac{x_2 - vt}{\sqrt{1-\beta^2}}.$$

where we imposed that the measurements were made at the same time: $t_1 = t_2 = t$. Hence, the length of the ruler is

$$\lambda = x_2 - x_1 = \sqrt{1-\beta^2}(x_2' - x_1') = \frac{\lambda_0}{\gamma},$$

that is,

$$\lambda = \sqrt{1 - \frac{v^2}{c^2}}\lambda_0 < \lambda_0.$$

This means that an observer in motion with respect to the ruler measures a smaller length than an observer who is at rest with the ruler: K considers the ruler to have shrunk by a factor $\sqrt{1-v^2}$. Length contraction is not an intrinsic property of the body, it is rather a relation between the measurements of two different observers in relative motion with respect to each other. This phenomenon is called length contraction.

1.3.7 The Doppler Effect

Although the speed of light does not depend on whether the observer is moving or not, its *frequency* does. This is the **Doppler effect**. If the source is approaching an observer, the frequency increases and light is shifted towards the ultraviolet side of the spectrum; if the source is receding, the frequency decreases and light is shifted towards the infrared. As a preliminary result, let us show that the *phase* of a plane wave is invariant. Consider a monochromatic plane wave in a reference frame S.

$$\psi = Ae^{i(\omega t - \mathbf{k}\cdot\mathbf{x})},$$

where A is the amplitude, $\mathbf{k}$ is the wave vector, and ω is the pulsation and, for simplicity, we denote with $\cdot$ the Euclidean scalar product. Let's suppose the wave moves in the xy plane; therefore, $\mathbf{k} = (k\cos\theta, k\sin\theta, 0)$; thus, since $ck = 2\pi\nu$, the equation becomes

$$\psi = Ae^{i(\omega t - kx\cos\theta - ky\sin\theta)} = Ae^{2\pi i\nu\left(t - \frac{x\cos\theta + y\sin\theta}{c}\right)}.$$

The phase is defined as follows:

$$\phi = \nu\left(t - \frac{x\cos\theta + y\sin\theta}{c}\right) \equiv \nu\left(t - \frac{l}{c}\right),$$

with $l = x\cos\theta + y\sin\theta$.

The physical meaning of the phase can be explained as follows. Assume that the crest of a wave passes through the origin O of K at $t = 0$. When that crest reaches point $P = (x, y)$, which happens at time $t = l/c$, an observer in P starts counting the crests passing through. At time t, it will have already counted a number of crests equal to the phase $\phi = \nu(t - l/c)$; ν is, in fact, the number of crests passing in a unit of time, and $t - l/c$ is the total counting time.
Let's see what happens in the reference K' which is in uniform motion with respect to K with velocity v along the x-axis. Assume that the two origins O and O' coincide at $t = 0$. If $P' = (x', y')$ is a point in K that coincides with P at time t of K, then the number of crests that an observer in P' counts between time l'/c and time t' is equal to the number of crests counted by an observer in P, that is, the two phases coincide:

$$\phi' = \nu'\left(t' - \frac{l'}{c}\right) = \phi = \nu\left(t - \frac{l}{c}\right)$$

where ν' is the wave frequency in K. Hence, the phase is an invariant quantity. We can explicitly write it as

$$\nu\left(t - \frac{x\cos\theta + y\sin\theta}{c}\right) = \nu'\left(t' - \frac{x'\cos\theta' + y'\sin\theta'}{c}\right).$$

Using Lorentz transformations

$$\begin{aligned} x' &= \frac{x - vt}{\sqrt{1-\beta^2}}, \\ y' &= y, \\ t' &= \frac{t - \dfrac{vx}{c^2}}{\sqrt{1-\beta^2}}, \end{aligned}$$

we can write:

$$\begin{aligned} \phi' &= \nu'\left(\frac{t - \dfrac{vx}{c^2}}{\sqrt{1-\beta^2}} - \frac{1}{c}\frac{x - vt}{\sqrt{1-\beta^2}}\cos\theta' - \frac{1}{c}y\sin\theta'\right) \\ &= \nu'\left(t\frac{1 + \dfrac{v}{c}\cos\theta'}{\sqrt{1-\beta^2}} - x\frac{\cos\theta' + \dfrac{v}{c}}{c\sqrt{1-\beta^2}} - \frac{y}{c}\sin\theta'\right). \end{aligned}$$

This must be equal to

$$\phi = \nu\left(t - \frac{x\cos\theta + y\sin\theta}{c}\right).$$

This implicates the following relations:

$$\nu = \nu' \frac{1 + \frac{v}{c}\cos\theta'}{\sqrt{1-\beta^2}},$$
$$\nu\cos\theta = \nu' \frac{\cos\theta' + \frac{v}{c}}{\sqrt{1-\beta^2}},$$
$$\nu\sin\theta = \nu'\sin\theta'.$$

By inverting the first of the three equations, one obtains the relation between the frequency ν emitted by a source at rest in an inertial frame K and the frequency ν' received by a reference frame K' in uniform motion with velocity v relative to K, that is:

$$\nu' = \nu \frac{\sqrt{1-\beta^2}}{1 + \frac{v}{c}\cos\theta'},$$

where θ' is the angle that the signal forms with the axis x' in K' (and θ was the same angle measured in K). Let us now consider a particular case: the *longitudinal Doppler effect*. In this case, the signal propagates along the direction of relative motion between the source and the receiver, thus $\theta' = 0$. We obtain:

$$\nu' = \nu \frac{\sqrt{1-\beta^2}}{1 + \frac{v}{c}} = \sqrt{\frac{1-\beta}{1+\beta}}.$$

If the receiver is receding from the source ($v > 0$), the frequency it measures is $\nu' < \nu$. On the contrary, if the receiver is approaching the source ($v < 0$), then $\nu > \nu$.

If the signal propagates perpendicularly to the direction of relative motion between the source and the observer, then one experiences the *transverse Doppler effect*. In this case, $\theta' = \pi/2$, hence

$$\nu' = \nu\sqrt{1-\beta^2}.$$

Since $\beta < 1$, $\nu' < \nu$ always. Note that this effect is not expected according to non-relativistic optics. The non relativistic formula is, in fact:

$$\nu' = \frac{\nu}{1 + \frac{v}{c}\cos\theta'},$$

which differs from the relativistic one by the factor $\sqrt{1-\beta^2}$.

1.4 Covariant Formulation of Dynamics

1.4.1 Newton's Laws of Motion

Kinetics is the study of the motion of material objects under the action of forces. Forces cause objects to accelerate, that is, to change their velocity. In spacetime, acceleration makes the worldlines curved. In this way, the physics of forces is tied to the geometry of spacetime. The starting point are Newton's three laws of motion.

1. Every body that is not subject to external forces continues in its state of rest or uniform motion in a straight line.
   This law is called the **principle of inertia**. Inertia is defined as the ability to resist any change in velocity.

2. The change of motion is proportional to the external force applied:

$$\mathbf{F} = \frac{d\mathbf{p}}{dt} = m\mathbf{a},$$

   where $\mathbf{p} = m\mathbf{v}$ is the momentum, and m is the *inertial mass* of the body.

3. To every action, there is an equal and opposite reaction. That is, if a body A imposes a force $\mathbf{F}_{AB}$ on body B, then B imposes a force $\mathbf{F}_{BA} = -\mathbf{F}_{AB}$ on A.

Newton's laws do not always correspond to reality. For instance, the definition of inertial mass is problematic: if m is assumed to be constant, then a constant force would eventually push the body beyond the speed of light. To avoid this, we need to accept the fact that the mass of a body grows larger as its velocity increases. As the velocity approaches c, the mass m should diverge to infinity:

$$m \xrightarrow{v \to c} \infty.$$

In this way, it becomes increasingly difficult to accelerate the body, and the velocity limit is being respected. Furthermore, studies showed that this law is not valid in the limit of ultrarelativistic velocities[3].
Another difficulty that arises involves the frames of reference. For example, an object near the surface of the earth that has no visible forces pushing it accelerates downward. To solve this conflict, we can assume the existence of an "invisible" force for each unexplained motion. The force responsible for the free fall of objects is gravity. Another way to solve the problem is to choose a coordinate system in which such forces disappear. Are there any coordinate systems that can eliminate gravity? For example, in a space station orbiting around earth, objects float without falling to the ground. In this case,

[3] A velocity is said to be *ultrarelativistic* when it is close to the speed of light c.

the law of inertia holds. Thus, whether or not Newton's laws of motion hold, depends on the reference frame. Those frames in which Newton's laws do hold, are called **inertial frames**.

1.4.2 The Fundamental Law of Dynamics

Experimental studies have shown that Newton's second law of motion

$$\frac{d(m\mathbf{v})}{dt} = \mathbf{F}$$

is not valid at high velocities, and the deviations grow larger as the velocity approaches c. Theoretically speaking, this law is in conflict with the assumption that c is the limit velocity. Experiments show that the true law of motion for a massive particle is

$$\frac{d(m\gamma\mathbf{v})}{dt} = \frac{d}{dt}\left(\frac{m\mathbf{v}}{\sqrt{1 - v^2/c^2}}\right) = \mathbf{F}, \tag{1.1}$$

which is called the **Minkowski equation**. Here, m is the **proper mass** of a body, that is, the mass measured in the reference frame at rest with the body. It is immediate to verify that if $v \ll c$, then $\gamma \to 1$, so Minkowski's equation simply tends to Newton's. The first essential difference between these two laws is the presence of the Lorentz factor γ, which lessens the velocity increment as $\mathbf{v}$ approaches c. The second difference is how the force transforms from one inertial frame to another. In order for $\mathbf{F} = \gamma m\mathbf{a}$ to be consistent with the relativity principle, $\mathbf{F}$ needs to change in the same way as the first member of the equation ($\gamma m\mathbf{a}$). Lorentz maps act on vectors in spacetime, so if we want the force to transform properly, we first have to express it as a spacetime vector. There is a natural way to do this, starting with velocity and momentum in the full $(1 + 3)$-dimensional spacetime.
The best way to do so, is to write all these quantities as 4-vectors in their covariant form. This approach is called **covariant formulation of dynamics**. We say that an equation is covariant if, with respect to a given transformation, both sides change in the same way under that transformation.

4-velocity

As we saw in Section 1.2.2, any event in $\mathcal{M}$ can be identified by a 4-vector called **4-position**: $x^\mu = (ct, x, y, z)$. The 4-velocity of a body whose motion is described by $x^\mu(t) = (ct, x(t), y(t), z(t))$ is simply given by

$$\mathbf{v} = \frac{dx^\mu}{dt} = \left(t, \frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt}\right).$$

However, this formulation loses all the information about the nature of the object $\mathbf{v}$: dx^μ is a 4-vector, but dt is not, so we do not immediately know the transformation of $\mathbf{v}$ from one frame to another, like we did for x^μ (which is instead a 4-vector and transforms according to the Lorentz transformations).
A solution to this problem is to parametrize the position of a particle with ds instead of dt, which is a scalar. Let's now see how the two parametrizations are connected. The quantity ds is the infinitesimal interval defined in Section 1.1 as

$$ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2.$$

In the frame of reference K' at rest with the particle, $dx' = dy' = dz' = 0$, hence, we have

$$ds'^2 = c^2d\tau^2,$$

where $\tau \equiv t'$ is the proper time. But ds is an invariant quantity, so it must have the same value in any reference frame:

$$ds'^2 = ds^2 \Longrightarrow c^2d\tau^2 = c^2dt^2 - dx^2 - dy^2 - dz^2.$$

The relation between ds and dt is given by

$$\begin{aligned} ds = \sqrt{c^2dt^2 - dx^2 - dy^2 - dz^2} &= \sqrt{c^2dt^2\left(1 - \frac{1}{c^2}\frac{dx^2 + dy^2 + dz^2}{dt^2}\right)} \\ &= \sqrt{c^2dt^2\left(1 - \frac{v^2}{c^2}\right)} = \frac{c}{\gamma}dt. \end{aligned}$$

so we have found that $ds = \dfrac{c}{\gamma}dt$. According to this parametrization, the 4-velocity of a particle becomes

$$u^\mu = \frac{dx^\mu}{ds}.$$

Since ds is a scalar and x^μ is a 4-vector, u^μ must be a 4-vector by construction. Let's see one by one the components of the 4-velocity.

$$\begin{aligned} u^0 &= \frac{dx^0}{ds} = \frac{dx^0}{dt}\frac{dt}{ds} = c\frac{\gamma}{c} = \gamma, \\ u^i &= \frac{dx^i}{ds} = \frac{dx^i}{dt}\frac{dt}{ds} = \frac{\gamma}{c}v^i. \end{aligned}$$

Hence, the 4-velocity is given by

$$u^\mu = \left(\gamma(v), \frac{\gamma(v)}{c}\mathbf{v}\right).$$

In the reference frame at rest with a body, the 4-velocity is

$$u^\mu = (\gamma(0), \mathbf{0}) = (1, \mathbf{0}),$$

and its norm, which is invariant, is equal to

$$u^\mu u_\mu = 1.$$

If we derive the 4-velocity with respect to ds, we obtain the 4-acceleration

$$w^\mu = \frac{du^\mu}{ds}.$$

4-momentum

Starting from the 4-velocity $u^\mu = \left(\gamma, \frac{\gamma}{c}\mathbf{v}\right)$, we can define the 4-momentum of a massive particle as

$$p^\mu = mcu^\mu,$$

which is, by construction, a 4-vector. Its components are

$$p^0 = mcu^0 = mc\gamma,$$
$$p^i = mcu^i = m\gamma v^i.$$

It is easy to show that the momentum $\mathbf{p}$ and the energy E of a particle are the components of a 4-vector.
The formula for the relativistic kinetic energy is

$$E = \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}} = m\gamma c^2,$$

which is exactly $p^0 c$. Hence,

$$p^\mu = \left(\frac{E}{c}, \mathbf{p}\right),$$

where $\mathbf{p}$ is defined as

$$\mathbf{p} = \frac{m\mathbf{v}}{\sqrt{1 - \frac{v^2}{c^2}}} = \gamma(\mathbf{v})m\mathbf{v}.$$

In the frame of reference at rest with the body, $\gamma = 1$ and $\mathbf{v} = \mathbf{0}$, so, the 4-momentum is $p^\mu = (mc, \mathbf{0})$. The square norm of p^μ is therefore

$$p^\mu p_\mu = m^2 c^2,$$

which must be equal to the norm computed in any other reference frame

$$p^\mu p_\mu = \frac{E^2}{c^2} - \mathbf{p}^2.$$

Hence, we find the notorious mass-energy relation

$$E^2 = p^2c^2 + m^2c^4.$$

Equations of Motion

The covariant equations of motion of a massive particle can be deduced starting from their non covariant form

$$\mathbf{F} = \frac{d\mathbf{p}}{dt},$$

and the 4-momentum

$$p^\mu = \left(\frac{E}{c}, \mathbf{p}\right), \qquad p^\mu p_\mu = m^2c^2 = \frac{E^2}{c^2} - \mathbf{p}^2.$$

The covariant form of the equation of motion must be

$$\frac{dp^\mu}{ds} = \mathcal{F}^\mu,$$

where $\mathcal{F}^\mu$ is a 4-vector called 4-force. The previous equation is called **Minkowski covariant equation**, and it is simply Equation (1.1) in its covariant formulation. In terms of the 4-velocity it becomes

$$mc\frac{du^\mu}{ds} = \mathcal{F}^\mu.$$

To deduce $\mathcal{F}^\mu$'s components, let's study the components of the other member of the equation

$$\frac{dp^\mu}{ds} = \left(\frac{d}{ds}\left(\frac{E}{c}\right), \frac{d\mathbf{p}}{ds}\right) = \frac{dt}{ds}\left(\frac{d}{dt}\left(\frac{E}{c}\right), \frac{d\mathbf{p}}{dt}\right) = \left(\frac{\gamma}{c^2}\frac{dE}{dt}, \frac{\gamma}{c}\frac{d\mathbf{p}}{dt}\right).$$

The non covariant equation

$$\frac{d\mathbf{p}}{dt} = \mathbf{F},$$

and the law

$$\frac{dE}{dt} = \mathbf{F} \cdot \mathbf{v},$$

allow us to rewrite the 4-force as

$$\mathcal{F}^\mu = \left(\frac{\gamma}{c^2}\mathbf{F} \cdot \mathbf{v}, \frac{\gamma}{c}\mathbf{F}\right).$$

The peculiar difference between the relativistic 4-force and the Newtonian force is the dependence of $\mathcal{F}^\mu$ on the velocity of the particle. This is crucial for the validity of Einstein's postulates: if the force did not depend on the velocity, then a constant force would produce an indefinite increase in v, which would eventually be pushed beyond the speed limit c.

Chapter 2

Regular Surfaces

In this chapter, we aim to define the notion of a regular surface in $\mathbb{R}^3$. Roughly speaking, surfaces are obtained by arranging and deforming pieces of planes in such a way that the resulting figure has no sharp edges, points, or self-intersections. We start Section 2.1 by introducing some elementary concepts of the theory of curves, such as curvature, parametrization, and the definition of the tangent vector to a curve. These definitions will be useful later in this chapter to fully understand surfaces and their properties. Then, we move on to Section 2.2 where we define regular surfaces, providing the reader with some criteria that should help when trying to decide whether a given subset of $\mathbb{R}^3$ is a regular surface or not. In Section 2.3, we begin to study the intrinsic geometry of the surface through the introduction of the metric tensor (or first fundamental form), a natural instrument to treat metric aspects like lengths, angles or areas. We conclude this section with the notion of orientability. After that, we open Section 2.4 by extending the concept of curvature to surfaces, followed by some relevant definitions (the Gauss map, principal curvatures and directions, Gaussian curvature, mean curvature). Then, in Section 2.5, we start the study of intrinsic geometry, that is, the study of those features which can be deduced directly from the metric without reference to the external embedding. A pivotal result of this section is Gauss's Theorema Egregium, which shows that the Gaussian curvature is actually an intrinsic property of surfaces. This opens the way for a more abstract theory of intrinsic differential geometry in which a surface patch, and likewise, the spacetime frame of an arbitrary observer, is simply an open set provided with a suitable metric. Finally, Section 2.6 is dedicated to the study of geodesics. First, we introduce some definitions, like the covariant derivative of a vector, and the concept of a parallel vector field, and then we conclude the chapter with the definition of geodesic curve. Geodesics are the generalization of "straight lines" on a curved surface.

This geometric framework, in particular the concepts of intrinsic geometry, geodesics, and curvature, provides the basic mathematical tools for transitioning from the flat spacetime of special relativity to the curved one of general relativity.

The leading source for the material presented in this chapter is [dC76], with occasional

reference to [Cal00].

2.1 Curves and Curvature

Before introducing the definition of a regular surface, let us recall some notions on curves in $\mathbb{R}^3$ that will be useful later on.
As we saw in the previous chapter, bodies that move under the effect of an external force experience changes in velocity, and since velocity is the slope of the worldline of a body, changes in velocity imply a curved worldline.
Let's now talk about curves in the ordinary Euclidean plane.

Definition 2.1.1. A parametrized differentiable curve is a differentiable map

$$\begin{aligned} &\alpha : I \to \mathbb{R}^3, \\ &t \mapsto \alpha(t) = (x(t), y(t), z(t)). \end{aligned}$$

of an open interval $I = (a, b)$ of the real line $\mathbb{R}$ into $\mathbb{R}^3$. By differentiable we mean that all the components of α have continuous derivatives up to a desired order.

A curve defined on a *closed* interval $[a, b]$ is a parametrized differential curve if it is continuous and it is differentiable on (a, b).

From now on, we will refer to parametrized differentiable curves simply as "curves".

As t moves along I, the point $\alpha(t)$ traces out a path in the space. That is, α is a correspondence that maps each $t \in I$ into a point $\alpha(t) = (x(t), y(t), z(t))$. The "coordinate" t allows us to label points along the path, and it is therefore called **parameter**.

Definition 2.1.2. The first derivative of a parametrized differentiable curve $\alpha : I \to \mathbb{R}^3$

$$\frac{d\alpha}{dt} = \alpha'(t) = (x'(t), y'(t), z'(t)) \in \mathbb{R}^3$$

is called **tangent velocity vector** and it is the tangent vector to the path at point $\alpha(t)$, at least when $\alpha'(t) \neq 0$.

Definition 2.1.3. A parametrized differentiable curve $\alpha : I \to \mathbb{R}^3$ is said to be **regular** if $\alpha'(t) \neq 0$, $\forall t \in I$.

The condition of regularity is very important to measure the length of a curve. In fact, if $\alpha'(t_0) \neq 0$, according to Taylor's theorem

$$\alpha(t_0 + \Delta t) \approx \alpha(t_0) + \alpha'(t_0)\Delta t$$

so the point $\alpha(t_0+\Delta t)$ is very close to the continuation of $\alpha(t_0)$ at a distance of $|\alpha'(t_0)|\Delta t$ along the tangent vector. So the straight-line distance

$$|\Delta\alpha| = |\alpha(t_0 + \Delta t) - \alpha(t_0)|$$

is well approximated by the length $|\alpha'(t_0)\Delta t|$ of the vector $\alpha'(t_0)\Delta t$. If Δt is sufficiently small, this length is a good approximation to the length of the curved arc from $\alpha(t_0)$ to $\alpha(t_0 + \Delta t)$. This is not possible if the curve is not regular. To measure the length of the entire curve C, we need to partition the interval $I = (a, b)$,

$$a = t_0 < t_1 < t_2 < ... < t_n < t_{n+1} = b,$$

in such a way that each segment $\Delta t_j = t_{j+1} - t_j$ is small enough for

$$|\alpha(t_{j+1}) - \alpha(t_j)| \approx |\alpha'(t_j)|\Delta t_j$$

to be a good approximation. Then the length of the entire curve C is approximately

$$\sum_{j=1}^{n} |\alpha'(t_j)|\Delta t_j.$$

As the partition gets finer and finer while $n \to \infty$, the sum approaches the integral

$$\int_a^b |\alpha'(t)|\, dt.$$

The above reasoning motivates the following definition.

Definition 2.1.4. Let $\alpha : I = [a, b] \to \mathbb{R}^3$ be a parametrized curve C. Then the **arc length** of C is by definition

$$s(t) = \int_a^b |\alpha'(t)|\, dt,$$

where $|\alpha'(t)| = \sqrt{(x'(t))^2 + (y'(t))^2 + (z'(t))^2}$ is the length of the vector $\alpha'(t)$. Since $\alpha'(t) \neq 0$, the arc length is a differentiable function of t, and $\dfrac{ds}{dt} = |\alpha'(t)|$.

It can happen that the parameter t is already the arc length measured from some point. In this case

$$\frac{ds}{dt} = |\alpha'(t)| = 1,$$

that is, the velocity vector has constant unit length. If we consider a curve α parametrized by arc length s, then, since its tangent vector $|\alpha'(s)|$ has unit length, the norm $|\alpha''(s)|$ of the second derivative measures the rate of change of the angle which neighboring tangents make with the tangent at s. In other words, $|\alpha''(s)|$ measures how rapidly the curves "pulls away"" from the tangent line at s in a neighborhood of S. More formally:

Definition 2.1.5. Let $\alpha : I \to \mathbb{R}^3$ be a curve parametrized by arc length $s \in I$. The number $|\alpha''(s)| = k(s)$ is called the *curvature* of α at s.

Moreover, $\alpha''(s)$ is orthogonal to $\alpha'(s)$. Indeed, by differentiating

$$|\alpha'(s)|^2 = \langle \alpha'(s), \alpha'(s) \rangle = 1,$$

one finds that $\langle \alpha''(s), \alpha'(s) \rangle = 0$. Thus, any vector $n(s)$ in the direction $\alpha''(s)$ is called *normal vector* at s.

2.2 Regular Surfaces

We now wish to introduce the formal definition of a regular surface and its parametrization, but before doing that, let us recall the definition of **differential** of a function.

Definition 2.2.1. Let $F : U \subset \mathbb{R}^n \to \mathbb{R}^m$ be a differentiable map. To each $p \in U$ we associate a linear map $dF_p : \mathbb{R}^n \to \mathbb{R}^m$, which is called the *differential* of F at p and is defined as follows. Let $w \in \mathbb{R}^n$ and let $\alpha : (-\epsilon, \epsilon) \to U$ be a differentiable curve such that $\alpha(0) = p$ and $\alpha'(0) = w$. By the chain rule, the curve $\beta = (F \circ \alpha) : (-\epsilon, \epsilon) \to \mathbb{R}^m$ is also differentiable. Then

$$dF_p(w) = \beta'(0),$$

and it is called the differential of F at point p.

The matrix of $dF_p : \mathbb{R}^n \to \mathbb{R}^m$ in the canonical bases of $\mathbb{R}^n$ and $\mathbb{R}^m$, that is the matrix

$$\frac{\partial f_i}{\partial x_j}, \qquad i = 1, 2, .., m, \quad j = 1, 2, .., n$$

is called the *Jacobian matrix* at p. When $n = m$, this is a square matrix, and its determinant is called the *Jacobian determinant*; it is usually denoted by

$$\det\left(\frac{\partial f_i}{\partial x_j}\right) = \frac{\partial(f_1, ..., f_n)}{\partial(x_1, ..., x_n)}.$$

Definition 2.2.2. Given a differentiable map $F : U \subset \mathbb{R}^n \to \mathbb{R}^m$ defined in an open set U of $\mathbb{R}^n$, we say that $p \in U$ is a critical point of F if the differential $dF_p : \mathbb{R}^n \to \mathbb{R}^m$ is not a surjective mapping. The image $F(p) \in \mathbb{R}^m$ of a critical point is called a *critical value* of F. A point of $\mathbb{R}^m$ which is not a critical value is called a *regular value* of F.

For example, given a 1-D differentiable function $f : U \subset \mathbb{R} \to \mathbb{R}$, a point $x_0 \in U$ is critical if $f'(x_0) = 0$, that is, if the differential df_{x_0} carries all the vectors in the domain (here $\mathbb{R}$) to the zero vector. Notice that any point $a \notin f(U)$ is trivially a regular value of f.

For clarity in future work, we should also recall the formulation of the inverse function theorem and the mean value theorem.

Theorem 2.2.1 (**Inverse function theorem**). Let $F : U \subset \mathbb{R}^n \to \mathbb{R}^n$ be a differentiable mapping, and suppose that at $p \in U$ the differential $dF_p : \mathbb{R}^n \to \mathbb{R}^n$ is an isomorphism. Then there exists a neighborhood V of p in U and a neighborhood W of $F(p)$ in $\mathbb{R}^n$ such that $F : V \to W$ has a differentiable inverse $F^{-1} : W \to V$.

Theorem 2.2.2 (**Mean Value Theorem for Definite Integrals**). Let $f : [a, b] \to \mathbb{R}$ be a continuous function. Then there exists $c \in [a, b]$ such that

$$\int_a^b f(x)\,dx = f(c)(b - a).$$

We can now introduce the notion of regular surface.

Definition 2.2.3. A subset $S \subset \mathbb{R}^3$ is a **regular surface** if, for each $p \in S$, there exists a neighborhood V in $\mathbb{R}^3$ and a map $\mathbf{x} : U \to V \cap S$ of an open set $U \subset \mathbb{R}^2$ onto $V \cap S \subset \mathbb{R}^3$ such that

1. $\mathbf{x}$ is **differentiable**. This means that if we write

   $$x(u, v) = (x(u, v), y(u, v), z(u, v)), \qquad (u, v) \in U,$$

   the functions $x(u, v)$, $y(u, v)$, $z(u, v)$ have continuous partial derivatives of all orders in U.

2. $\mathbf{x}$ is a **homeomorphism**. Since $\mathbf{x}$ is continuous by condition 1., this means that $\mathbf{x}$ has an inverse $\mathbf{x}^{-1} : V \cap S \to U$ which is continuous; that is, $\mathbf{x}^{-1}$ is the restriction of a continuous map $F : W \subset \mathbb{R}^3 \to \mathbb{R}^2$ defined on an open set W containing $V \cap S$.

3. (**the regularity condition**) For each $q \in U$, the differential $d\mathbf{x}_q : \mathbb{R}^2 \to \mathbb{R}^3$ is **one-to-one**.

The mapping $\mathbf{x}$ is called *parametrization* or *system of local coordinates* in a neighborhood of p. The neighborhood $V \cap S$ of p in S is called a *coordinate neighborhood*.

The following proposition shows the relation between a regular surface and the graph of a function $z = f(x, y)$.

Proposition 2.2.1. If $f : U \to \mathbb{R}$ is a differentiable function in an open set $U \subset \mathbb{R}^2$, then the graph of f, that is, the subset of $\mathbb{R}^3$ given by $(x, y, f(x, y))$ for $(x, y) \in U$, is a regular surface.

To set an example, let us consider the following function

$$f(x, y) = x^2 + y^2.$$

Its graph is given by the subset (x, y, x^2+y^2) of $\mathbb{R}^3$, that is the regular surface $z = x^2+y^2$ shown in Figure 2.1.

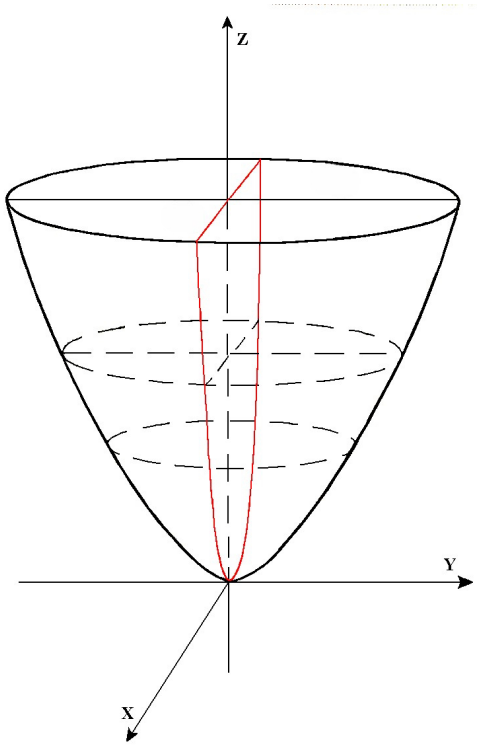

Figure 2.1: Graph of the function $z = x^2 + y^2$ representing an elliptic paraboloid.

As shown in the following proposition, another way to obtain surfaces is to consider subsets of the form $f(x, y, z) = \text{const}$.

Proposition 2.2.2. If $f : U \subset \mathbb{R}^3 \to \mathbb{R}$ is a differentiable function and $a \in f(U)$ is a regular value of f, then $f^{-1}(a)$ is a regular surface in $\mathbb{R}^3$.

Proof. Let $p = (x_0, y_0, z_0) \in f^{-1}(a)$. Since a is a regular value of f, it is always possible to assume (by renaming the axis if necessary) that $f_z = \dfrac{\partial f}{\partial z} \neq 0$ at p, that is,

$$\frac{\partial f}{\partial z}(p) = \frac{\partial f}{\partial z}(x_0, y_0, z_0) \neq 0.$$

Let's define a map $F : U \subset \mathbb{R}^3 \to \mathbb{R}^3$ by

$$F(x, y, z) = (x, y, f(x, y, z)),$$

and indicate by (u, v, t) the coordinates of a point in $\mathbb{R}^3$ where F takes its values. So we have:

$$\begin{cases} u = x, \\ v = y, \\ t = f(x, y, z) \end{cases}$$

The differential of F at p is given by

$$dF_p = \begin{pmatrix} \frac{\partial F_1}{\partial x} & \frac{\partial F_1}{\partial y} & \frac{\partial F_1}{\partial z} \\ \frac{\partial F_2}{\partial x} & \frac{\partial F_2}{\partial y} & \frac{\partial F_2}{\partial z} \\ \frac{\partial F_3}{\partial x} & \frac{\partial F_3}{\partial y} & \frac{\partial F_3}{\partial z} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ f_x & f_y & f_z \end{pmatrix},$$

hence, $\det(dF_p) \neq 0$. We can therefore apply Theorem 2.2.1 which guarantees the existence of neighborhoods V of p and W of $F(p)$ such that $F : V \to W$ is invertible and the inverse $F^{-1} : W \to V$ is differentiable. It follows that the coordinate functions of F^{-1}, that is the functions

$$x = u, \qquad y = v, \qquad z = f^{-1}(x, y, t) = g(u, v, t) \qquad (u, v, t) \in W,$$

are all differentiable. In particular, by fixing a particular value of t, let's say $t = a$, we get $z = g(u, v, a) = h(x, y)$ which is a differentiable function defined in the projection of V onto the xy plane. The following figure should clarify ideas.

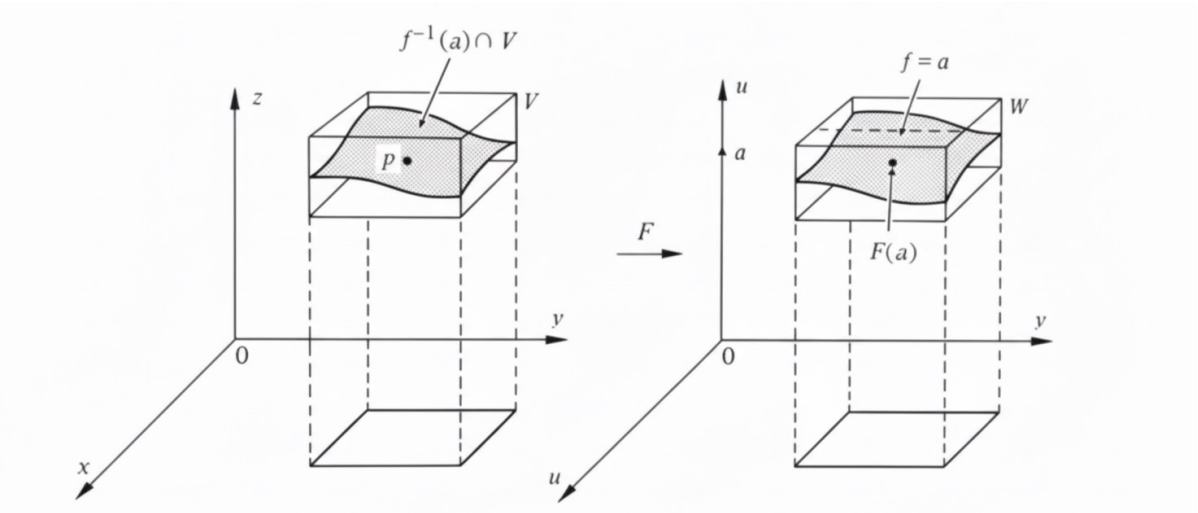

Figure 2.2: From [dC76, page 60].

Since

$$F(f^{-1}(a) \cap V) = W \cap \{(u, v, t) : t = a\},$$

we conclude that the graph of h is $f^{-1}(a) \cap V$, and by Proposition 2.2.1 it is a coordinate neighborhood of p. Therefore, every $p \in f^{-1}(a)$ can be covered by a coordinate neighborhood. But this is the definition of regular surface provided in Definition 2.2.3. So $f^{-1}(a)$ is a regular surface. □

The proof consists essentially of using the inverse function theorem "to solve for z" in the equation $f(x, y, z) = a$, which can be done in a neighborhood of p if $f_z(p) \neq 0$. In this way, we were able to express one variable as a function of the others. It can be shown that this process is permitted for every regular surface.

Example 2.2.1. The right cylinder

$$x^2 + y^2 = r^2$$

is a regular surface (see Figure 2.3). In fact, the set $f^{-1}(0)$ where

$$f(x, y, z) = x^2 + y^2 - r^2$$

is a regular function and 0 is a regular value of f. This follows from the fact that the partial derivatives $f_x = 2x$, $f_y = 2y$, and $f_z = 0$ vanish simultaneously in $(0, 0, 0)$, which does not belong to $f^{-1}(0)$ provided $r \neq 0$.

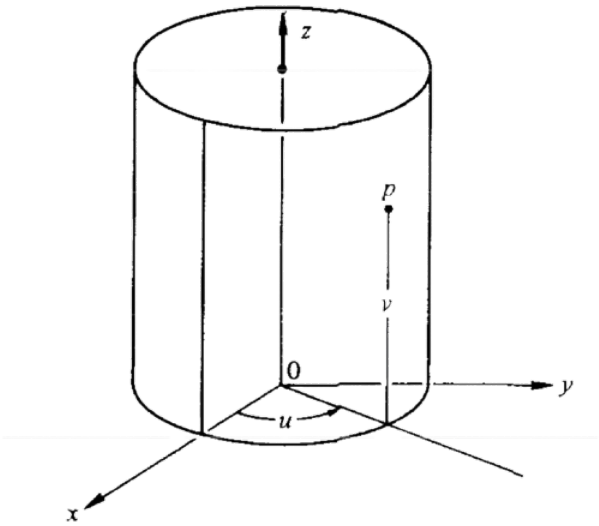

Figure 2.3: Graph of the function $x^2 + y^2 = r^2$, $r \neq 0$ representing a right cylinder. From [dC76, page 93].

All the points of this cylinder, except the ones belonging to the straight line

$$\begin{cases} x = r, \\ y = 0, \\ z = t, \end{cases}$$

admit the parametrization $\mathbf{x} : U \to \mathbb{R}^3$ where

$$\begin{aligned} &\mathbf{x}(u, v) = (r \cos u, r \sin u, v), \quad r > 0, \\ &U = \{(u, v) \in \mathbb{R}^2 : \quad 0 < u < 2\pi, \quad -\infty < v < \infty\}. \end{aligned}$$

Example 2.2.2. The hyperboloid of one sheet

$$x^2 + y^2 - z^2 = 1$$

is a regular surface (see Figure 2.4). In fact, it is given by $S = f^{-1}(0)$, where 0 is a regular value of

$$f(x, y, z) = x^2 + y^2 - z^2 - 1.$$

This follows from the fact that the partial derivatives $f_x = 2x$, $f_y = 2y$ and $f_z = -2z$ vanish simultaneously in $(0, 0, 0)$ which does not belong to S.

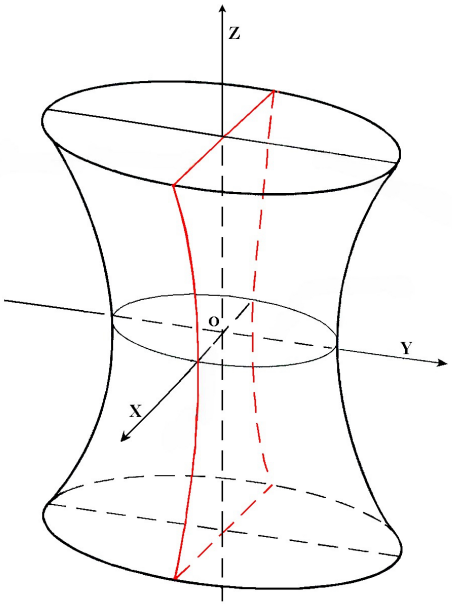

Figure 2.4: Graph of the function $x^2 + y^2 - z^2 = 1$ representing a hyperboloid of one sheet.

All the points of the hyperboloid of one sheet except the ones belonging to the line

$$\begin{cases} x = -\cosh u, \\ y = 0, \\ z = \sinh u \end{cases}$$

admit the following parametrization $\mathbf{x} : U \to \mathbb{R}^3$ where

$$\begin{aligned} &\mathbf{x}(u, v) = (\cosh(u)\cos(v), \cosh(u)\sin(v), \sinh(u)), \\ &U = \{(u, v) \in \mathbb{R}^2 : \quad -\infty < u < \infty, \quad -\pi < v < \pi\}. \end{aligned}$$

Proposition 2.2.1 states that the graph of a differentiable function is a regular surface. The following proposition says, at least locally, the inverse: any regular surface is locally the graph of a differentiable function.

Proposition 2.2.3. Let $S \subset \mathbb{R}^3$ be a regular surface and let $p \in S$. Then there exists a neighborhood V of p in S such that V is the graph of a differentiable function which has one of the following three forms:

$$z = f(x, y), \quad y = g(x, z), \quad x = h(y, z).$$

Proof. Let $\mathbf{x} : U \subset \mathbb{R}^2 \to S \subset \mathbb{R}^3$ be a parametrization of S in p, so that $\mathbf{x}(u, v) = (x(u, v), y(u, v), z(u, v)) \in U$. By condition 3 of Definition 2.2.3, one of the following Jacobian determinants

$$\frac{\partial(x, y)}{\partial(u, v)}, \qquad \frac{\partial(y, z)}{\partial(u, v)}, \qquad \frac{\partial(x, z)}{\partial(u, v)}$$

is nonzero at $\mathbf{x}^{-1}(p) = q$. Suppose that $\dfrac{\partial(x, y)}{\partial(u, v)}(q) \neq 0$. Let's consider the map $\pi \circ \mathbf{x} : U \subset \mathbb{R}^2 \to \mathbb{R}^2$, where π is the projection $\pi(x, y, z) = (x, y)$. Then $\pi \circ \mathbf{x}(u, v) = (x(u, v), y(u, v))$, and since $\dfrac{\partial(x, y)}{\partial(u, v)}$ is exactly its differential and it is nonzero at q, we can apply the inverse function theorem to guarantee the existence of the neighborhoods $V_1 \subset U$ of q and V_2 of $(\pi \circ \mathbf{x})(q)$ such that $\pi \circ \mathbf{x}$ maps V_1 diffeomorphically onto V_2 (this is what the theorem guarantees). It follows that π restricted to $\mathbf{x}(V_1) = V$ is one-to-one and that there is a differentiable inverse $(\pi \circ \mathbf{x})^{-1} : V_2 \to V_1$. Since $\mathbf{x}$ is a homeomorphism, V is a neighborhood of p in S. Now, if we compose the map $(\pi \circ \mathbf{x})^{-1} : (x, y) \to (u(x, y), v(x, y))$ with the function $(u, v) \to z(u, v)$, we find that V is the graph of the differentiable function $z = z(u(x, y), v(x, y)) = f(x, y)$. This settles the first case because we have found a way to express z as a function of x and y. The other cases can be treated in the same way. □

What we have done in this proof is to show that, since z is a function of (u, v), and (u, v) is a function of (x, y), then z is also a function of (x, y). To do so, we need a projection function π to build a diffeomorphism on which we can apply the inverse function theorem. It is this auxiliary invertible function $\pi \circ \mathbf{x}$ that allows us to say that (x, y) is a function of (u, v), and not just vice versa.

According to the definition of regular surface, each point p of a regular surface S belongs to a coordinate neighborhood. The points of such a neighborhood are characterized by their coordinates which should also define their local properties. However, we have seen that there may be more system of coordinates characterizing a point, so, in order for the definition to make sense, it is necessary that such properties do not depend on the chosen system of coordinates.
It must be shown that when p belongs to two different coordinate neighborhoods with parametrizations (u, v) and (ξ, η), it is possible to pass from one to the other by a differentiable function. The following proposition shows that this is true.

Proposition 2.2.4 (**Change of Parameters**)**.** Let p be a point of a regular surface S, and let $\mathbf{x} : U \subset \mathbb{R}^2 \to S$, $\mathbf{y} : V \subset \mathbb{R}^2 \to S$ be two parametrizations of S such that $p \in \mathbf{x}(U) \cap \mathbf{y}(V) = W$. Then the change of coordinates

$$h = \mathbf{x}^{-1} \circ \mathbf{y} : \mathbf{y}^{-1}(W) \to \mathbf{x}^{-1}(W)$$

is a diffeomorphism; that is, h is differentiable and has a differentiable inverse h^{-1}.

Proof. Let $r \in \mathbf{y}^{-1}(W)$ and set $q = h(r)$. Since $\mathbf{x}(u, v) = (x(u, v), y(u, v), z(u, v))$ is a parametrization, we can assume that

$$\frac{\partial(x, y)}{\partial(u, v)}(q) \neq 0.$$

We extend $\mathbf{x}$ to a map $F : U \times \mathbb{R} \to \mathbb{R}^3$ defined by

$$F(u, v, t) = (x(u, v), y(u, v), z(u, v) + t), \qquad (u, v) \in U, \quad t \in \mathbb{R}.$$

Geometrically, F maps a vertical cylinder C over U into a "vertical cylinder" over $\mathbf{x}(U)$ by mapping each section of C with height t into the surface $\mathbf{x}(u, v) + te_3$, where e_3 is the unit vector of the z-axis. It is clear the F is differentiable, and that the restriction of F to $t = 0$ is

$$F_{|U\times 0} = \mathbf{x}.$$

Moreover, computing the determinant of the differential dF_q, we obtain

$$\det(dF_q) = \det \begin{pmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} & 0 \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} & 0 \\ \dfrac{\partial x}{\partial u} & \dfrac{\partial z}{\partial v} & 1 \end{pmatrix} = \frac{\partial(x, y)}{\partial(u, v)} \neq 0.$$

So we can apply the inverse function theorem (Therorem 2.2.1) to F, which guarantees the existence of a neighborhood M of $\mathbf{x}(q)$ in $\mathbb{R}^3$ such that F^{-1} exists and is differentiable in M.
By the continuity of $\mathbf{y}$, there exists a neighborhood N of r in V such that $\mathbf{y}(N) \subset M$. Notice that, restricted to N, the function $h = F^{-1} \circ \mathbf{y}$, so it is a composition of differentiable maps. Thus, h is differentiable at r. Since r is arbitrary, h is differentiable on $\mathbf{y}^{-1}(W)$. This very same argument can be used to show that the map h^{-1} is differentiable, so we conclude that h is a diffeomorphism. □

In other words, if $\mathbf{x}$ and $\mathbf{y}$ are two parametrizations given by

$$\mathbf{x}(u, v) = (x(u, v), y(u, v), z(u, v)), \qquad (u, v) \in U,$$
$$\mathbf{y}(\xi, \eta) = (x(\xi, \eta), y(\xi, \eta), z(\xi, \eta)), \qquad (\xi, \eta) \in V,$$

then the change of coordinates h given by

$$u = u(\xi, \eta), \qquad v = v(\xi, \eta), \qquad (\xi, \eta) \in y^{-1}(W),$$

has the property that the functions u and v have continuous partial derivatives of all orders, and the map h can be inverted, giving

$$\xi = \xi(u, v), \qquad \eta = \eta(u, v), \qquad (u, v) \in \mathbf{x}^{-1}(W),$$

where the functions ξ and η also have partial derivatives of all orders. Since

$$\frac{\partial(u, v)}{\partial(\xi, \eta)} \cdot \frac{\partial(\xi, \eta)}{\partial(u, v)} = 1,$$

this implies that the Jacobian determinants of the change of coordinates for both h and h^{-1} are nonzero everywhere.

The reason we did not deduce the differentiability of $h = \mathbf{x}^{-1} \circ \mathbf{y}$ by the chain rule is that we do not yet know what is meant by "differentiable function" on S since we have only defined differentiability for an open set of $\mathbb{R}^n$.

2.3 The Metric

2.3.1 Geometry in the Tangent Plane

We can use the notion of tangent vector to a curve (see Definition 2.1.2) to define a tangent vector to a surface: by tangent vector to a regular surface S at point $p \in S$, we mean the tangent vector $\alpha'(0)$ to a differentiable parametrized curve $\alpha : (-\epsilon, \epsilon) \to S$ with $\alpha(0) = p$.

Let T_pS denote the set of all vectors in $\mathbb{R}^3$ that are tangent to a surface S at the point p. We call T_pS the **tangent plane to S at p.**

The following proposition shows that the tangent plane is indeed a plane, since it coincides with the vector space spanned by the image of the differential of a parametrization for that surface.

Proposition 2.3.1. Let $\mathbf{x} : U \in \mathbb{R}^2 \to S$ be a parametrization of a regular surface S, and let $q \in U$. The 2-dimensional vector subspace

$$d\mathbf{x}_q(\mathbb{R}^2) \subset \mathbb{R}^3,$$

coincides with the set of tangent vectors to S at $\mathbf{x}(q)$.

Proof. Since $\mathbf{x}$ is a parametrization, it is differentiable, that is, its differential $d\mathbf{x}_q$ has full rank $\forall q \in U$. So $d\mathbf{x}_q$ is one-to-one, at least in a neighborhood V of q. This means that the only vector mapped onto the null vector by $d\mathbf{x}_q$ is $\mathbf{0}$ itself:

$$d\mathbf{x}_q(\mathbf{0}) = \mathbf{0}.$$

Hence, the kernel of the differential has dimension zero. According to the dimension theorem, the dimension of the image of a linear application is equal to the dimension of its kernel plus the dimension of the domain.

$$f : A \to B \Longrightarrow \dim(f(A)) = \dim(A) + \dim(\ker(f).$$

In this case, since $\dim(\ker(f)) = 0$, the dimension of the image must be the same as the dimension of the domain:

$$\dim(d\mathbf{x}_q(\mathbb{R}^2)) = \dim(\mathbb{R}^2) = 2.$$

So the vector subspace given by the image of the differential is indeed 2-dimensional. This means that the images of all the vectors by $d\mathbf{x}_q$ lie on a plane. It is our intention now, to show that this plane is exactly the set of the tangent vectors to S at point q. Let w be a tangent vector at $\mathbf{x}(q) = p$, that is, let $w = \alpha'(0)$, where

$$\alpha : (-\epsilon, \epsilon) \to \mathbf{x}(U) \in S$$

is differentiable and $\alpha(0) = \mathbf{x}(q) = p$. Then the curve

$$\beta = \mathbf{x}^{-1} \circ \alpha : (-\epsilon, \epsilon) \to U$$

is differentiable. By definition of the differential, we have $d\mathbf{x}_q(\beta'(0)) = w \in S$, so $w \in d\mathbf{x}_q$. Hence, we have just shown that $T_pS \subset d\mathbf{x}_q(\mathbb{R}^2)$. We want now to prove that

the vice versa is also true.
On the other hand, $w = d\mathbf{x}_q(v)$, where $v \in \mathbb{R}^2$ is the velocity vector of the curve

$$\gamma : (-\epsilon, \epsilon) \to U$$

given by

$$\gamma(y) = tv + q, \qquad t \in (-\epsilon, \epsilon).$$

By the definition of the differential, $w = \alpha'(0)$, where $\alpha = \mathbf{x} \circ \gamma$. This shows that w is a tangent vector, so it is true that $d\mathbf{x}_q(\mathbb{R}^2) \subset T_pS$. □

By the above proposition, the plane $d\mathbf{x}_q(\mathbb{R}^2)$ passing through $\mathbf{x}(q) = p$ is independent of the parametrization $\mathbf{x}$. However, the choice of a parametrization determines a basis for the plane, called the (ordered) basis associated to $\mathbf{x}$:

$$\left(\left(\frac{\partial \mathbf{x}}{\partial u}\right)(q), \left(\frac{\partial \mathbf{x}}{\partial v}\right)(q)\right) = \left(\mathbf{x}_u(q), \mathbf{x}_v(q)\right).$$

So, the vectors lying on the tangent plane are 3-dimensional objects; however, it might be more useful to identify them through their coordinates with respect to the basis of the plane T_pS rather than through their canonical 3D coordinates.

The coordinates of a vector $w \in T_pS$ in the basis associated to a parametrization $\mathbf{x}$ are determined as follows: w is the velocity vector $\boldsymbol{\alpha}'(0)$ of a curve $\boldsymbol{\alpha} = \mathbf{x} \circ \beta : (-\epsilon, \epsilon) \to S$ where

$$\beta : (-\epsilon, \epsilon) \to U$$

is given by $\beta(t) = (u(t), v(t))$ with $\beta(0) = q = \mathbf{x}^{-1}(p)$. Thus,

$$\begin{aligned}\boldsymbol{\alpha}'(0) &= \frac{d}{dt}(\mathbf{x} \circ \beta)(0) = \frac{d}{dt}\mathbf{x}(u(t), v(t))(0) \\ &= \mathbf{x}_u(q)u'(0) + \mathbf{x}_v(q)v'(0) = w.\end{aligned}$$

Thus, in this basis, w has coordinates $(u'(0), v'(0))$, where $(u(t), v(t))$ is the representation in the parametrization $\mathbf{x}$ of a curve whose velocity vector at $t = 0$ is w.

We want to introduce a metric in the tangent plane. Let $(\mathbf{x}_u, \mathbf{x}_v)$ be a basis of the tangent plane T_pS, induced by a parametrization $\mathbf{x}$ and let

$$\mathbf{a} = a^u\mathbf{x}_u + a^v\mathbf{x}_v, \qquad \mathbf{w} = b^u\mathbf{x}_u + b^v\mathbf{x}_v$$

be two vectors on the tangent plane to S at p. With respect to the basis, we can also write them as:

$$\mathbf{a} = \begin{pmatrix} a^u \\ a^v \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} b^u \\ b^v \end{pmatrix}.$$

We can express the standard Euclidean inner product of $\mathbf{v}$ and $\mathbf{w}$, as vectors of $\mathbb{R}^3$, in terms of these coordinates:

$$\begin{aligned}\langle \mathbf{a}, \mathbf{b}\rangle &= \langle (a^u \mathbf{x}_u + a^v \mathbf{x}_v), (b^u \mathbf{x}_u + b^v \mathbf{x}_v)\rangle \\ &= a^u b^u \langle \mathbf{x}_u, \mathbf{x}_u\rangle + a^u b^v \langle \mathbf{x}_u, \mathbf{x}_v\rangle + a^v b^u \langle \mathbf{x}_v, \mathbf{x}_u)\rangle + a^v b^v \langle \mathbf{x}_v, \mathbf{x}_v\rangle \\ &= (a^u, a^v) \begin{pmatrix} \langle \mathbf{x}_u, \mathbf{x}_u\rangle & \langle \mathbf{x}_u, \mathbf{x}_v\rangle \\ \langle \mathbf{x}_v, \mathbf{x}_u\rangle & \langle \mathbf{x}_v, \mathbf{x}_v\rangle \end{pmatrix} \begin{pmatrix} b^u \\ b^v \end{pmatrix} \\ &= \mathbf{a}^t G \mathbf{b}.\end{aligned}$$

We have expressed the inner product $\langle \mathbf{x}, \mathbf{y}\rangle$ as a matrix multiplication. In this case, G is equal to

$$G = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} = \begin{pmatrix} \langle \mathbf{x}_u, \mathbf{x}_u\rangle & \langle \mathbf{x}_u, \mathbf{x}_v\rangle \\ \langle \mathbf{x}_v, \mathbf{x}_u\rangle & \langle \mathbf{x}_v, \mathbf{x}_v\rangle \end{pmatrix} = \begin{pmatrix} |\mathbf{x}_u|^2 & \langle \mathbf{x}_u, \mathbf{x}_v\rangle \\ \langle \mathbf{x}_v, \mathbf{x}_v\rangle & |\mathbf{x}_v|^2 \end{pmatrix}.$$

We call G the *metric* on T_pS. Each tangent plane has its own metric G, which, therefore, that is a differentiable function of the parameters on which the parametrization depends. The metric is also called *metric tensor* or **first fundamental form**. For convenience, we define $g = \det(G)$.

Definition 2.3.1. The quadratic form $I_p : T_pS \to \mathbb{R}$ on T_pS defined by

$$I_p(v) = \langle v, v\rangle_p = |v|^2 \geq 0$$

is called the **first fundamental form** of the regular surface $S \subset \mathbb{R}^3$ at $p \in S$.

Therefore, the first fundamental form is simply the expression of how the surface S inherits the natural Euclidean inner product of $\mathbb{R}^3$. The first fundamental form allows us to make measurements on the surface (lengths, areas, etc...) without referring back to the ambient space $\mathbb{R}^3$ where the surface lies. As we have seen in Section 1.2.1, we can express the inner product and the metric using Einstein's summation convention:

$$\begin{aligned}\langle \mathbf{a}, \mathbf{b}\rangle &= \left\langle \left(\sum_{i=u,v} a^i \mathbf{x}_i \right), \left(\sum_{j=u,v} b^j \mathbf{x}_j \right) \right\rangle \\ &= \sum_{i=u,v} \sum_{j=u,v} a^i b^j \langle \mathbf{x}_i, \mathbf{x}_j\rangle = \sum_i \sum_j a^i b^j g_{ij} = a^i b^j g_{ij}\end{aligned}$$

In the same way, the length of a vector is

$$|\mathbf{a}| = \sqrt{\langle \mathbf{a}, \mathbf{a}\rangle} = \sqrt{\sum_{i,j} a^i a^j g_{ij}} = \sqrt{a^i a^j g_{ij}}.$$

And the angle θ between two non-zero vectors $\mathbf{a}$ and $\mathbf{b}$, is

$$\cos\theta = \frac{\sum_{i,j} a^i b^j g_{ij}}{\sqrt{\sum_{i,j} a^i a^j g_{ij}} \sqrt{\sum_{i,j} b^i b^j g_{ij}}} = \frac{a^i b^j g_{ij}}{\sqrt{a^i a^j g_{ij}} \sqrt{b^i b^j g_{ij}}}.$$

Example 2.3.1. Let's now calculate, as an example, the first fundamental form of the right cylinder over the unit circle $x^2+y^2=1$ (Figure 2.3). We know from Example 2.2.1 that it admits the parametrization

$$\begin{aligned}&\mathbf{x}(u,v)=(\cos u,\sin u,v),\\&U=\{(u,v)\in\mathbb{R}^2:\quad 0<u<2\pi,\quad -\infty<v<\infty\}.\end{aligned}$$

To compute the first fundamental form we need to derive the parametrization $\mathbf{x}$ with respect to its parameters:

$$\begin{aligned}\mathbf{x}_u&=(-\sin u,\cos u,0),\\\mathbf{x}_v&=(0,0,1).\end{aligned}$$

Since a tangent vector $w\in T_pS$ to S in p is the tangent vector to a parametrized curve $\alpha(t)=\mathbf{x}(u(t),v(t))$, $t\in(-\epsilon,\epsilon)$ with $p=\alpha(0)$ and $w=\alpha'(0)$, the first fundamental form of w is simply

$$\begin{aligned}I_p(w)=I_p(\alpha'(0))&=\langle\alpha'(0),\alpha'(0)\rangle=\langle u'\mathbf{x}_u+v'\mathbf{x}_v,u'\mathbf{x}_u+v'\mathbf{x}_v\rangle\\&=(u')^2\langle\mathbf{x}_u,\mathbf{x}_u\rangle+2u'v'\langle\mathbf{x}_u,\mathbf{x}_v\rangle+(v')^2\langle\mathbf{x}_v,\mathbf{x}_v\rangle\\&=(u')^2(\sin^2u+\cos^2u)+2u'v'(0)+(v')^2(1)^2=(u')^2+(v')^2.\end{aligned}$$

So the first fundamental form of the right cylinder on the unit circle in the basis $(\mathbf{x}_u,\mathbf{x}_v)$ is

$$I(\alpha')=(u')^2+(v')^2.$$

The metric can also be used to calculate areas. Let us begin by considering a vector plane. Given two linearly independent vectors $\mathbf{a}$ and $\mathbf{b}$, we can associate to their exterior product $\mathbf{a}\wedge\mathbf{b}$ the parallelogram U spanned by these vectors, with an orientation (of its boundary) determined by the oriented angle from $\mathbf{a}$ to $\mathbf{b}$; then, $\mathbf{b}\wedge\mathbf{a}=-\mathbf{a}\wedge\mathbf{b}$ is associated to the very same parallelogram with the opposite orientation. In both cases the area of the parallelogram, which we require to be positive, is given by

$$\text{Area}(U)=\text{Area}(\mathbf{a}\wedge\mathbf{b})=|\mathbf{a}\wedge\mathbf{b}|.$$

The area of the parallelogram spanned by $(\mathbf{x}_u,\mathbf{x}_v)$, in the tangent space T_pS, is determined using the metric as follows.

Proposition 2.3.2. Given a parallelogram $U=\mathbf{x}_u\wedge\mathbf{x}_v$, its area is

$$\text{Area}(U)=\text{Area}(\mathbf{x}_u\wedge\mathbf{x}_v)=|\mathbf{x}_u\wedge\mathbf{x}_v|=\sqrt{g}$$

Proof. By definition,

$$|\mathbf{x}_u \wedge \mathbf{x}_v| = \text{Area}(U).$$

But we also have

$$\text{Area}(U) = |\mathbf{x}_u||\mathbf{x}_v| \sin\theta,$$

which is exactly $\sqrt{\det(G)} = \sqrt{g}$, in fact the metric is

$$G = \begin{pmatrix} |\mathbf{x}_u|^2 & \langle \mathbf{x}_u, \mathbf{x}_v \rangle \\ \langle \mathbf{x}_v, \mathbf{x}_u \rangle & |\mathbf{x}_v|^2 \end{pmatrix},$$

and its determinant is

$$\begin{aligned} g &= |\mathbf{x}_u|^2|\mathbf{x}_v|^2 - \langle \mathbf{x}_u, \mathbf{x}_v \rangle \langle \mathbf{x}_v, \mathbf{x}_u \rangle \\ &= |\mathbf{x}_u|^2|\mathbf{x}_v|^2 - \langle \mathbf{x}_u, \mathbf{x}_v \rangle^2 \\ &= |\mathbf{x}_u|^2|\mathbf{x}_v|^2 - |\mathbf{x}_u|^2|\mathbf{x}_v|^2 \cos^2\theta \\ &= |\mathbf{x}_u|^2|\mathbf{x}_v|^2 \sin^2\theta = \text{Area}^2(U^*) \end{aligned}$$

□

The above definition and proposition allow us to define (and compute) locally the area of a surface.

Definition 2.3.2. Let $R \subset \mathbf{x}(U) \subset S$ be a bounded region of a regular surface contained in the coordinate neighborhood of the parametrization $\mathbf{x} : U \subset R^2 \to S$ of S. The positive number

$$\iint_Q |\mathbf{x}_u \wedge \mathbf{x}_v| \, du \, dv = A(R), \quad Q = \mathbf{x}^{-1}(R),$$

is called the **area** of R.

In order to have a global notion of area we have to restrict our discussion to oriented surfaces.

2.3.2 Oriented Surfaces

We shall now introduce the concept of *oriented surface*. Since every point of a regular surface S has a tangent plane T_pS, the choice of an orientation of such a plane induces an orientation on a neighborhood of that point on the surface itself (via a parametrization $\mathbf{x}$). If the local orientations thus defined are coherent on the overlapping coordinate neighborhoods, they determine an orientation of the surface, and the surface is said to be orientable. More formally:

Definition 2.3.3. A regular surface S is called **orientable** if it is possible to cover it with a family of neighborhoods in such a way that if a point $p \in S$ belongs to two neighborhoods of this family, then the change of coordinates has positive Jacobian at p. The choice of such a family is called *orientation* of S and in this case, S is called *oriented*. If such a choice is not possible, S is called *nonorientable*.

Let's make this idea more precise by fixing a parametrization $\mathbf{x}(u, v)$ of a neighborhood of $p \in S$. We determine an orientation of the tangent plane T_pS associated with the orientation of the corresponding basis $(\mathbf{x}_u, \mathbf{x}_v)$. If p also belongs to a neighborhood of another parametrization $\bar{\mathbf{x}}(\bar{u}, \bar{v})$, the new basis $(\bar{\mathbf{x}}_{\bar{u}}, \bar{\mathbf{x}}_{\bar{v}})$ may induce a different orientation. We can write the second basis in terms of the first one by

$$\bar{\mathbf{x}}_{\bar{u}} = \mathbf{x}_u \frac{\partial u}{\partial \bar{u}} + \mathbf{x}_v \frac{\partial v}{\partial \bar{u}},$$
$$\bar{\mathbf{x}}_{\bar{v}} = \mathbf{x}_u \frac{\partial u}{\partial \bar{v}} + \mathbf{x}_v \frac{\partial v}{\partial \bar{v}}.$$

where $u = u(\bar{u}, \bar{v})$ and $v = v(\bar{u}, \bar{v})$ are the expressions of the change of coordinates. The bases $(\mathbf{x}_u, \mathbf{x}_v)$ and $(\bar{\mathbf{x}}_{\bar{u}}, \bar{\mathbf{x}}_{\bar{v}})$ determine the same orientation of T_pS if and only if the Jacobian of the coordinate change is positive:

$$\frac{\partial(u, v)}{\partial(\bar{u}, \bar{v})} > 0.$$

For a surface the property of being orientable is related to the notion of normal field, as the following proposition shows.

Proposition 2.3.3. A regular surface $S \subset \mathbb{R}^3$ is orientable if and only if there exists a differentiable field of unit normal vectors $N : S \to \mathbb{R}^3$ on S.

Proof. If S is orientable, it is possible to cover it with a family of coordinate neighborhoods so that in the intersection of any two of them, the change of coordinates has a positive Jacobian.
At the points $p = \mathbf{x}(u, v)$ of each neighborhood, we define the unit normal vector N at p by

$$N = \frac{\mathbf{x}_u \wedge \mathbf{x}_v}{|\mathbf{x}_u \wedge \mathbf{x}_v|}.$$

$N(p)$ is well defined, since if p belongs to two different coordinate neighborhoods, with parameters (u, v) and $(\bar{u}, \bar{v})$, the normal vector $N(u, v)$ and $N(\bar{u}, \bar{v})$ coincide; in fact:

$$\bar{\mathbf{x}}_{\bar{u}} \wedge \bar{\mathbf{x}}_{\bar{v}} = (\mathbf{x}_u \wedge \mathbf{x}_v) \frac{\partial(u, v)}{\partial(\bar{u}, \bar{v})}$$

preserves its sign ($\partial(u, v)/\partial(\bar{u}, \bar{v}) = +1$). Moreover, the coordinates of $N(u, v)$ are differentiable functions of (u, v), and thus the mapping $N : S \to \mathbb{R}^3$ is differentiable, as desired.
On the other hand, let $N : S \to \mathbb{R}^3$ be a differentiable field of unit normal vectors, and consider a family of connected coordinate neighborhoods covering S. For the points

$p = \mathbf{x}(u, v)$ of each coordinate neighborhood $\mathbf{x}(U)$, with $U \subset \mathbb{R}^2$, it is possible, by the continuity of N, to arrange that

$$N(p) = \frac{\mathbf{x}_u \wedge \mathbf{x}_v}{|\mathbf{x}_u \wedge \mathbf{x}_v|}.$$

In fact, the inner product

$$\left\langle N(p), \frac{\mathbf{x}_u \wedge \mathbf{x}_v}{|\mathbf{x}_u \wedge \mathbf{x}_v|} \right\rangle = f(p) = \pm 1$$

is a continuous function on $\mathbf{x}(U)$. Since $\mathbf{x}(U)$ is connected, the sign of f is constant. If $f(p) = -1$, we interchange u and v in the parametrization, and the assertion follows. Proceeding in this manner for all the coordinate neighborhoods, we have that in the intersection of any two of them, let's say $\mathbf{x}(u, v)$ and $\mathbf{x}(\bar{u}, \bar{v})$ the Jacobian of the coordinate change

$$\frac{\partial(u, v)}{\partial(\bar{u}, \bar{v})}$$

is certainly positive. Otherwise, we would have

$$\frac{\mathbf{x}_u \wedge \mathbf{x}_v}{|\mathbf{x}_u \wedge \mathbf{x}_v|} = N(p) = -\frac{\bar{\mathbf{x}}_{\bar{u}} \wedge \bar{\mathbf{x}}_{\bar{v}}}{|\bar{\mathbf{x}}_{\bar{u}} \wedge \bar{\mathbf{x}}_{\bar{v}}|} = -N(p).$$

which is a contradiction. Hence, the given family of coordinate neighborhoods after undergoing certain interchanges of u and v satisfies the condition of the previous definition, therefore S is orientable. □

Another important proposition regarding orientable surfaces is the following.

Proposition 2.3.4. If a regular surface is given by $S = \{(x, y, z) \in \mathbb{R}^3 : f(x, y, z) = a\}$, where $f : U \subset \mathbb{R}^3 \to \mathbb{R}$ is differentiable, and a is a regular value of f, then S is orientable.

Proof. Given a point $p = (x_0, y_0, z_0) \in S$, consider the parametrized curve $(x(t), y(t), z(t)), \quad t \in I$ on S passing through p for $t = t_0$. Since the curve is on S, we have

$$f(x(t), y(t), z(t)) = a$$

for all $t \in I$. By differentiating both sides we get

$$f_x(p)\left(\frac{dx}{dt}\right)_{t_0} + f_y(p)\left(\frac{dy}{dt}\right)_{t_0} + f_z(p)\left(\frac{dz}{dt}\right)_{t_0} = \frac{da}{dt} = 0.$$

This shows that the tangent vector to the curve at $t = t_0$ is perpendicular to the vector (f_x, f_y, f_z) at p. We conclude that

$$N(x, y, z) = \left(\frac{f_x}{\sqrt{f_x^2 + f_y^2 + f_z^2}}, \frac{f_y}{\sqrt{f_x^2 + f_y^2 + f_z^2}}, \frac{f_z}{\sqrt{f_x^2 + f_y^2 + f_z^2}}\right)$$

is a differentiable field of unit normal vectors on S. □

Thus, the sphere ($S = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = r^2\}$) and the cylinder ($C = \{(x, y, z) \in \mathbb{R}^3 : x^2+y^2 = r^2\}$) are orientable surfaces. A final remark we ought to make is that orientation is not a local property of a regular surface. Locally, every regular surface is diffeomorphic to an open set in the plane, and, hence, is orientable. Orientation is a global property, in the sense that it involves the whole surface.

2.4 Curvature of a Surface

The rate of change of the tangent line to a curve C is an important geometric entity called *curvature of* C (Definition 2.1.5). It is natural to try to define the curvature of a regular surface by analogy with the curvature of a curve. The idea is to try to measure how rapidly a surface S pulls away from the tangent plane T_pS in a neighborhood of a point $p \in S$. This is equivalent to measuring the rate of change at p of a unit normal vector field N in a neighborhood of p. We shall see that this rate of change is given by a linear map on the tangent plane which cannot be an isometry since isometries preserve distances (and distances from a flat plane to a curved surface will necessarily be distorted), but it is self-adjoint.
To do so, we should recall the concept of orientation. By Proposition 2.3.3, an orientation on an orientable surface S is determined by the choice of a differentiable unit normal field N on S. An orientation on S induces an orientation on each tangent space T_pS, $p \in S$ as follows: choose a basis $(v, w) \in T_pS$; this is defined to be positive if $\langle v \wedge w, N\rangle$ is positive.

We can now introduce the concept of the Gauss map.

Definition 2.4.1 (The Gauss Map). Let $S \subset \mathbb{R}^3$ be a surface with orientation N. The map $N : S \to \mathbb{R}^3$ takes its values in the unit sphere

$$S^2 = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}.$$

The map $N : S \to S^2$ is called the **Gauss map** of S (Figure 2.5).

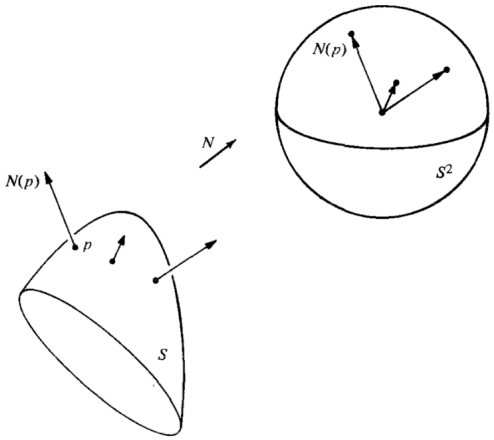

Figure 2.5: The Gauss map. From [Cal00, page 137].

It is straightforward to verify that the Gauss map is differentiable. The differential dN_p of N at $p \in S$ is a linear map from T_pS to $T_{N(p)}S^2$. Since T_pS and $T_{N(p)}S^2$ are parallel planes, dN_p can be looked upon as a linear map on T_pS.

The fact that these two planes are parallel is proven in the following proposition.

Proposition 2.4.1. For each $p \in \mathbb{R}$, the two planes T_pS and $T_{N(p)}S^2$ are parallel to each other.

Proof. The plane T_pS has basis $(\mathbf{x}_u, \mathbf{x}_v)$, while the basis of $T_{N(p)}S^2$ is $(\mathbf{N}_u, \mathbf{N}_v)$. In order to prove that the two planes are parallel, we need to show that they are orthogonal to the same vector, in this case $\mathbf{N}$. Since

$$\mathbf{N} = \frac{\mathbf{x}_u \wedge \mathbf{x}_v}{|\mathbf{x}_u \wedge \mathbf{x}_v|}$$

is normal to the plane spanned by $\mathbf{x}_u$ and $\mathbf{x}_v$ by construction, it is sufficient to show that $\mathbf{N}$ is orthogonal also to the basis of $T_{N(p)}S^2$. Since $|\mathbf{N}|^2 = \langle \mathbf{N}, \mathbf{N} \rangle = 1$, differentiation gives:

$$\frac{\partial}{\partial u}\langle \mathbf{N}, \mathbf{N} \rangle = \left\langle \frac{\partial \mathbf{N}}{\partial u}, \mathbf{N} \right\rangle + \left\langle \mathbf{N}, \frac{\partial \mathbf{N}}{\partial u} \right\rangle = 2\langle \mathbf{N}_u, \mathbf{N} \rangle = 0.$$

Hence, $\mathbf{N}_u \perp \mathbf{N}$; and the same happens deriving with respect to v. □

Based on the above statement, from now on we will identify $T_{N(p)}S^2$ with T_pS. The linear map $dN_p : T_pS \to T_pS$ operates as follows. For each parametrized curve $\alpha(t)$ in S with $\alpha(0) = p$, we consider the parametrized curve $N \circ \alpha(t) = \alpha(t)$ in the sphere S^2. This is equivalent to restricting the normal vector N to the curve $\alpha(t)$. The tangent vector $N'(0) = dN_p(\alpha'(0))$ is a vector in $T_{N(p)}S^2$ restricted to the curve $\alpha(t)$ at $t = 0$.

Thus, dN_p measures how N pulls away from $N(p)$ in a neighborhood of p. In the case of curves, this measure is given by a number; in the case of surfaces, it is characterized by a linear map.
We said earlier that the differential of the map measuring the curvature of a surface is self-adjoint[1]. Let's now prove it.

Proposition 2.4.2. The differential $dN_p : T_pS \to T_pS$ of the Gauss map is a self-adjoint linear map.

Proof. Since dN_p is linear, it is enough to verify that

$$\langle dN_p(w_1), w_2\rangle = \langle w_1, dN_p(w_2)\rangle$$

for a basis (w_1, w_2) of T_pS. Let $\mathbf{x}(u, v)$ be a parametrization of S at p, and let $(\mathbf{x}_u, \mathbf{x}_v)$ be the associated basis of T_pS. If $\alpha(t) = \mathbf{x}(u(t), v(t))$ is a parametrized curve in S, with $\alpha(0) = p$, then

$$\begin{aligned} dN_p(\alpha'(0)) &= dN_p(\mathbf{x}_u u'(0) + \mathbf{x}_v v'(0)) \\ &= dN_p(\mathbf{x}_u)u'(0) + dN_p(\mathbf{x}_v)v'(0) \\ &= \frac{d}{dt} N(u(t), v(t))\Big|_{t=0} \\ &= N_u u'(0) + N_v v'(0); \end{aligned}$$

in particular, $dN_p(\mathbf{x}_u) = N_u$ and $dN_p(\mathbf{x}_v) = N_v$. Therefore, to prove that dN_p is self-adjoint, we only need to show that

$$\langle N_u, \mathbf{x}_u\rangle = \langle \mathbf{x}_u, N_v\rangle.$$

Take the derivatives of $\langle N, \mathbf{x}_u\rangle = 0$ and $\langle N, \mathbf{x}_v\rangle = 0$, relative to v and u and obtain

$$\begin{aligned} \frac{\partial\langle N, \mathbf{x}_u\rangle}{\partial v} &= \langle N_v, \mathbf{x}_u\rangle + \langle N, \mathbf{x}_{uv}\rangle = 0, \\ \frac{\langle N, \mathbf{x}_v\rangle}{\partial u} &= \langle N_u, \mathbf{x}_v\rangle + \langle N, \mathbf{x}_{vu}\rangle = 0. \end{aligned}$$

Thus,

$$\langle N_u, \mathbf{x}_v\rangle = -\langle N, \mathbf{x}_{uv}\rangle = \langle N_v, \mathbf{x}_u\rangle.$$

□

The fact that dN_p is a self-adjoint linear map allows us to associate a quadratic form Q in T_pS given by $Q(v) = \langle dN_p(v), v\rangle$, $v \in T_pS$ to it.

[1] A linear map $A : V \to V$ is self-adjoint if $A = A^\dagger$, that is, if $\langle Av, w\rangle = \langle v, Aw\rangle \quad \forall\, v, w \in V$

Definition 2.4.2. The quadratic form II_p defined in T_pS by

$$II_p(v) = -\langle dN_p(v), v\rangle$$

is called the **second fundamental form** of S at p.

In the previous section, we introduced the notion of curvature of a curve (see Definition 2.1.5). We now wish to extend this concept to a curve lying *on* a surface.

Definition 2.4.3. Let C be a regular curve in S passing through $p \in S$, k the curvature of C at point p, and $\cos\theta = \langle n, N\rangle$, where n and N are the normal vectors at p to the curve C and the surface S respectively.
The number $k_n = k\cos\theta = k\langle n, N\rangle$ is called **normal curvature** of $C \subset S$ at p.

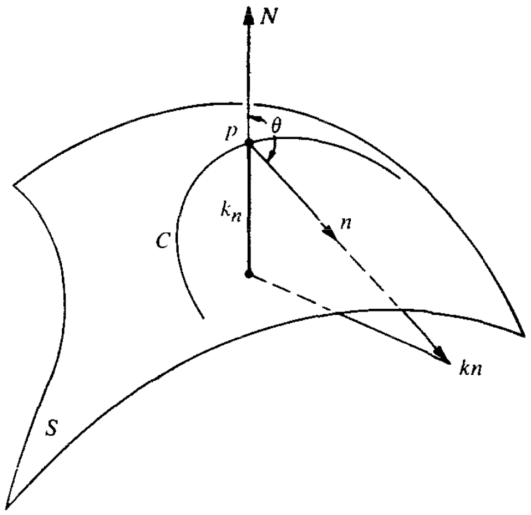

Figure 2.6: Normal curvature of a curve C on a surface S. From [dC76, page 141].

In other words, referring to Figure 2.6, the normal curvature k_n is the projection of the vector kn over the normal to the surface at p, with a sign given by the orientation N of S at p. One important remark is that the normal curvature of C does not depend on its orientation, but it changes sign with a change of orientation of the surface. Now let's give an interpretation to the second fundamental form II_p. Consider a regular curve $C \subset S$ parametrized by $\alpha(s)$, where s is the arc length of C, and with $\alpha(0) = p \in C$. Denoting by $N(s) = N \circ \alpha(s)$ the restriction of N to the curve $\alpha(s)$, we have $\langle N(s), \alpha'(s)\rangle = 0$ because N is the normal vector and $\alpha'(s)$ is the tangent vector to the curve at p, so they are orthogonal. Thus,

$$\frac{d}{dt}(\langle N(s), \alpha'(s)\rangle) = \langle N'(s), \alpha'(s)\rangle + \langle N(s), \alpha''(s)\rangle = 0,$$

which becomes $\langle N'(s), \alpha'(s)\rangle = -\langle N(s), \alpha''(s)\rangle$. Therefore,

$$\begin{aligned} II_p(\alpha'(0)) &= -\langle dN_p(\alpha'(0)), \alpha'(0)\rangle \\ &= -\langle N'(0), \alpha'(0)\rangle = \langle N(0), \alpha''(0)\rangle \\ &= \langle N(0), kn\rangle = k\langle N, n\rangle(p) = k_n(p) = k\cos(\theta)(p). \end{aligned}$$

So the normal curvature k_n is linked to the second fundamental form II_p by this relation: the second fundamental form of a unit vector $v \in T_p(S)$ at p is equal to the normal curvature of a regular curve passing through p and tangent to v.

Since dN_p is a self-adjoint linear map (see Proposition 2.4.2), it is similar to a diagonal matrix, that is, it has two linearly independent eigenvectors and, therefore, two (potentially equal) eigenvalues:

$$\forall\, p \quad \exists\, \mathbf{e}_1, \mathbf{e}_2 \in T_pS \quad \text{and} \quad k_1, k_2 \in \mathbb{R}: \quad dN_p(\mathbf{e}_1) = -k_1\mathbf{e}_1, \quad dN_p(\mathbf{e}_2) = -k_2\mathbf{e}_2.$$

This means that in the orthonormal basis $(\mathbf{e}_1, \mathbf{e}_2)$ of T_pS, the differential dN_p assumes the form:

$$dN_p = \begin{pmatrix} -k_1 & 0 \\ 0 & -k_2 \end{pmatrix}.$$

Moreover, we know from linear algebra that k_1 and k_2 are the maximum and the minimum of the second fundamental form $II_p(v) = -\langle dN_p(v), v\rangle$ restricted to the unit circle of T_pS. This is the key factor in the following definition.

Definition 2.4.4. Let p be a point on a regular surface S, and let $dN_p : T_pS \to T_pS$ be the differential of the Gauss map.

The maximum normal curvature k_1 and the minimum normal curvature k_2 are called the **principal curvatures** at p. The corresponding directions (which are given by the eigenvectors $\mathbf{e}_1$ and $\mathbf{e}_2$) are called *principal directions* at p.

The determinant of dN_p is the **Gaussian curvature** K of S at p.
The negative of half of the trace of dN_p is called the **mean curvature** H of S at p.

In terms of the principal curvatures, we can write

$$K = \det(dN_p) = k_1k_2, \qquad H = -\frac{1}{2}\mathrm{Tr}(dN_p) = -\frac{k_1 + k_2}{2}.$$

For instance, in the plane and in the sphere, all directions at all points are principal directions.

Example 2.4.1. Let's consider, as an example, the right cylinder on the unit circle (Figure 2.3) given by

$$S = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 = 1\}.$$

We already know that it admits a parametrization (see Example 2.3.1.)

$$\mathbf{x}(u, v) = (\cos u, \sin u, v).$$

To compute the orientation we need

$$\begin{aligned} \mathbf{x}_u &= (-\sin u, \cos u, 0), \\ \mathbf{x}_v &= (0, 0, 1), \end{aligned}$$

which give

$$N = \frac{\mathbf{x}_u \wedge \mathbf{x}_v}{|\mathbf{x}_u \wedge \mathbf{x}_v|} = (\cos u, \sin u, 0) = (x, y, 0).$$

So the possible unit normal vector fields are

$$\begin{aligned} N &= (-x, -y, 0), \\ \bar{N} &= (x, y, 0). \end{aligned}$$

We fix an orientation by choosing one of them as normal vector field, let's say $N = (-x, -y, 0)$.
Let's consider a curve $\alpha(t) = (x(t), y(t), z(t))$ on the cylinder, that is with $x(t)^2 + y(t)^2 = 1$. Along this curve, the normal vector is $N(t) = (-x, -y, 0)$, and therefore,

$$dN(x'(t), y'(t), z'(t)) = N'(t) = (-x'(t), -y'(t), 0).$$

If $\mathbf{v}$ is a vector tangent to the cylinder at $p = (x_0, y_0, z_0)$ and parallel to the z-axis, then it must be proportional to $\mathbf{v} = (0, 0, 1)$. To compute $dN(v)$, we need to calculate $N'(t)$ along a curve $\alpha(t)$ with velocity v and such that $\alpha(0) = p$. Let's consider

$$\alpha(t) = (x_0, y_0, z_0 + t),$$

which satisfies $\alpha(0) = (x_0, y_0, z_0) = p$, and $\alpha'(0) = (0, 0, 1) = v$; so the tangent vector to the curve α at p is exactly v.
Let's now compute $N(\alpha(t))$:

$$N(\alpha(t)) = N(x_0, y_0, z_0 + t) = (-x_0, -y_0, 0),$$

and its derivative is

$$N'(0) = \frac{d}{dt} N(\alpha(t))\Big|_{t=0} = (-x_0', -y_0', 0) = (0, 0, 0)$$

because x_0 and y_0 are constant. So the differential of v is given by

$$dN(v) = dN(\alpha'(0)) = N'(0) = (0, 0, 0) = 0\, v.$$

Hence, v, which is parallel to the z-axis, is an eigenvector with eigenvalue 0.
On the other hand, if w is a vector tangent to the cylinder at $p = (x_0, y_0, z_0)$ and parallel to the xy plane, it must be of the form

$$w = (-y_0, x_0, 0)$$

which is in fact perpendicular to $N = (-x(t), -y(t), 0)$. Just like we did before, to compute the differential at w, we need to calculate $N'(t)$ along a curve $\beta(t)$ with velocity w and such that $\beta(0) = p$. Let's consider

$$\beta(t) = (x_0 - y_0 t, y_0 + x_0 t, z_0),$$

which satisfies $\beta(0) = (x_0, y_0, z_0) = p$, and $\beta'(0) = (-y_0, x_0, 0) = w$, so the tangent vector to the curve β at p is exactly w.
Let's now compute $N(\beta(t))$:

$$N(\beta(t)) = N(x_0 - y_0 t, y_0 + x_0 t, z_0) = (-x_0 + y_0 t, -y_0 - x_0 t, 0),$$

and its derivative is

$$N'(0) = \frac{d}{dt} N(\beta(t)\Big|_{t=0} = (y_0, -x_0, 0).$$

So the differential of w is given by

$$dN(w) = dN(\beta'(0)) = N'(0) = (y_0, x_0, 0) = -w.$$

Hence, w, which is parallel to the xy-plane, is an eigenvector with eigenvalue -1.

As one can see from Figure 2.7, the normal sections at a point p vary from a circle perpendicular to the axis of the cylinder to a straight line parallel to it.

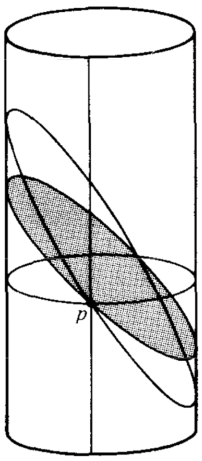

Figure 2.7: Normal sections on a cylinder. Adapted from [dC76, page 144].

Let's calculate the normal curvatures of these sections:

$$k_n = k\cos\theta,$$

where $\cos\theta = 1$ because the normal vector to these curves coincides with the normal vector to the cylinder. In the case of the circle, $k = 1$, so its normal curvature is $k_1 = 1$. Instead, if we consider the straight line, $k = 0$, hence $k_2 = 0$. Thus, the normal curvatures of these sections vary from $k_1 = 1$ (circle) to $k_2 = 0$ (straight line), passing through a family of ellipses with intermediate normal curvature. These values are the maximum and the minimum of the normal curvature at p, so, according to Definition 2.4.4, they are the *principal curvatures* at p, and the corresponding directions, w and v respectively, are the *principal directions* at p.
From the principal curvatures we can calculate the Gaussian and the mean curvature:

$$K = k_1 k_2 = 0, \qquad H = \frac{1}{2}(k_1 + k_2) = \frac{1}{2}.$$

Finally, let's calculate the second fundamental form of the cylinder. Since a tangent vector $w \in T_pS$ is the tangent vector to a parametrized curve $\alpha(t) = \mathbf{x}(u(t), v(t)) = (\cos u, \sin u, v)$ with $p = \alpha(0)$ and $w = \alpha'(0)$, the second fundamental form of w is simply

$$II_p(w) = II_p(\alpha'(0)) = -\langle dN(\alpha'(0)), \alpha'(0)\rangle = -\langle N'(0), \alpha'(0)\rangle,$$

where,

$$\begin{aligned}
\alpha' &= u'\mathbf{x}_u + v'\mathbf{x}_v,\\
\mathbf{x}_u &= (-\sin u, \cos u, 0),\\
\mathbf{x}_v &= (0, 0, 1),
\end{aligned}$$

and

$$N = (-\cos u, -\sin u, 0),$$

with derivative

$$\begin{aligned} N' &= u'N_u + v'N_v, \\ N_u &= (\sin u, -\cos u, 0), \\ N_v &= (0, 0, 0). \end{aligned}$$

Hence, the second fundamental form becomes

$$\begin{aligned} II_p(\alpha'(t)) &= -\langle N'(t), \alpha'(t)\rangle = -\langle u'N_u + v'N_v, u'\mathbf{x}_u + v'\mathbf{x}_v\rangle \\ &= -(u')^2\langle N_u, \mathbf{x}_u\rangle - u'v'\langle N_u, \mathbf{x}_v\rangle - u'v'\langle N_v, \mathbf{x}_u\rangle - (v')^2\langle N_v, \mathbf{x}_v\rangle \\ &= (u')^2\langle N, \mathbf{x}_{uu}\rangle + 2u'v'\langle N, \mathbf{x}_{uv}\rangle + (v')^2\langle N, \mathbf{x}_{vv}\rangle, \end{aligned}$$

where we used the fact that $\langle N_u, \mathbf{x}_v\rangle = \langle N_v, \mathbf{x}_u\rangle = -\langle N, \mathbf{x}_{uv}\rangle$ proved in Proposition 2.4.2, and the similar properties $\langle N_u, \mathbf{x}_u\rangle = -\langle N, \mathbf{x}_{uu}\rangle$, $\langle N_v, \mathbf{x}_v\rangle = -\langle N, \mathbf{x}_{vv}\rangle$ obtained deriving $\langle N, \mathbf{x}_u\rangle = 0$ and $\langle N, \mathbf{x}_v\rangle = 0$ by u and v respectively. And since

$$\begin{aligned} \mathbf{x}_{uu} &= (-\cos u, -\sin u, 0), \\ \mathbf{x}_{uv} &= (0, 0, 0), \\ \mathbf{x}_{vv} &= (0, 0, 0), \end{aligned}$$

we have

$$II_p(\alpha'(0)) = (u')^2.$$

The knowledge of the principal curvatures at p allows us to compute the normal curvature along a given direction of T_pS. In fact, let $v \in T_pS$, $|v| = 1$. Since (e_1, e_2) forms an orthonormal basis of T_pS, we have

$$v = e_1 \cos\theta + e_2 \sin\theta,$$

where θ is the angle between v and e_1 in the orientation of T_pS. The normal curvature along v is given by

$$\begin{aligned} k_n &= II_p(v) = -\langle dN_p(v), v\rangle \\ &= -\langle dN_p(e_1\cos\theta + e_2\sin\theta), e_1\cos\theta + e_2\sin\theta\rangle \\ &= \langle k_1e_1\cos\theta + k_2e_2\sin\theta, e_1\cos\theta + e_2\sin\theta\rangle \\ &= k_1\cos^2\theta + k_2\sin^2\theta. \end{aligned}$$

This expression is known as the **Euler formula**, and it represents the second fundamental form in the basis (e_1, e_2).

When mapped onto the Gauss map, a surface undergoes changes such as stretching and compressing, resulting in an image with a different area compared to the original one. This is the geometric interpretation of the Gaussian curvature K, for $K \neq 0$ discussed in the following proposition.

Proposition 2.4.3. Let p be a point of a surface S such that the Gaussian curvature $K(p) \neq 0$, and let V be a connected neighborhood of p where K does not change sign. Then the Gaussian curvature of S at p is given by:

$$K(p) = \lim_{A \to 0} \frac{A'}{A},$$

where A is the area of a region $B \subset V$ containing p, and A' is the area of $N(B)$, which is the image of B by the Gauss map $N : S \to S^2$. This limit is taken through a sequence of regions B_n that converges to p.

Proof. According to Definition 2.3.2, the area A of B is given by

$$A = \iint_R |\mathbf{x}_u \wedge \mathbf{x}_v| \, du \, dv,$$

where $\mathbf{x}(u, v)$ is a parametrization in p, whose coordinate neighborhood contains V, and R is the region in the uv plane corresponding to B (that is, $B = \mathbf{x}(R)$). The area A' of $N(B)$ is

$$A' = \iint_R |N_u \wedge N_v| \, du \, dv.$$

According to Proposition 2.4.1, the two tangent planes T_pS and $T_{N(p)}S^2$ are parallel, so the generators of the latter can be expressed as a linear combination of the basis vectors of the former:

$$\begin{aligned} N_u &= a_{11}\mathbf{x}_u + a_{21}\mathbf{x}_v, \\ N_v &= a_{12}\mathbf{x}_u + a_{22}\mathbf{x}_v. \end{aligned}$$

Hence, the integrand becomes:

$$\begin{aligned} |N_u \wedge N_v| &= |(a_{11}\mathbf{x}_u + a_{21}\mathbf{x}_v) \wedge (a_{12}\mathbf{x}_u + a_{22}\mathbf{x}_v)| \\ &= |a_{11}a_{22}\mathbf{x}_u \wedge \mathbf{x}_v + a_{21}a_{12}\mathbf{x}_v \wedge \mathbf{x}_u| \\ &= |(a_{11}a_{22} - a_{12}a_{21})\mathbf{x}_u \wedge \mathbf{x}_v| = \det(a_{ij})|\mathbf{x}_u \wedge \mathbf{x}_v|, \end{aligned}$$

where a_{ij} is the matrix of the base change from one plane to the other:

$$a_{ij} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

By Definition 2.4.4, we know that the Gaussian curvature is $K = \det(dN_p)$, which happens to be exactly $\det(a_{ij})$. Hence, we have

$$A' = \iint_R K|\mathbf{x}_u \wedge \mathbf{x}_v|\, du\, dv.$$

Going to the limit, we find

$$\begin{aligned}\lim_{A\to 0}\frac{A'}{A} &= \lim_{R\to 0}\frac{A'/R}{A/R}\\ &= \frac{\lim_{R\to 0}(1/R)\iint_R K|\mathbf{x}_u \wedge \mathbf{x}_v|\, du\, dv}{\lim_{R\to 0}(1/R)\iint_R |\mathbf{x}_u \wedge \mathbf{x}_v|\, du\, dv}\\ &= \frac{K|\mathbf{x}_u \wedge \mathbf{x}_v|}{|\mathbf{x}_u \wedge \mathbf{x}_v|} = K.\end{aligned}$$

Note that we have used the mean value theorem for double integrals (see Theorem 2.2.2). This proves the proposition. □

2.5 The Intrinsic Geometry of Surfaces

In the previous sections we introduced the first fundamental form of a surface S (see Definition 2.3.1) and showed how this can be used to compute simple metric concepts (lengths, angles, areas, ...). Many of these concepts can be expressed only in terms of the first fundamental form, hence they are said to be *intrinsic* to the surface. The study of such concepts is called the *intrinsic geometry* of the surface. It is therefore convenient that we state what is meant by two regular surface having equal fundamental forms.

Definition 2.5.1. Given two regular surfaces S and $\bar{S}$, a diffeomorphism $\phi : S \to \bar{S}$ is an **isometry** if $\forall p \in S$ and $\forall$ pairs $w_1, w_2 \in T_pS$ we have

$$\langle w_1, w_2\rangle_p = \langle d\phi_p(w_1).d\phi_p(w_2)\rangle_{\phi(p)}.$$

The surfaces S and $\bar{S}$ are said to be *isometric*.

In other words, a diffeomorphism is an isometry if it preserves the inner product defined on the surface. It follows that an isometry also preserves the first fundamental form:

$$I_p(w) = \langle w, w, \rangle_p = \langle d\phi_p(w), d\phi_p(w)\rangle_{\phi(p)} = I_{\phi(p)}(d\phi_p(w)), \qquad \forall w \in T_pS.$$

Vice versa, if a diffeomorphism preserves the first fundamental form, it is an isometry. If ϕ preserves I, then

$$I_p(w) = I_{\phi(p)}(d\phi_p(w)) \qquad \forall w \in T_pS.$$

So, we have

$$\begin{aligned}&I_p(w_1+w_2)-I_p(w_1)-I_p(w_2)=\\&\langle w_1+w_2,w_1+w_2\rangle-\langle w_1,w_1\rangle-\langle w_2,w_2\rangle=\\&\langle w_1,w_1+w_2\rangle+\langle w_2,w_1+w_2\rangle-\langle w_1,w_1\rangle-\langle w_2,w_2\rangle=\\&2\langle w_1,w_2\rangle.\end{aligned}$$

On the other hand

$$\begin{aligned}&I_{\phi(p)}(d\phi_p(w_1+w_2))-I_{\phi(p)}(d\phi_p(w_1))-I_{\phi(p)}(d\phi_p(w_2))=\\&2\langle d\phi_p(w_1),d\phi_p(w_2)\rangle.\end{aligned}$$

But the two expressions must coincide, therefore, we have

$$\begin{aligned}&I_p(w_1+w_2)-I_p(w_1)-I_p(w_2)=\\&I_{\phi(p)}(d\phi_p(w_1+w_2))-I_{\phi(p)}(d\phi_p(w_1))-I_{\phi(p)}(d\phi_p(w_2))=\\&2\langle w_1,w_2\rangle=2\langle d\phi_p(w_1),d\phi_p(w_2)\rangle,\end{aligned}$$

which becomes

$$\langle w_1,w_2\rangle=\langle d\phi_p(w_1),d\phi_p(w_2)\rangle,$$

so the scalar product is preserved.

Definition 2.5.2. A map $\phi: V\to\bar{S}$ of a neighborhood V of $p\in S$ is a *local isometry* at p if there exists a neighborhood $\bar{V}$ of $\phi(p)\in\bar{S}$ such that $\phi: V\to\bar{V}$ is an isometry. If there exists a local isometry into $\bar{S}$ at every $p\in S$, the surface S is said to be locally isometric.
Two surfaces S and $\bar{S}$ are locally isometric if S is locally isometric to $\bar{S}$ and $\bar{S}$ is locally isometric to S.

Hence, isometries are distance-preserving maps. When two surfaces S and $\bar{S}$ are connected by an isometry $\phi: S\to\bar{S}$, we say that they are isometric. In other words, they can be *developed* on each other. This means that they can be wrapped around each other without stretching or tearing. But under what conditions is this "developing" possible? Gauss took up this question and determined that the surfaces must have the same curvature at the same points. In broad terms, this is the content of one of his most important theorems: the so called **Theorema Egregium**.

Theorem 2.5.1 (**Theorema Egregium**)**.** Let $\mathbf{x}: U\to\mathbb{R}^3$ be a parametrization of the surface S. Then, the Gaussian curvature K can be expressed entirely in terms of the derivatives of the metric tensor $g_{ij}=\langle\mathbf{x}_i,\mathbf{x}_j\rangle$, and thus is an intrinsic feature of S. In other words, the Gaussian curvature of a surface is invariant by local symmetries.

2.6 Geodesics

Special relativity is applied to observers who move at a constant speed along a straight path. However, on curved surfaces there may not be straight paths at all. Nevertheless, there still are paths that are acceleration-free. These paths are the **geodesics**.
It is easy to show that the acceleration of a particle moving along a curve on a parametrized surface has two components. There is a normal component which cannot be cancelled, so any curve has to accelerate to stay on the surface; but there also is a tangent one, which, instead, depends on the curve and will equal zero if the parametrization is properly chosen.
First, we shall recall the definition of a differentiable vector field: a vector field w is differentiable at p if, for some parametrization $\mathbf{x}(u, v)$ in p, the components a and b of $w = a\mathbf{x}_u + b\mathbf{x}_v$ are differentiable functions at p.

Definition 2.6.1. Let S be a regular parametrized surface, and let w be a differentiable vector field in an open set $U \subset S$ and $p \in U$. Let $y \in T_pS$. Consider a parametrized curve

$$\alpha : (-\epsilon, \epsilon) \to U,$$

with $\alpha(0) = p$ and $\alpha'(0) = y$, and let $w(t)$, $t \in (-\epsilon, \epsilon)$, be the restriction of the vector field w to the curve α. The vector obtained by the normal projection of

$$\frac{dw}{dt}(0)$$

onto the plane T_pS is called the **covariant derivative** at p of the vector field w relative to the vector y (Figure 2.8). This covariant derivative is denoted

$$\frac{Dw}{dt}(0) = D_y w(p)$$

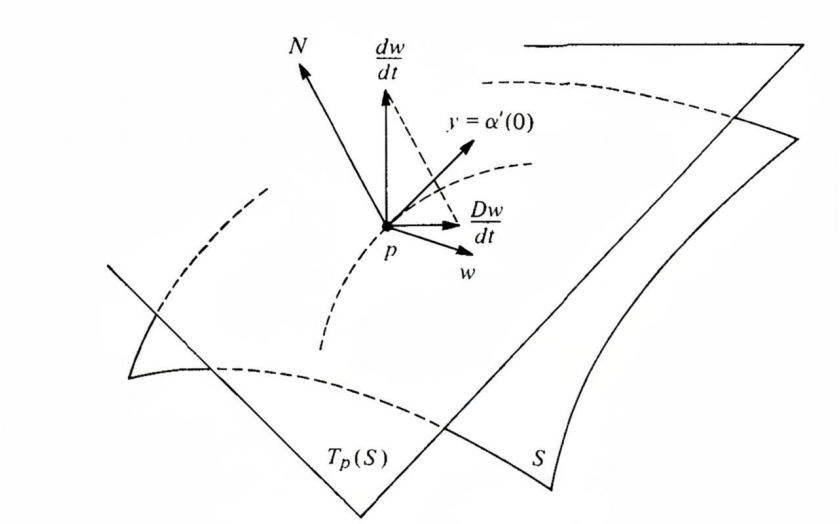

Figure 2.8: Covariant derivative of a vector field w. From [dC76, page 238],.

From a point of view external to the surface, in order to obtain the covariant derivative of a field w along a curve $\alpha : I \to S$ at $t \in I$, one needs to project the ordinary derivative $(dw/dt)(t)$ onto the tangent plane $T_{\alpha(t)}S$. It follows that when two surfaces are tangent along a curve α, then the covariant derivative of a vector field w along α is the same for both surfaces.
If $\alpha(t)$ is a curve on a regular surface S, we can think of it as the trajectory of a point moving on the surface. $\alpha'(t)$ is then the speed of the point, and $\alpha''(t)$ is its acceleration. The covariant derivative $(D\alpha'/dt)$ of the field $\alpha'(t)$ is the tangential component of the acceleration $\alpha''(t)$.

Definition 2.6.2. Let $\alpha : I \to S$ be a parametrized curve in a regular surface S. A vector field w along α is a correspondence that assigns to each $t \in I$ a vector

$$w(t) \in T_{\alpha(t)}S,$$

that is, a vector tangent to the curve α at t.
The vector field w is differentiable at $t_0 \in I$ if, for some parametrization $\mathbf{x}(u, v)$ in $\alpha(t_0)$, the components $a(t)$ and $b(t)$ of $w(t) = a(t)\mathbf{x}_u + b(t)\mathbf{x}_v$ are differentiable functions of t at t_0. w is differentiable in I if it is differentiable $\forall\, t \in I$.

Definition 2.6.3. A vector field w along a parametrized curve $\alpha : I \to S$ is said to be *parallel* if

$$\frac{Dw}{dt} = 0, \quad \forall t \in I.$$

One important property of parallel vector fields is given by the following proposition.

Proposition 2.6.1. Let v and w be parallel vector fields along $\alpha : I \to S$. Then $\langle w(t), v(t)\rangle$ is constant, and in particular $|w(t)|$ and $|v(t)|$ are constant, and the angle between $v(t)$ and $w(t)$ is constant.

Proof. If a vector field w is parallel along α, then its derivative dw/dt is normal to the plane which is tangent to the surface at $\alpha(t)$, and the covariant derivative Dw/dt is null. But the vector field v lies along the tangent plane to the surface, so

$$\langle v(t), w'(t)\rangle = 0.$$

However, this argument is also true the other way around, so

$$\langle w(t), v'(t)\rangle = 0.$$

The two results combined give

$$\langle v(t), w'(t)\rangle + \langle v'(t), w(t)\rangle = \langle v(t), w(t)\rangle' = 0;$$

that is, $\langle v(t), w(t)\rangle$ is constant. □

The next proposition shows that there exist parallel vector fields along a parametrized curve $\alpha(t)$ and that they are completely determined by their values at an arbitrary point t_0.

Proposition 2.6.2. Let $\alpha : I \to S$ be a parametrized curve on a regular surface S, and let $w_0 \in T_{\alpha(t_0)}S$, $t_0 \in I$. Then, there exists a unique parallel vector field $w(t)$ along $\alpha(t)$ with $w(t_0) = w_0$.

We can now introduce the definition of geodesic.

Definition 2.6.4. A nonconstant parametrized curve $\gamma : I \to S$ is said to be **geodesic** at $t \in I$ if the field of its tangent vectors $\gamma'(t)$ is parallel along γ at t, that is, if

$$\frac{D\gamma'(t)}{dt} = 0.$$

γ is a *parametrized geodesic* if it is geodesic for all $t \in I$.

Since $\gamma'(t)$ is a parallel vector field, we immediately know that $|\gamma'(t)| = c \neq 0$ (see Proposition 2.6.1). Therefore, we may introduce the arc length parametrization $s = ct$, and we can conclude that the parameter t of a parametrized geodesic γ is proportional to the arc length of γ. The tangent vector of a geodesic is never zero, thus, the parametrization is regular. The notion of geodesic is clearly local.

Proposition 2.6.3. Given a point $p \in S$ and a vector $w \in T_pS$, $w \neq 0$, there exists an $\epsilon > 0$ and a unique parametrized geodesic $\gamma : (-\epsilon, \epsilon) \to S$ such that $\gamma(0) = p$ and $\gamma'(0) = w$.

Chapter 3

De Sitter Spacetime

In this chapter, we will discuss the geometric and physical interpretations of De Sitter spacetime, one of the simplest and yet most fundamental examples of a curved universe. While special relativity describes the physical phenomena in the flat Minkowski spacetime, this model extends this description to a curved universe. We will start by describing it as a (1+2)-dimensional spacelike hyperboloid of one sheet embedded in the (1+4)-dimensional Minkowski space, and then describe it through a proper parametrization. By studying its realization in a higher-dimensional space, we can explore the behavior of observers, light cones, and photon worldlines within its curved geometry. Finally, this analysis will lead to interesting properties, such as the fact that no photon can travel more than halfway around the spatial circle.

The primary source for the material in this chapter is [Cal00].

3.1 De Sitter Spacetime

The De Sitter universe is one of the simplest curved spacetimes possible. It is the spacelike unit sphere in a (1+4)-dimensional Minkowski space: it is the set of spacelike unit vectors. To see what we mean more concretely, let's consider a (1+2)-dimensional slice of the (1+4)-dimensional ambient space and examine it.

Let $\mathbf{w} = (ct, x, y)$ be a point in the (1+2)-dimensional Minkowski space with the standard metric $|\mathbf{w}|^2 = c^2t^2 - x^2 - y^2$. The spacelike unit vectors $\mathbf{w}$ are such that $|\mathbf{w}|^2 = c^2t^2 - x^2 - y^2 = -1$. They lie on a surface S that in ordinary $\mathbb{R}^3$ forms a *hyperboloid of one sheet* (see Example 2.2.2). So this is the De Sitter spacetime: it extends indefinitely far into the past and into the future, whereas in the spatial direction it is just a circle. We are not used to contemplating a spacetime in which space is finite but has no boundary or edge: it seems possible for an observer or a photon to circumnavigate the entire space making journeys that return to the starting point without ever reversing direction.

3.2 Parametrization of de Sitter Spacetime

To explore the new spacetime, we might want to use the parametrization introduced in Example 2.2.2. We will refer to Figure 2.4.

$$\mathbf{w}(u,v) = (ct, x, y) = (\sinh(u), \cosh(u)\cos(v), \cosh(u)\sin(v))$$

where $-\infty < u < \infty, -\pi \leq v \leq \pi$. The basis vectors of the tangent space are

$$\begin{aligned}\mathbf{w}_u &= (\cosh(u), \sinh(u)\cos(v), \sinh(u)\sin(v)),\\ \mathbf{w}_v &= (0, -\cosh(u)\sin(v), \cosh(u)\cos(v)).\end{aligned}$$

So the components of the metric tensor are

$$\begin{aligned}g_{11} &= \langle \mathbf{w}_u, \mathbf{w}_u\rangle = \cosh^2(u) - \sinh^2(u)\cos^2(v) - \sinh^2(u)\sin^2(v) = 1,\\ g_{12} &= \langle \mathbf{w}_u, \mathbf{w}_v\rangle = \sinh(u)\cosh(u)\cos(v)\sin(v) - \sinh(u)\cosh(u)\cos(v)\sin(v) = 0,\\ g_{22} &= \langle \mathbf{w}_v, \mathbf{w}_v\rangle = -\cosh^2(u)\sin^2(v) - \cosh^2(u)\cos^2(v) = -\cosh^2(u).\end{aligned}$$

So we have just shown that each tangent plane is a (1+1)-dimensional Minkowski space[1] in which, according to the classification made in Section 1.2.2, $\mathbf{w}_u$ is future-timelike, $\mathbf{w}_v$ is spacelike and the two vectors are Minkowski-orthogonal. Moreover, any curve whose tangent vector is always a future-timelike vector can be taken as the worldcurve of an observer. This implies that the coordinate lines v = const are worldcurves, because their tangents $\mathbf{w}_u$ are always future-timelike. In particular, we take the u-axis to be the worldcurve of the observer K. In fact, u is K's proper time τ, as we can see from this calculation:

$$\tau(u) = \int_0^u |\mathbf{w}_u| du = \int_0^u 1 \cdot du = u.$$

Hence, we will use τ and u interchangeably.

According to the shape of the universe, one might conclude that it is possible for a body or a photon to circumnavigate the entire universe since space is infinite but has no boundaries. To address the issue of circumnavigation, let us first measure the circumference of space. At time τ,

$$C = \int_{-\pi}^{\pi} |\mathbf{w}_v| dv = \int_{-\pi}^{\pi} \cosh(\tau) dv = 2\pi\cosh(\tau).$$

So the radius of space at time τ is $\cosh(\tau)$, a value that grows exponentially with $\tau > 0$.

[1] Quadratic forms are classified according to their signature, and in this case the signature of the metric on the plane is $(1+1)$, which is equal to the signature of Minkowski spacetime in 2 dimensions

Now consider, in the (u, v)-plane, a photon emitted by K at the event in O in the positive v-direction and detected at the event E_2. It might have a worldcurve where event E_1 appears twice on the worldcurve simply because the chart "wraps around" in the v-direction, as shown in Figure 3.1.

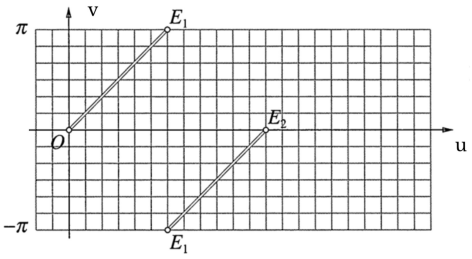

Figure 3.1: Hypothesis for the worldcurve for a photon on the (u, v)-plane. Adapted from [Cal00, page 233].

However, this does not happen: in fact, we can prove that no photon can travel more than halfway around the circle. Let's consider the light cone determined in each tangent plane by the metric tensor:

$$G = \begin{pmatrix} 1 & 0 \\ 0 & -\cosh^2(u) \end{pmatrix}, \quad g = \det(G) = -\cosh^2(u) \leq -1.$$

Note that, when $u = 0$, this is exactly the metric of the $(1 + 1)$-dimension Minkowski spacetime.

One should check that the light-like vectors that separate the timelike from the spacelike vectors at the point (u, v) are multiples of

$$L_\pm = \mathbf{w}_u \pm \mathbf{w}_v \text{sech}(u) = \begin{pmatrix} 1 \\ \pm\text{sech}(u) \end{pmatrix}.$$

In fact, their squared norm is equal to

$$\begin{aligned} \langle L_\pm, L_\pm \rangle &= \langle \mathbf{w}_u \pm \mathbf{w}_v \text{sech}(u), \mathbf{w}_u \pm \mathbf{w}_v \text{sech}(u) \rangle \\ &= \langle \mathbf{w}_u, \mathbf{w}_u \rangle \pm 2\text{sech}(u) \langle \mathbf{w}_u, \mathbf{w}_v \rangle + \text{sech}^2(u) \langle \mathbf{w}_v, \mathbf{w}_v \rangle \\ &= 1 + \text{sech}^2(u)(-\cosh^2(u)) = 0. \end{aligned}$$

Hence, $L_\pm$ are lightlike vectors, and, according to what we said in Section 1.2.2, they are the generators of the light cone. Geometrically, L_+ corresponds to the direction of motion of a light ray moving forward along $+v$, while L_- corresponds to that of a light ray moving forward along $-v$.

The slopes of $L_\pm$ are $\pm\text{sech}(u)$ respectively, and they rapidly approach zero as $|u|$ increases. This means that the light cone flattens as $|u|$ increases. There is such a light cone at each point (u, v) in the parameter plane. We can think of the cone as a pair of vector fields (given by $L_\pm$) that define the possible directions of photon worldcurves (see Figure 3.2).

Now suppose that the wordlcurve of a photon is the graph of a function $v = \phi(u) = \phi(\tau)$ in the (u, v)-plane. This graph must be everywhere tangent to one of the light cone fields. Consider, for example, the L_+ field, with slope $\text{sech}(\tau)$. Then

$$\frac{d\phi}{d\tau} = \text{sech}(\tau), \qquad v = \phi(\tau) = \int \text{sech}(u)\, d\tau.$$

To integrate this, we write the integrand as

$$\text{sech}(\tau) = \frac{1}{\cosh(\tau)} = \frac{2}{e^\tau + e^{-\tau}} = \frac{2e^\tau}{e^{2\tau} + 1},$$

and then make the substitution $a = e^\tau$, $da = e^\tau d\tau$.

$$\begin{aligned}\phi(\tau) &= \int \text{sech}(\tau)\, d\tau = \int \frac{2e^\tau}{e^{2\tau}+1}\, d\tau = \int \frac{2}{a^2+1}\, du = 2\arctan(u) + C \\ &= 2\arctan(e^\tau) + C.\end{aligned}$$

When $\tau \to -\infty$, then $e^\tau \to 0$ and $\phi(\tau) \to C$. When $\tau \to \infty$, then $e^\tau \to +\infty$, $\arctan(e^\tau) \to \pi/2$ and $\phi(\tau) \to \pi + C$. As shown in Figure 3.2, each graph $u = \phi(\tau) = 2\arctan(e^\tau) + C$ lies in a horizontal band whose vertical width is π, and they are all vertical translations of one another obtained by changing the value of C.

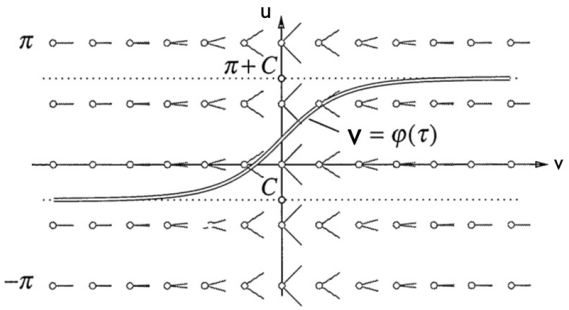

Figure 3.2: Worldcurve of a photon in given by the equation $v = \phi(\tau) = 2\arctan(e^\tau) + C$. The worldcurve is superimposed on the vector fields $L_\pm$ representing the light cones at any point of the (u, v)-plane. Adapted from [Cal00, page 235].

Since every photon worldcurve lies in a horizontal band of vertical width $\Delta v = \pi$, no photon ever travels more than halfway around the circle.

Bibliography

[Bar04] Vincenzo Barone. *Relatività. Principi e Applicazioni.* Bollati Boringhieri, 2004.

[Cal00] James J. Callahan. *The Geometry of Spacetime: An Introduction to Special and General Relativity.* Springer Science & Business Media, 2000.

[dC76] Manfredo P. do Carmo. *Differential Geometry of Curves and Surface.* Prentice-Hall, Inc., 1976.

[Lor04] Hendrik Antoon Lorentz. Electromagnetic phenomena in a system moving with any velocity smaller than that of light. *Proceedings of the Royal Netherlands Academy of Arts and Sciences*, 6, 1904.